Selected Works in Probability and Statistics

Itai Benjamini
Olle Häggström *Editors*

Selected Works of Oded Schramm

Volume 1

Springer

Selected Works in Probability and Statistics

For other titles published in this series, go to
www.springer.com/series/8556

Itai Benjamini · Olle Häggström
Editors

Selected Works of Oded Schramm

Springer

Editors
Itai Benjamini
Weizmann Institute of Science
Department of Mathematics
Rehovot 76100, Israel
itai.benjamini@weizmann.ac.il

Olle Häggström
Department of Mathematical Statistics
Chalmers University of Technology
Göteborg, Sweden
olleh@chalmers.se

Printed in 2 volumes
ISBN 978-1-4419-9674-9 e-ISBN 978-1-4419-9675-6
DOI 10.1007/978-1-4419-9675-6
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011931468

© Springer Science+Business Media, LLC 2011
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface to the Series

Springer's Selected Works in Probability and Statistics series offers scientists and scholars the opportunity of assembling and commenting upon major classical works in statistics, and honors the work of distinguished scholars in probability and statistics. Each volume contains the original papers, original commentary by experts on the subject's papers, and relevant biographies and bibliographies.

Springer is committed to maintaining the volumes in the series with free access of SpringerLink, as well as to the distribution of print volumes. The full text of the volumes is available on SpringerLink with the exception of a small number of articles for which links to their original publisher is included instead. These publishers have graciously agreed to make the articles freely available on their websites. The goal is maximum dissemination of this material.

The subjects of the volumes have been selected by an editorial board consisting of Anirban DasGupta, Peter Hall, Jim Pitman, Michael Sörensen, and Jon Wellner.



Preface

Oded Schramm was born on December 10, 1961, in Jerusalem, and died at the age of 46 in a climbing accident on Guye Peak, WA, on September 1, 2008. In between, he made profound and beautiful contributions to mathematics that will have a lasting influence.

In these two volumes, we have collected some of his papers, supplemented with three survey papers by Steffen Rohde, Olle Häggström and Cristophe Garban that further elucidate his work. Despite the seemingly generous size of the collection, spatial considerations neverthelss forced us to omit most of Oded's papers, and the mere fact that all of them are inspiring pieces of work led to some difficult issues on what to include and what to omit. The reader should not view our choices as an attempt to separate his best works from those that are merely great. Rather, we have tried to put together a representative collection that shows the breadth, depth, enthusiasm and clarity of his work. Others may have diverging opinions on what should or should not have been included, but we do hope that Oded himself would not have been too displeased by our choices. The papers we have included speak for themselves; let us just say a few words about how we have organized them into five sections.

Oded began his mathematical career as a geometer, making GEOMETRY the natural topic for Section 1. This section opens with Oded's first two papers from his Master's thesis in 1987 under Gil Kalai at the Hebrew University. We then move on to circle packing and conformal geometry, including examples of his extraordinarily fruitful collaboration with Zheng-Xu Hu, and end the section with the joint paper with Mario Bonk on embeddings of hyperbolic spaces. Of course geometric aspects permeate also all of the following sections. In fact, Oded once mentioned to one of us that in order to be able to think about a problem he always liked it to have a geometric component.

In the mid 1990's, Oded became interested in the topic of probability theory, which dominates Sections 2-5 of this collection. Section 2 deals in particular with the study of NOISE SENSITIVITY, pioneered in a joint paper with Itai Benjamini and Gil Kalai that opens the section. Noise sensitivity turned out to be a rich topic with applications ranging from voting systems to percolation. In the two other papers of this section, the first coauthored with Jeff Steif and the second with Christophe Garban and Gabor Pete, Oded went progressively deeper into noise sensitivity in a percolation setting, arriving at surprisingly detailed insights.

In Section 3 we have collected some of Oded's papers on RANDOM WALKS AND GRAPH LIMITS. This includes (i) a paper with Itai Benjamini establishing recurrence of random walks on suitably defined limits of finite planar graphs, (ii) the joint paper with his student Omer Angel on distributional limits of triangulations, (iii) a singly-authored paper on compositions of random transpositions, (iv) a collaboration with Yuval Peres, Scott Sheffield and David Wilson on the remarkably fruitful interplay between the infinity Laplacian and certain board games, and (v) the concise 2008 paper on so called hyperfinite graph limits.

One of the probabilistic objects that caught the strongest grip on Oded's imagination was PERCOLATION, which appeared prominently in Section 2 as a major testing ground for noise sensitivity. Papers on other aspects of percolation are collected in Section 4. Here we will see how Oded joined forces with Itai Benjamini, Russ Lyons and Yuval Peres in order to uncover many of the new

and interesting phenomena that happen when we move beyond the usual setting of percolation in a Euclidean geometry, and instead study what happens on hyperbolic lattices and other nonamenable graph structures. Most of the papers in this section are joint work with (subsets of) this team of coauthors, plus Harry Kesten who joined them in establishing a beautiful result on uniform spanning trees. We end the section on a different note, namely Oded's joint paper with Itai Benjamini and Gil Kalai making progress on the important open problem of determining the order of magnitude of fluctuations of first passage percolation on the Euclidean lattice.

Despite strong competition from Oded's other works, there seems to be a consensus view that the most important of all his contributions to mathematics is his discovery and subsequent study of SCHRAMM-LOEWNER EVOLUTION (or stochastic Loewner avolution as Oded himself preferred to call it; conveniently, the abbreviation SLE works either way), which is the topic of Section 5. SLE is a family of conformally invariant random processes in the plane that turn out to appear as the scaling limit of percolation and a variety of other critical models. We begin this section with the famous paper, published in 2000 in the Israel Journal of Mathematics, where Schramm singlehandedly discovered SLE and obtained the first preliminary results on scaling limits. Then followed a series of papers with Greg Lawler and Wendelin Werner in which SLE was exploited to deduce deep results on intersection properties of random walks; here we include only some of the highlights. We furthermore include important joint papers with Steffen Rohde, Scott Sheffield, David Wilson and Stanislav Smirnov, plus Oded's contribution to the International Congress of Mathematicians in Madrid, 2006, in which he gives a survey of the field with an emphasis on open problems.

There were no signs of a decrease in creativity or productivity on Oded's part until the untimely and tragic end, and we can only guess what further discoveries we miss because of it. Substantial parts of his work, joint with others, is still not completely written up and will appear in the coming years.

However, Oded will be missed not just because of his mathematics, but even more because of the gentle, warm-hearted and generous person that he was. Of course, the loss of him is felt most strongly by his wife Avivit, his daughter Tselil and his son Pele. But he was very much loved by the mathematical community and by everyone who knew him, as amply witnessed on the memorial blog which was set up shortly after his death: http://odedschramm.wordpress.com/

We hope that this collection will contribute, however modestly, to keeping the memory of Oded alive, and to nourishing the mathematical heritage he left for all of us. Yehi zichro baruch - may the memory of him be a blessing.

Itai Benjamini
Olle Häggström

Acknowledgements

This series of selected works is possible only because of the efforts and cooperation of many people, societies, and publishers. The series editors originated the series and directed its development. The volume editors spent a great deal of time organizing the volumes and compiling the previously published material. The contributors provided comments on the significance of the papers. The societies and publishers who own the copyright to the original material made the volumes possible and affordable by their generous cooperation:

American Institute of Mathematical Sciences
American Mathematical Society
Duke University Press
Hebrew University Magnes Press
Institute of Mathematical Statistics
Institute of Mathematical Studies
Institut des Hautes Ètudes Scientifiques
International Mathematical Union
The London Mathematical Society
Princeton University and the Institute for Advanced Studies
Springer Science+Business Media
Yuval Peres

Contents

Volume 1

Part 1: Geometry

Commentary: Oded Schramm: From circle packing to SLE, by Steffen Rohde. To appear in *Ann. Probab.* Reprinted with permission of the Institute of Mathematical Statistics 3

O. Schramm. Illuminating sets of constant width. *Mathematika* Vol. 35 No. 2, 180–189 (1988). Reprinted with permission of The London Mathematical Society 47

O. Schramm. On the volume of sets having constant width. *Israel J. Math*. Vol. 63 No. 2, 178–182 (1988). Reprinted with permission of Hebrew University Magnes Press 57

O. Schramm. Rigidity of infinite (circle) packings. *J. Amer. Math. Soc*. Vol. 4 No. 1, 127–149 (1991). Reprinted with permission of the American Mathematical Society 63

O. Schramm. How to cage an egg. *Invent. Math*. Vol. 107 No. 3, 543–560 (1992). Reprinted with permission of Springer Science+Business Media 87

Zheng-Xu He and O. Schramm. Fixed points, Koebe uniformization and circle packings. *Ann. of Math*. (2) Vol. 137 No. 2, 369–406 (1993). Reprinted with permission of Princeton University and the Institute for Advanced Study 105

Zheng-Xu He and O. Schramm. Hyperbolic and parabolic packings. *Discrete Comput. Geom*. Vol. 14 No. 2, 123–149 (1995). Reprinted with permission of Springer Science+Business Media 143

O. Schramm. Circle patterns with the combinatorics of the square grid. *Duke Math. J.* Vol. 86 No. 2, 347–389 (1997). Reprinted with permission of Duke University Press. An electronic version is available at doi: 10.1215/S0012-7094-97-08611-7 171

Zheng-Xu He and O. Schramm. The C∞-convergence of hexagonal disk packings to the Riemann map. *Acta Math.* Vol. 180 No. 2, 219–245 (1998). Reprinted with permission of Springer Science+Business Media 215

M. Bonk and O. Schramm. Embeddings of Gromov hyperbolic spaces. *Geom. Funct. Anal.* Vol. 10 No. 2, 266–306 (2000). Reprinted with permission of Springer Science+Business Media 243

Part 2: Noise Sensitivity

Commentary: Oded Schramm's contributions to noise sensitivity, by Christophe Garban. To appear in *Ann. Probab.* Reprinted with permission of the Institute of Mathematical Statistics 287

I. Benjamini, G. Kalai, and O. Schramm. Noise sensitivity of Boolean functions and applications to percolation. *Inst. Hautes Études Sci. Publ. Math*. No. 90, 5–43 (1999). Reprinted with permission of Institut Des Hautes Études Scientifiques 351

O. Schramm and J. E. Steif. Quantitative noise sensitivity and exceptional times for percolation. *Annals of Mathematics,* Pages 619–672 from Volume 171 (2010), Issue 2. Reprinted with permission of Princeton University and the Institute for Advanced Study 391

C. Garban, G. Pete, and O. Schramm. The Fourier spectrum of critical percolation. *Acta Math*. 205 (2010), 19–104. Reprinted with permission of Springer Science+Business Media 445

Part 3: Random Walks and Graph Limits

I. Benjamini and O. Schramm. Recurrence of distributional limits of finite planar graphs. *Electron. J. Probab*. Vol. 6, No. 23, 13 pp. (electronic) (2001). Reprinted with permission of the Institute of Mathematical Statistics 533

O. Angel and O. Schramm. Uniform infinite planar triangulations. *Comm. Math. Phys*. Vol. 241 No. 2–3, 191–213 (2003). Reprinted with permission of Springer Science+Business Media 547

O. Schramm. Compositions of random transpositions. *Israel J. Math*. Vol. 147, 221–243 (2005). Reprinted with permission of Hebrew University Magnes Press 571

Y. Peres, O. Schramm, S. Sheffield, and D. B. Wilson. Tug-of-war and the infinity Laplacian. *Journal of the American Mathematical Society* 22(1):167–210, (2009). Reprinted with permission of Yuval Peres 595

O. Schramm. Hyperfinite graph limits. *Electron. Res. Announc. Math. Sci.* Vol. 15, 17–23 (2008). Reprinted with permission of the American Institute of Mathematical Sciences 639

Volume 2

Part 4: Percolation

Commentary: Percolation beyond Z^d: The contributions of Oded Schramm, by Olle Häggström. To appear in *Ann. Probab.* Reprinted with permission of the Institute of Mathematical Statistics 649

I. Benjamini and O. Schramm. Percolation beyond Z^d, many questions and a few answers. *Electron. Comm. Probab*. Vol. 1, No. 8, 71–82 (electronic) (1996). Reprinted with permission of the Institute of Mathematical Statistics 679

I. Benjamini, R. Lyons, Y. Peres, and O. Schramm. Critical percolation on any nonamenable group has no infinite clusters. *Ann. Probab*. Vol. 27 No. 3, 1347–1356 (1999). Reprinted with permission of the Institute of Mathematical Statistics 691

R. Lyons and O. Schramm. Indistinguishability of percolation clusters. *Ann. Probab*. Vol. 27 No. 4, 1809–1836 (1999). Reprinted with permission of the Institute of Mathematical Statistics 701

I. Benjamini and O. Schramm. Percolation in the hyperbolic plane. *J. Amer. Math. Soc*. Vol. 14 No. 2, 487–507 (electronic) (2001). Reprinted with permission of the American Mathematical Society 729

I. Benjamini, H. Kesten, Y. Peres, and O. Schramm. Geometry of the uniform spanning forest: transitions in dimensions 4, 8, 12, ... *Ann. of Math*. (2) Vol. 160 No. 2, 465–491 (2004). Reprinted with permission of Princeton University and the Institute for Advanced Study 751

I. Benjamini, G. Kalai, and O. Schramm. First passage percolation has sublinear distance variance. *Ann. Probab.* Vol. 31 No. 4, 1970–1978 (2003). Reprinted with permission of the Institute of Mathematical Statistics 779

Part 5: Schramm-Loewner Evolution

O. Schramm. Scaling limits of loop-erased random walks and uniform spanning trees. *Israel J. Math*. Vol. 118, 221–288 (2000). Reprinted with permission of Hebrew University Magnes Press 791

G. F. Lawler, O. Schramm, and W. Werner. Values of Brownian intersection exponents I: Half-plane exponents. *Acta Math*. Vol. 187 No. 2, 237–273 (2001). Reprinted with permission of Springer Science+Business Media 859

G. F. Lawler, O. Schramm, and W. Werner. Values of Brownian intersection exponents II: Plane exponents. *Acta Math*. Vol. 187 No. 2, 275–308 (2001). Reprinted with permission of Springer Science+Business Media 897

G. F. Lawler, O. Schramm, and W. Werner. Conformal invariance of planar loop-erased random walks and uniform spanning trees. *Ann. Probab*. Vol. 32 No. 1B, 939–995 (2004). Reprinted with permission of the Institute of Mathematical Statistics 931

S. Rohde and O. Schramm. Basic properties of SLE. *Ann. of Math*. (2) Vol. 161 No. 2, 883–924 (2005). Reprinted with permission of Princeton University and the Institute for Advanced Study 989

O. Schramm and S. Sheffield. Contour lines of the two-dimensional discrete Gaussian free field. *Acta Mathematica*, Vol. 202 No. 1, 21–137 (2009). Reprinted with permission of Springer Science+Business Media 1031

O. Schramm, S. Sheffield, and D. B. Wilson. Conformal radii for conformal loop ensembles. *Commun.Math.Phys.*288: 43–53, (2009). Reprinted with permission of Springer Science+Business Media 1149

O. Schramm. Conformally invariant scaling limits: an overview and a collection of problems. In *International Congress of Mathematicians*. Vol. I, 513–543, Eur. Math. Soc., Zürich (2007). Reprinted with permission of the International Mathematical Union 1161

O. Schramm and S. Smirnov. On the scaling limits of planar percolation. To appear in *Ann. Probab*. Reprinted with permission of the Institute of Mathematical Statistics 1193

Contributors

Itai Benjamini The Weizmann Institute of Science, Rehovot POB 76100, Israel, itai.benjamini@weizmann.ac.il

Cristophe Garban Département de mathématiques, CNRS, 46, allée d'Italie, 69364 Lyon Cedex 07, France, christophe.garban@umpa.ens-lyon.fr

Olle Häggström Mathematical Sciences, Chalmers University of Technology, 412 96 Göteborg, Sweden, olleh@chalmers.se

Steffen Rohde University of Washington, Department of Mathematics, C-337 Padelford Hall, Box 354350 Seattle, WA 98195-4350, United States, rohde@math.washington.edu

Author Bibliography

Oded Schramm's publications, as of January 2011

*O. Schramm. Illuminating sets of constant width. *Mathematika* Vol. 35 No. 2, 180–189 (1988).

*O. Schramm. On the volume of sets having constant width. *Israel J. Math*. Vol. 63 No. 2, 178–182 (1988).

O. Schramm. Packing two dimensional bodies with prescribed combinatorics and applications to the construction of conformal and quasiconformal mappings. Ph. D. thesis. Princeton University (1990).

*O. Schramm. Rigidity of infinite (circle) packings. *J. Amer. Math. Soc*. Vol. 4 No. 1, 127–149 (1991).

O. Schramm. Existence and uniqueness of packings with specified combinatorics. *Israel J. Math*. Vol. 73 No. 3, 321–341 (1991).

*O. Schramm. How to cage an egg. *Invent. Math*. Vol. 107 No. 3, 543–560 (1992).

O. Schramm. Square tilings with prescribed combinatorics. *Israel J. Math*. Vol. 84 No. 1–2, 97–118 (1993).

*Zheng-Xu He and O. Schramm. Fixed points, Koebe uniformization and circle packings. *Ann. of Math*. (2) Vol. 137 No. 2, 369–406 (1993).

G. Kuperberg and O. Schramm. Average kissing numbers for non-congruent sphere packings. *Math. Res. Lett.* Vol. 1 No. 3, 339–344 (1994).

Zheng-Xu He and O. Schramm. Rigidity of circle domains whose boundary has σ-finite linear measure. *Invent. Math*. Vol. 115 No. 2, 297–310 (1994).

Zheng-Xu He and O. Schramm. Koebe uniformization for "almost circle domains". *Amer. J. Math*. Vol. 117 No. 3, 653–667 (1995).

*Zheng-Xu He and O. Schramm. Hyperbolic and parabolic packings. *Discrete Comput. Geom*. Vol. 14 No. 2, 123–149 (1995).

Zheng-Xu He and O. Schramm. The inverse Riemann mapping theorem for relative circle domains. *Pacific J. Math*. Vol. 171 No. 1, 157–165 (1995).

O. Schramm. Transboundary extremal length. *J. Anal. Math*. Vol. 66, 307–329 (1995).

O. Schramm. Conformal uniformization and packings. *Israel J. Math*. Vol. 93, 399–428 (1996).

Z. Reich, O. Schramm, V. Brumfeld, and A. Minsky. Chiral discrimination in DNA-peptide interactions involving chiral DNA mesophases: a geometric analysis. *Jour. Amer. Chem. Soc*. Vol. 118, 6345–6349 (1996).

Zheng-Xu He and O. Schramm. On the convergence of circle packings to the Riemann map. *Invent. Math*. Vol. 125 No. 2, 285–305 (1996).

---

*Indicates article found in Vol.1

+Indicates article found in Vol.2

+I. Benjamini and O. Schramm. Percolation beyond Zd, many questions and a few answers. *Electron. Comm. Probab*. Vol. 1, No. 8, 71–82 (electronic) (1996).

I. Benjamini and O. Schramm. Harmonic functions on planar and almost planar graphs and manifolds, via circle packings. *Invent. Math*. Vol. 126 No. 3, 565–587 (1996).

I. Benjamini and O. Schramm. Random walks and harmonic functions on infinite planar graphs using square tilings. *Ann. Probab*. Vol. 24 No. 3, 1219–1238 (1996).

*O. Schramm. Circle patterns with the combinatorics of the square grid. *Duke Math. J.* Vol. 86 No. 2, 347–389 (1997).

I. Benjamini and O. Schramm. Every graph with a positive Cheeger constant contains a tree with a positive Cheeger constant. *Geom. Funct. Anal.* Vol. 7 No. 3, 403–419 (1997).

Zheng-Xu He and O. Schramm. On the distortion of relative circle domain isomorphisms. *J. Anal. Math.* Vol. 73, 115–131 (1997).

*Zheng-Xu He and O. Schramm. The C∞-convergence of hexagonal disk packings to the Riemann map. *Acta Math.* Vol. 180 No. 2, 219–245 (1998).

I. Benjamini and O. Schramm. Exceptional planes of percolation. *Probab. Theory Related Fields* Vol. 111 No. 4, 551–564 (1998).

I. Benjamini and O. Schramm. Conformal invariance of Voronoi percolation. *Comm. Math. Phys*. Vol. 197 No. 1, 75–107 (1998).

I. Benjamini and O. Schramm. Recent progress on percolation beyond Zd. (1998). revised Jan 18 2000 [Web paper]

I. Benjamini, R. Lyons, Y. Peres, and O. Schramm. Group-invariant percolation on graphs. *Geom. Funct. Anal*. Vol. 9 No. 1, 29–66 (1999).

R. Lyons and O. Schramm. Stationary measures for random walks in a random environment with random scenery. *New York J. Math*. Vol. 5, 107–113 (electronic) (1999).

O. Schramm and B. Tsirelson. Trees, not cubes: hypercontractivity, cosiness, and noise stability. *Electron. Comm. Probab*. Vol. 4, 39–49 (electronic) (1999).

+I. Benjamini, R. Lyons, Y. Peres, and O. Schramm. Critical percolation on any nonamenable group has no infinite clusters. *Ann. Probab*. Vol. 27 No. 3, 1347–1356 (1999).

I. Benjamini, R. Lyons, and O. Schramm. Percolation perturbations in potential theory and random walks. In *Random walks and discrete potential theory (Cortona, 1997),* Sympos. Math., XXXIX, 56–84, Cambridge Univ. Press, Cambridge (1999).

+R. Lyons and O. Schramm. Indistinguishability of percolation clusters. *Ann. Probab*. Vol. 27 No. 4, 1809–1836 (1999).

*I. Benjamini, G. Kalai, and O. Schramm. Noise sensitivity of Boolean functions and applications to percolation. *Inst. Hautes Études Sci. Publ. Math*. No. 90, 5–43 (1999).

I. Benjamini, O. Häggström, and O. Schramm. On the effect of adding ε-Bernoulli percolation to everywhere percolating subgraphs of Zd. *J. Math. Phys*. Vol. 41 No. 3, 1294–1297 (2000).

*M. Bonk and O. Schramm. Embeddings of Gromov hyperbolic spaces. *Geom. Funct. Anal.* Vol. 10 No. 2, 266–306 (2000).

+O. Schramm. Scaling limits of loop-erased random walks and uniform spanning trees. *Israel J. Math*. Vol. 118, 221–288 (2000).

J. Jonasson and O. Schramm. On the cover time of planar graphs. *Electron. Comm. Probab.* Vol. 5, 85–90 (electronic) (2000).

+I. Benjamini and O. Schramm. Percolation in the hyperbolic plane. *J. Amer. Math. Soc*. Vol. 14 No. 2, 487–507 (electronic) (2001).

I. Benjamini, R. Lyons, Y. Peres, and O. Schramm. Uniform spanning forests. *Ann. Probab.* Vol. 29 No. 1, 1–65 (2001).

+G.F. Lawler, O. Schramm, and W. Werner. Values of Brownian intersection exponents I: Half-plane exponents. *Acta Math*. Vol. 187 No. 2, 237–273 (2001).

+G.F. Lawler, O. Schramm, and W. Werner. Values of Brownian intersection exponents II: Plane exponents. *Acta Math*. Vol. 187 No. 2, 275–308 (2001).

G.F. Lawler, O. Schramm, and W. Werner. The dimension of the planar Brownian frontier is 4/3. *Math. Res. Lett.* Vol. 8 No. 4, 401–411 (2001).

*I. Benjamini and O. Schramm. Recurrence of distributional limits of finite planar graphs. *Electron. J. Probab*. Vol. 6, No. 23, 13 pp. (electronic) (2001).

O. Schramm. A percolation formula. *Electron. Comm. Probab*. Vol. 6, 115–120 (electronic) (2001).

G.F. Lawler, O. Schramm, and W. Werner. Values of Brownian intersection exponents III: Two-sided exponents. *Ann. Inst. H. Poincaré Probab. Statist*. Vol. 38 No. 1, 109–123 (2002).

G.F. Lawler, O. Schramm, and W. Werner. Analyticity of intersection exponents for planar Brownian motion. *Acta Math*. Vol. 189 No. 2, 179–201 (2002).

G.F. Lawler, O. Schramm, and W. Werner. Sharp estimates for Brownian non-intersection probabilities. In *In and out of equilibrium* (Mambucaba, 2000), Progr. Probab. Vol. 51, 113–131, Birkhäuser Boston, Boston, MA (2002).

G.F. Lawler, O. Schramm, and W. Werner. One-arm exponent for critical 2D percolation Electron. *J. Probab.* Vol. 7, No. 2, 13 pp. (electronic) (2002).

R. Lyons, Y. Peres, and O. Schramm. Markov chain intersections and the loop-erased walk *Ann. Inst. H. Poincaré Probab. Statist*. Vol. 39 No. 5, 779–791 (2003).

+I. Benjamini, G. Kalai, and O. Schramm. First passage percolation has sublinear distance variance. *Ann. Probab.* Vol. 31 No. 4, 1970–1978 (2003).

*O. Angel and O. Schramm. Uniform infinite planar triangulations. *Comm. Math. Phys*. Vol. 241 No. 2–3, 191–213 (2003).

G.F. Lawler, O. Schramm, and W. Werner. Conformal restriction: the chordal case. *J. Amer. Math. Soc*. Vol. 16 No. 4, 917–955 (electronic) (2003).

+I. Benjamini, H. Kesten, Y. Peres, and O. Schramm. Geometry of the uniform spanning forest: transitions in dimensions 4, 8, 12, ... *Ann. of Math*. (2) Vol. 160 No. 2, 465–491 (2004).

+G.F. Lawler, O. Schramm, and W. Werner. Conformal invariance of planar loop-erased random walks and uniform spanning trees. *Ann. Probab*. Vol. 32 No. 1B, 939–995 (2004).

G.F. Lawler, O. Schramm, and W. Werner. On the scaling limit of planar self-avoiding walk. In *Fractal geometry and applications: a jubilee of Benoît Mandelbrot, Part 2, Proc. Sympos. Pure Math. Vol. 72, 339–364, Amer. Math. Soc*., Providence, RI (2004).

I. Benjamini, S. Merenkov, and O. Schramm. A negative answer to Nevanlinna's type question and a parabolic surface with a lot of negative curvature. *Proc. Amer. Math. Soc.* Vol. 132 No. 3, 641–647 (electronic) (2004).

I. Benjamini and O. Schramm. Pinched exponential volume growth implies an infinite dimensional isoperimetric inequality. In *Geometric aspects of functional analysis, Lecture Notes in Math.* Vol. 1850, 73–76, Springer, Berlin (2004).

+S. Rohde and O. Schramm. Basic properties of SLE. *Ann. of Math*. (2) Vol. 161 No. 2, 883–924 (2005).

O. Schramm and S. Sheffield. Harmonic explorer and its convergence to SLE(4). *Ann. Probab.* Vol. 33 No. 6, 2127–2148 (2005).

*O. Schramm. Compositions of random transpositions. *Israel J. Math*. Vol. 147, 221–243 (2005).

I. Benjamini, O. Schramm, and D.B.Wilson. Balanced Boolean functions that can be evaluated so that every input bit is unlikely to be read. In *STOC'05: Proceedings of the 37th Annual ACM Symposium on Theory of Computing*, 244–250, ACM, New York (2005).

O. Schramm. Emergence of Symmetry: Conformal invariance in scaling limits of random systems. In *European congress of mathematics (ECM), Stockholm, Sweden, June 27–July 2, 2004*, Ari Laptev editor, 783–786, European Mathematical Society, Zurich (2005).

O. Schramm and D.B. Wilson. SLE coordinate changes. *New York J. Math*. Vol. 11 :659–669, (2005).

R. O'Donnell, M. Saks, O. Schramm, and R. Servedio. Every decision tree has an influential variable. In *Proceedings of the 46th Annual Symposium on Foundations of Computer Science (FOCS)* (2005).

A. Naor, Y. Peres, O. Schramm, and S. Sheffield. Markov chains in smooth Banach spaces and Gromov-hyperbolic metric spaces. *Duke Math. J.* Vol. 134 No. 1, 165–197 (2006).

R. Lyons, Y. Peres, and O. Schramm. Minimal spanning forests. *Ann. Probab*. Vol. 34 No. 5, 1665–1692 (2006).

Y. Peres, O. Schramm, S. Sheffield, and D.B. Wilson. Random-turn hex and other selection games. *Amer. Math. Monthly* Vol. 114 No. 5, 373–387 (2007).

+O. Schramm. Conformally invariant scaling limits: an overview and a collection of problems. In *International Congress of Mathematicians*. Vol. I, 513–543, Eur. Math. Soc., Zürich (2007).

R. Lyons, B.J. Morris, and O. Schramm. Ends in uniform spanning forests. *Electron. J. Probab*. 13 Paper 58 (2008), 1701–1725.

R. Lyons, R. Peled, and O. Schramm. Growth of the number of spanning trees of the Erdős-Rényi giant component. *Combinatorics, Probability and Computing*,Volume 17 Issue 5, 711–726, (2008).

I. Benjamini, O. Schramm, and A. Shapira. Every minor-closed property of sparse graphs is testable. Proceedings of the 40th annual ACM symposium on Theory of computing, May 17–20, 2008, Victoria, British Columbia, Canada.

*O. Schramm. Hyperfinite graph limits. *Electron. Res. Announc. Math. Sci.* Vol. 15, 17–23 (2008).

*Y. Peres, O. Schramm, S. Sheffield, and D.B. Wilson. Tug-of-war and the infinity Laplacian. *Journal of the American Mathematical Society* 22(1):167–210, (2009).

+O. Schramm and S. Sheffield. Contour lines of the two-dimensional discrete Gaussian free field. *Acta Mathematica*, Volume 202, Number 1, 21–137, (2009).

+O. Schramm, S. Sheffield, and D.B. Wilson. Conformal radii for conformal loop ensembles *Commun.Math.Phys.* 288:43–53, (2009).

Y. Peres, O. Schramm, and J.E. Steif. Dynamical sensitivity of the infinite cluster in critical percolation. *Ann. Inst. H. Poincaré Probab. Statist*. Volume 45, Number 2 (2009), 491–514.

O. Schramm and W. Zhou. Boundary proximity of SLE. *Probability Theory and Related Fields* Volume 146, Numbers 3–4, 435–450, (2009).

A.E. Holroyd, R. Pemantle, Y. Peres, and O. Schramm. Poisson matching. *Ann. Inst. Henri Poincaré Probab. Stat.*, 45(1):266–287, (2009).

*C. Garban, G. Pete, and O. Schramm. The Fourier spectrum of critical percolation. *Acta Math*. 205 (2010), 19–104.

I. Benjamini and O. Schramm. KPZ in one dimensional random geometry of multiplicative cascades. *Commun. Math. Phys*, 289:46–56 (2009).

I. Benjamini, J. Jonasson, O. Schramm, and J. Tykesson.Visibility to infinity in the hyperbolic plane, despite obstacles. *ALEA Lat. Am. J. Probab. Math. Stat.*, Vol. 6 232–343 (2009).

*O. Schramm and J.E. Steif. Quantitative noise sensitivity and exceptional times for percolation. *Annals of Mathematics,* Pages 619–672 from Volume 171 (2010), Issue 2.

C. Garban, G. Pete, and O. Schramm. The scaling limit of the minimal spanning tree - a preliminary report. *XVIth International Congress on Mathematical Physics* (Prague, 2009), ed. P. Exner, pp. 475–480., World Scientific, Singapore, 2010.

I. Benjamini, O. Schramm, and S. Sodin. Poisson asymptotics for random projections of points on a high-dimensional sphere. *Israel J. Math.*, to appear.

+O. Schramm and S. Smirnov. On the scaling limits of planar percolation. *Ann. Probab.*, to appear.

C. Garban, S. Rohde, and O. Schramm. Continuity of the SLE trace in simply connected domains. arXiv:0810.4327

I. Benjamini, O. Gurel-Gurevich, O. Schramm. Cutpoints and resistance of random walk paths. *Ann. Probab.*, to appear.

I. Benjamini and O. Schramm. Lack of sphere packing of graphs via non-linear potential theory. *Jour. of Topology and Analysis*, to appear.

N. Berestycki, O. Schramm, and O. Zeitouni. Mixing times for random k-cycles and coalescence-fragmentation chains. *Ann. Probab.*, to appear.

I. Benjamini, O. Schramm, and A. Timar. Separation. *Groups, Geometry and Dynamics*, to appear.

C. Garban, G. Pete, and O. Schramm. Pivotal, cluster and interface measures for critical planar percolation. arXiv:1008.1378

O. Schramm and S. Sheffield. A contour line of the continuum Gaussian free field. arXiv:1008.2447

Part I
Geometry

Oded Schramm: From Circle Packing to SLE

Steffen Rohde*

July 12, 2010

Contents

1 Introduction **2**

2 Circle Packing and the Koebe Conjecture **3**
2.1 Background . . . 3
2.2 Why are Circle Packings interesting? . . . 5
2.3 Existence of Packings . . . 7
2.4 Uniqueness of Packings . . . 9
2.4.1 Extremal length and the conformal modulus of a quadrilateral . . . 10
2.4.2 The Incompatibility Theorem . . . 11
2.4.3 A simple uniqueness proof . . . 13
2.5 Koebe's Kreisnormierungsproblem . . . 13
2.6 Convergence to conformal maps . . . 14
2.7 Other topics . . . 16

3 The Schramm-Loewner Evolution **18**
3.1 Pre-history . . . 18
3.2 Definition of SLE . . . 19
3.2.1 The (radial) Loewner equation . . . 19
3.2.2 The scaling limit of LERW . . . 21
3.2.3 The chordal Loewner equation, percolation, and the UST . . . 23
3.3 Properties and applications of SLE . . . 25
3.3.1 Locality . . . 25
3.3.2 Intersection exponents and dimensions . . . 27
3.3.3 Path properties . . . 29
3.3.4 Discrete processes converging to SLE . . . 31
3.3.5 Restriction measures . . . 33
3.3.6 Other results . . . 35
3.4 Problems . . . 36
3.5 Conclusion . . . 37

*University of Washington, Supported in part by NSF Grant DMS-0800968.

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_1, © Springer Science+Business Media, LLC 2011

1 Introduction

When I first met Oded Schramm in January 1991 at the University of California, San Diego, he introduced himself as a "Circle Packer". This modest description referred to his Ph.D. thesis around the Koebe-Andreev-Thurston theorem and a discrete version of the Riemann mapping theorem, explained below. In a series of highly original papers, some joint with Zhen-Xu He, he created powerful new tools out of thin air, and provided the field with elegant new ideas. At the time of his deadly accident on September 1st, 2008, he was widely considered as one of the most innovative and influential probabilists of his time. Undoubtedly, he is best known for his invention of what is now called the Schramm-Loewner Evolution (SLE), and for his subsequent collaboration with Greg Lawler and Wendelin Werner that led to such celebrated results as a determination of the intersection exponents of two-dimensional Brownian motion and a proof of Mandelbrot's conjecture about the Hausdorff dimension of the Brownian frontier. But already his previous work bears witness to the brilliance of his mind, and many of his early papers contain both deep and beautifully simple ideas that deserve better knowing.

In this note, I will describe some highlights of his work in circle packings and the Koebe conjecture, as well as on SLE. As Oded has co-authored close to 20 papers related to circle packings and more than 20 papers involving SLE, only a fraction can be discussed in detail here. The transition from circle packing to SLE was through a long sequence of influential papers concerning probability on graphs, many of them written jointly with Itai Benjamini. I will present almost no work from that period (some of these results are described elsewhere in this volume, for instance in Christophe Garban's article on Noise Sensitivity). In that respect, the title of this note is perhaps misleading.

In order to avoid getting lost in technicalities, arguments will be sketched at best, and often ideas of proofs will be illustrated by analogies only. In an attempt to present the evolution of Oded's mathematics, I will describe his work in essentially chronological order.

Oded was a truly exceptional person: not only was his clear and innovative way of thinking an inspiration to everyone who knew him, but also his caring, modest and relaxed attitude generated a comfortable atmosphere. As inappropriate as it might be, I have included some personal anecdotes as well as a few quotes from email exchanges with Oded, in order to at least hint at these sides of Oded that are not visible in the published literature.

This note is not meant to be an overview article about circle packings or SLE. My prime concern is to give a somewhat self-contained account of Oded's contributions. Since SLE has been featured in several excellent articles and even a book, but most of Oded's work on circle packing is accessible only through his original papers, the first part is a bit more expository and contains more background. The expert in either field will find nothing new, and will find a very incomplete list of references. My apology to everyone whose contribution is either unmentioned or, perhaps even worse, mentioned without proper reference.

Acknowledgement: I would like to thank Mario Bonk, Jose Fernández, Jim Gill, Joan Lind, Don Marshall, Wendelin Werner and Michel Zinsmeister for helpful comments on a first draft. I would also like to thank Andrey Mishchenko for generating Figure 3, and Don Marshall for Figure 6.

2 Circle Packing and the Koebe Conjecture

Oded Schramm was able to create, seemingly without effort, ingenious new ideas and methods. Indeed, he would be more likely to invent a new approach than to search the literature for an existing one. In this way, in addition to proving wonderful new theorems, he rediscovered many known results, often with completely new proofs. We will see many examples throughout this note.

Oded received his Ph.D. in 1990 under William Thurston's direction at Princeton. His thesis, and the majority of his work until the mid 90's, was concerned with the fascinating topic of circle packings. Let us begin with some background and a very brief overview of some highlights of this field prior to Oded's thesis. Other surveys are [Sa] and [Ste2].

2.1 Background

According to the Riemann mapping theorem, every *simply connected* planar domain, except the plane itself, is conformally equivalent to a disc. The conformal map to the disc is unique, up to postcomposition with an automorphism of the disc (which is a Möbius transformation). The standard proof exhibits the map as a solution of an extremal problem (among all maps of the domain *into* the disc, maximize the derivative at a given point). The situation is quite different for *multiply connected* domains, partly due to the lack of a standard target domain. The standard proof can be modified to yield a conformal map onto a *parallel slit domain* (each complementary component is a horizontal line segment or a point). Koebe showed that every *finitely connected* domain is conformally equivalent to a *circle domain* (every boundary component is a circle or a point), in an essentially unique way. No proof similar to the standard proof of the Riemann mapping theorem is known.

Theorem 2.1 ([Ko1])**.** *For every domain $\Omega \subset \mathbb{C}$ with finitely many connected boundary components, there is a conformal map f onto a domain $\Omega' \subset \mathbb{C}$ all of whose boundary components are circles or points. Both f and Ω' are unique up to a Möbius transformation.*

Koebe conjectured (p. 358 of [Ko1]) that the same is true for infinitely connected domains. It later turned out that *uniqueness* of the circle domain can fail (for instance, it fails whenever the set of point-components of the boundary has positive area, as a simple application of the measurable Riemann mapping theorem shows). But *existence* of a conformally equivalent circle domain is still open, and is known as Koebe's conjecture or "Kreisnormierungsproblem". It motivated a lot of Oded's research.

There is a close connection between Koebe's theorem and circle packings. A *circle packing* P is a collection (finite or infinite) of closed discs D in the two dimensional plane $\mathbb{C}$, or in the two dimensional sphere S^2, with disjoint interiors. Associated with a circle packing is its *tangency graph* or *nerve* $G = (V, E)$, whose vertices correspond to the discs, and such that two vertices are joined by an edge if and only if the corresponding discs are tangent. We will only consider packings whose tangency graph is connected.

Conversely, the Koebe-Andreev-Thurston *Circle Packing Theorem* guarantees the existence of packings with prescribed combinatorics. Loosely speaking, a planar graph is a graph that can be drawn in the plane so that edges do not cross. Our graphs will not have double edges (two edges with the same endpoints) or loops (an edge whose endpoints coincide).

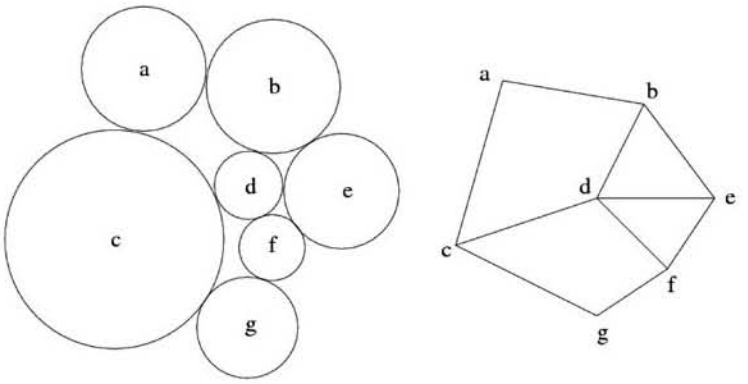


Figure 1: A circle packing and its tangency graph.

Theorem 2.2 ([Ko2], [T], [A1]). *For every finite planar graph G, there is a circle packing in the plane with nerve G. The packing is unique (up to Möbius transformations) if G is a triangulation of S^2.*

See the following sections for the history of this theorem, and sketches of proofs. In particular, in Section 2.3 we will indicate how the Circle Packing Theorem 2.2 can be obtained from the Koebe Theorem 2.1, and conversely that the Koebe theorem can be deduced from the Circle Packing Theorem. Every finite planar graph can be extended (by adding vertices and edges as in Figure 3(c)) to a triangulation, hence packability of triangulations implies packability of finite planar graphs (there are many ways to extend a graph to a triangulation, and uniqueness of the packing is no longer true). The situation is more complicated for infinite graphs. Oded wrote several papers dealing with this case. Thurston conjectured that circle packings approximate conformal maps, in the following

sense: Consider the *hexagonal packing* H_ε of circles of radius ε (a portion is visible in Fig. 2 and Fig. 3(a)). Let $\Omega \subset \mathbb{C}$ be a domain (a connected open set). Approximate Ω from the inside by a circle packing P_ε of circles of $\Omega \cap H_\varepsilon$, as in Fig. 2 and Fig. 3(a) (more precisely, take the connected component containing p of the union of those circles whose six neighbors are still contained in Ω). Complete the nerve of this packing by adding one vertex for each connected component of the complement to obtain a triangulation of the sphere (there are three new vertices v_1, v_2, v_3 in Fig. 3(c); the three copies of v_3 are to be identified). By the Circle Packing Theorem, there is a circle packing P'_ε of the sphere with the same tangency graph (Figures 2 and 3(d) show these packings after stereographic projection from the sphere onto the plane; the circle corresponding to v_3 was chosen as the upper hemisphere and became the outside of the large circle after projection). Notice that each of the complementary components now corresponds to one ("large") circle of P'_ε, and the circles in the boundary of P_ε are tangent to these complementary circles. Now consider the map f_ε that sends the centers of the circles of P_ε to the corresponding centers in P'_ε, and extend it in a piecewise linear fashion. Rodin and Sullivan proved Thurston's conjecture that f_ε approximates

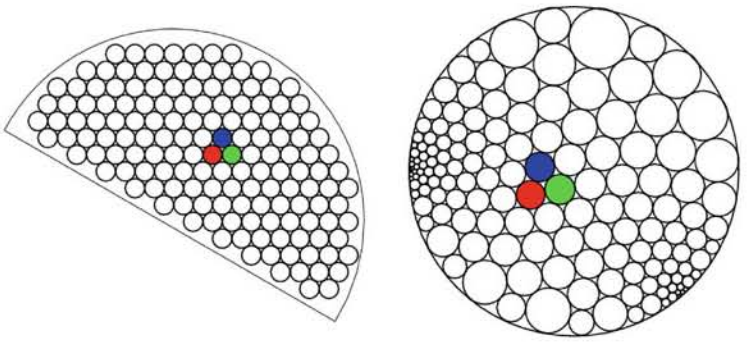


Figure 2: A circle packing approximation to a Riemann map.

the Riemann map, if Ω is simply connected (see Fig. 2):

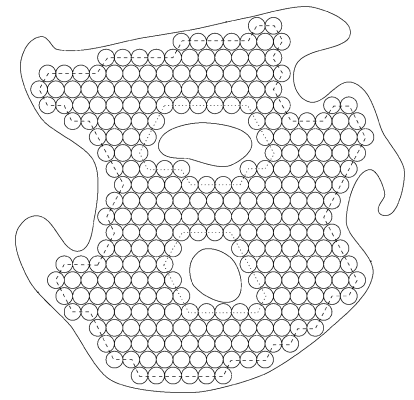
Theorem 2.3. *[RSu] Let Ω be simply connected, $p, q \in \Omega$, and P'_ε normalized such that the complementary circle is the unit circle, and such that the circle closest to p (resp. q) corresponds to a circle containing 0 (resp. some positive real number). Then the above maps f_ε converge to the conformal map $f : \Omega \to \mathbb{D}$ that is normalized by $f(p) = 0$ and $f(q) > 0$, uniformly on compact subsets of Ω as $\varepsilon \to 0$.*

Their proof depends crucially on the non-trivial *uniqueness* of the hexagonal packing as the only packing in the plane with nerve the triangular lattice. Oded found remarkable improvements and generalizations of this theorem. See Section 2.6 for further discussion.

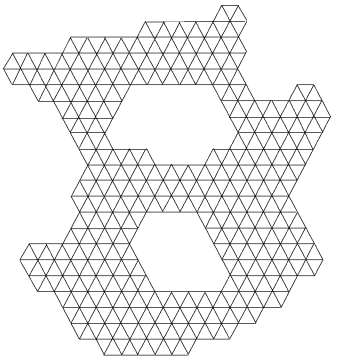
2.2 Why are Circle Packings interesting?

Despite their intrinsic beauty (see the book [Ste2] for stunning illustrations and an elementary introduction), circle packings are interesting because they provide a canonical and conformally natural way to embed a planar graph into a surface. Thus they have applications to combinatorics (for instance the proof of Miller and Thurston [MT] of the Lipton-Tarjan separator theorem, see e.g. the slides of Oded's circle packing talk on his memorial webpage), to differential geometry (for instance the construction of minimal surfaces by Bobenko, Hoffmann and Springborn [BHS] and their references), to geometric analysis (for instance, the Bonk-Kleiner [BK] quasisymmetric parametrization of Ahlfors 2-regular LLC topological spheres) to discrete probability theory (for instance, through the work of Benjamini and Schramm on harmonic functions on graphs and recurrence on random planar graphs [BS1],[BS2], [BS3]) and of course to complex analysis (discrete analytic functions, conformal mapping). However, Oded's work on circle packing did not follow any "main-stream" in conformal geometry or geometric function theory. I believe he continued to work on them just because he liked it. His interest never wavered, and many of his numerous late contributions to Wikipedia were about this topic.

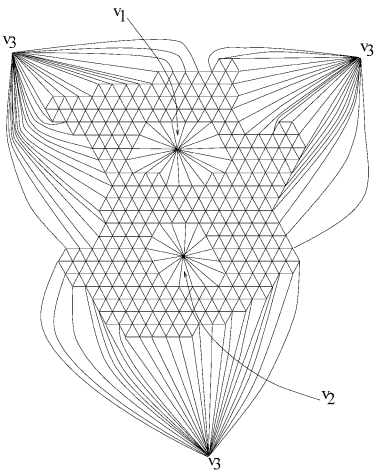
Existence and uniqueness are intimately connected. Nevertheless, for better readability I will discuss them in two separate sections.



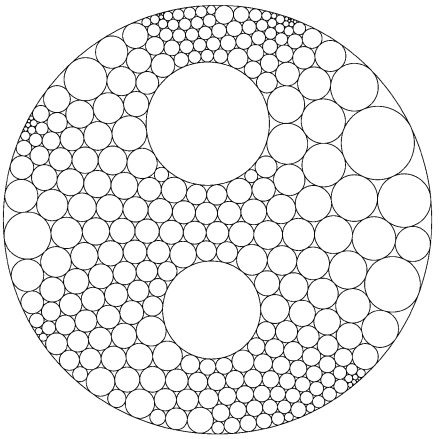
(a)



(b)



(c)



(d)

Figure 3: A circle packing approximation of a triply connected domain, its nerve, its completion to a triangulation of S^2, and a combinatorially equivalent circle packing; (a)-(c) are from Oded's thesis; thanks to Andrey Mishchenko for creating (d)

2.3 Existence of Packings

Oded applied the highest standards to his proofs and was not satisfied with "ugly" proofs. As we shall see, he found four (!) different new existence proofs for circle packings with prescribed combinatorics. Before discussing them, let us have a glance at previous proofs.

The Circle Packing Theorem was first proved by Koebe [Ko2] in 1936. Koebe's proof of existence was based on his earlier result that every planar domain Ω with *finitely many* boundary components, say m, can be mapped conformally onto a *circle domain*. A simple iterative algorithm, due to Koebe, provides an infinite sequence Ω_n of domains conformally equivalent to Ω and such that Ω_n converges to a circle domain. To obtain Ω_{n+1} from Ω_n, just apply the Riemann mapping theorem to the simply connected domain (in $\mathbb{C} \cup \{\infty\}$) containing Ω_n whose boundary corresponds to the $(n \mod m)-$th boundary component of Ω. With the conformal equivalence of finitely connected domains and circle domains established, a circle packing realizing a given tangency pattern can be obtained as a limit of circle domains: Just construct a sequence of m-connected domains so that the boundary components approach each other according to the given tangency pattern. For instance, if the graph $G = (V, E)$ is embedded in the plane by specifying simple curves $\gamma_e : [0, 1] \to S^2$, $e \in E$, then the complement Ω_ε of the set

$$\bigcup_{e \in E} \gamma_e[0, 1/2 - \varepsilon] \cup \bigcup_{e \in E} \gamma_e[1/2 + \varepsilon, 1]$$

is such an approximation. It is not hard to show that the (suitably normalized) conformally equivalent circle domains Ω'_ε converge to the desired circle packing when $\varepsilon \to 0$.

Koebe's theorem was nearly forgotten. In the late 1970's, Thurston rediscovered the circle packing theorem as an interpretation of a result of Andreev [A1], [A2] on convex polyhedra in hyperbolic space, and obtained uniqueness from Mostow's rigidity theorem. He suggested an algorithm to compute a circle packing (see [RSu]) and conjectured Theorem 2.3, which started the field of circle packing. Convergence of Thurston's algorithm was proved in [dV1]. Other existence proofs are based on a Perron family construction (see [Ste2]) and on a variational principle [dV2].

Oded's thesis [S1] was chiefly concerned with a generalization of the existence theorem to packings with prescribed convex shapes instead of discs, and to applications. A consequence ([S1], Proposition 8.1) of his "Monster packing theorem" is, roughly speaking, that the circle packing theorem still holds if discs are replaced by smooth convex sets.

Theorem 2.4. *([S1], Proposition 8.1) For every triangulation $G = (V, E)$ of the sphere, every $a \in V$, every choice of smooth strictly convex sets D_v for $v \in V \setminus \{a\}$, and every smooth simple closed curve C, there is a packing $P = \{P_v : v \in V\}$ with nerve G, such that P_a is the exterior of C and each $P_v, v \in V \setminus \{a\}$ is positively homothetic to D_v.*

Sets A and B are positively homothetic if there is $r > 0$ and $s \in \mathbb{C}$ with $A = rB + s$. Strict convexity (instead of just convexity) was only used to rule out that three of the prescribed sets could meet in one point (after dilation and translation), and thus his packing theorem applied in much more generality. Oded's approach was topological in nature: Based on a cleverly constructed

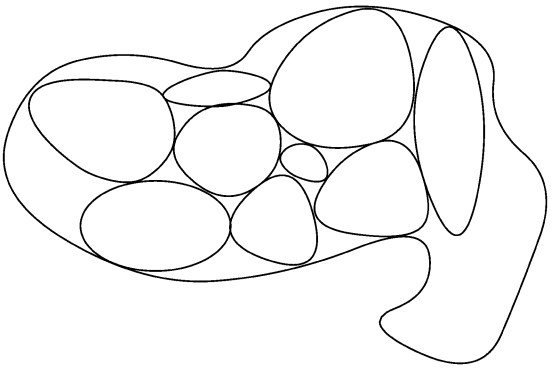


Figure 4: A packing of convex shapes in a Jordan domain, from Oded's thesis

spanning tree of G, he constructed what he called a "monster". This refers to a certain $|V|$-dimensional space of configurations of sets homothetic to the given convex shapes, with tangencies according to the tree, and certain non-intersection properties. Existence of a packing was then obtained as a consequence of Brower's fixed point theorem. Here is a poetic description, quoted from his thesis:

One can just see the terrible monster swinging its arms in sheer rage, the tentacles causing a frightful hiss, as they rub against each other.

Applying Theorem 2.4 to the situation of Figure 3, with D_v chosen as circles when $v \notin \{v_1, v_2, v_3\}$, and arbitrary convex sets D_{v_j}, Oded adopted the Rodin-Sullivan convergence proof to obtain a new proof of the following generalization of Koebe's mapping theorem. The original proof of Courant, Manel and Shiffman [CMS] employed a very different (variational) method.

Theorem 2.5. *([S1], Theorem 9.1; [CMS]) For every $n+1$-connected domain Ω, every simply connected domain $D \subset \mathbb{C}$ and every choice of n convex sets D_j, there are sets D'_j which are positively homothetic to D_j such that Ω is conformally equivalent to $D \setminus \cup_1^n D'_j$.*

Later [S7] he was able to dispose of the convexity assumption, and proved the packing theorem for smoothly bounded but otherwise arbitrary shapes. As a consequence, he was able to generalize Theorem 2.5 to arbitrary (not neccessarily convex) compact connected sets D_j, thus rediscovering a theorem due to Brandt [Br] and Harrington [Ha].

Oded then developed a differentiable approach to the circle packing theorem. In [S3] he shows

Theorem 2.6. *([S3],Theorem 1.1) Let P be a 3-dimensional convex polyhedron, and let $K \subset \mathbb{R}^3$ be a smooth strictly convex body. Then there exists a convex polyhedron $Q \subset \mathbb{R}^3$ combinatorially equivalent to P which midscribes K.*

Here "Q midscribes K" means that all edges of Q are tangent to ∂K. He also shows that the space of such Q is a six-dimensional smooth manifold, if the boundary of K is smooth and has

positive Gaussian curvature. For $K = S^2$, Theorem 2.6 has been stated by Koebe [Ko2] and proved by Thurston [T] using Andreev's theorem [A1], [A2]. Oded notes that Thurston's midscribability proof based on the circle packing theorem can be reversed, so that Theorem 2.6 yields a new proof of the Circle Packing Theorem (given a triangulation, just take $K = S^2$, Q the midscribing convex polyhedron with the combinatorics of the packing, and for each vertex $v \in V$, let D_v be the set of points on S^2 that are visible from v).

One defect of the continuity method in his thesis was that it did not provide a proof of uniqueness (see next section). In [S4] he presented a completely different approach to prove a far more general packing theorem, that had the added benefit of yielding uniqueness, too. A quote from [S4]:

It is just about the most general packing theorem of this kind that one could hope for (it is more general than I have ever hoped for).

A consequence of [S4] (Theorem 3.2 and Theorem 3.5) is

Theorem 2.7. *Let G be a planar graph, and for each vertex $v \in V$, let $\mathcal{F}_v$ be a proper 3-manifold of smooth topological disks in S^2, with the property that the pattern of intersection of any two sets in $\mathcal{F}_v$ is topologically the pattern of intersection of two circles. Then there is a packing P whose nerve is G and which satisfies $P_v \in \mathcal{F}_v$ for $v \in V$.*

The requirement that $\mathcal{F}_v$ is a 3-manifold requires specification of a topology on the space of subsets of S^2 : Say that subsets $A_n \subset S^2$ converge to A if $\limsup A_n = \liminf A_n = A$ and $A^c = \mathrm{int}(\limsup \mathrm{A_n^c})$. An example is obtained by taking a smooth strictly convex set K in $\mathbf{R}^3$ and letting $\mathcal{F}$ be the family of intersections $H \cap \partial K$, where H is any (affine) half-space intersecting the interior of K. Specializing to $K = S^2$, $\mathcal{F}$ is the familiy of circles and the choice $\mathcal{F}_v = \mathcal{F}$ for all v reduces to the circle packing theorem.

The proof of Theorem 2.7 is based on his *incompatibility theorem*, described in the next section. It provides uniqueness of the packing (given some normalization), which is key to proving existence, using continuity and topology (in particular invariance of domains).

2.4 Uniqueness of Packings

I was always impressed by the flexibility of Oded's mind, in particular his ability to let go of a promising idea. If an idea did not yield a desired result, it did not take long for him to come up with a completely different, and in many cases more beautiful, approach. He once told me that if he did not make progress within three days of thinking about a problem, he would move on to different problems.

Following Koebe and Schottky, **uniqueness** of finitely connected circle domains (up to Möbius images) is not hard to show, using the reflection principle: If two circle domains are conformally equivalent, the conformal map can be extended by reflection across each of the boundary circles, to obtain a conformal map between larger domains (that are still circle domains). Continuing in this fashion, one obtains a conformal map between complements of limit sets of reflection groups. As they are Cantor sets of area zero, the map extends to a conformal map of the whole sphere, hence is a Möbius transformation. Uniqueness of the (finite) circle packing can be proved in a similar

fashion. To date, the strongest rigidity result whose proof is based on this method is the following theorem of He and Schramm. See [Bo] for the related rigidity of Sierpinski carpets.

Theorem 2.8 ([HS2], Theorem A)**.** *If Ω is a circle domain whose boundary has $\sigma-$finite length, then Ω is rigid (any conformal map to another circle domain is Möbius).*

For *finite* packings, there are several technically simpler proofs. The shortest and most elementary of them is deferred to the end of this section, since I believe it has been discovered last. Rigidity of *infinite* packings lies deeper. The rigidity of the hexagonal packing, crucial in the proof of the Rodin-Sullivan theorem as elaborated in Section 2.6 below, was originally obtained from deep results of Sullivan's concerning hyperbolic geometry. He's thesis [He] gave a quantitative and simpler proof, still using the above reflection group arguments and the theory of quasiconformal maps. In one of his first papers [S2], Oded gave an elegant combinatorial proof that at the same time was more general:

Theorem 2.9 ([S2], Theorem 1.1)**.** *Let G be an infinite, planar triangulation and P a circle packing on the sphere S^2 with nerve G. If $S^2 \setminus \text{carrier}(P)$ is at most countable, then P is rigid (any other circle packing with the same combinatorics is Möbius equivalent).*

The *carrier* of a packing $\{D_v : v \in V(G)\}$ is the union of the (closed) discs D_v and the "interstices" (bounded by three mutually touching circles) in the complement of the packing. The rigidity of the hexagonal packing follows immediately, since its carrier is the whole plane.

The ingenious new tool is his *Incompatibility Theorem*, a combinatorial analog to the conformal modulus of a quadrilateral. To fully appreciate it, lets first look at its classical continuous counterpart, and defer the statement of the Theorem to Section 2.4.2 below.

2.4.1 Extremal length and the conformal modulus of a quadrilateral

If you conformally map a 3x1-rectangle to a disc, such that the center maps to the center, what fraction of the circle does the image of one of the two short sides occupy? Despite having known the effect of "crowding" in numerical conformal mapping, I was surprised to learn of the numerical value of 0.0114... from Don Marshall (see [MS].) Of course, the precise value can be easily computed as an elliptic integral, but if asked for a rough guess, most answers are around 1/10 (the uniform measure with respect to length would give 1/8). Oded's answer, after a moments thought (during a tennis match in the early 90's), was 1/64, reasoning that this is the probability of a planar random walker to take each of his first three steps "to the right".

An important classical conformal invariant, masterfully employed by Oded in many of his papers, is the *modulus* of a quadrilateral. Let Ω be a simply connected domain in the plane that is bounded by a simple closed curve, and let p_1, p_2, p_3 and p_4 be four consecutive points on $\partial\Omega$. Then there is a unique $M > 0$ such that there is a conformal map $f : \Omega \to [0, M] \times [0, 1]$ and such that f takes the p_j to the four corners with $f(p_1) = 0$ (by a classical theorem of Caratheodory, f extends homeomorphically to the boundary of the domains). There are several quite different instructive proofs of uniqueness of M. Each of the following three techniques has a counterpart in the circle packing world that has been employed by Oded. Suppose we are given two rectangles and a conformal map f between them taking corners to corners.

One method to prove uniqueness is to repeatedly reflect f across the sides of the rectangles. The resulting extention is a conformal map of the plane, hence linear, and it follows that the aspect ratio is unchanged. This is similar to the aforementioned Schottky group argument.

A second method is to explicitly define a quantity λ depending on a configuration $(\Omega, p_1, ..., p_4)$ in such a way that it is conformally invariant and such that one can compute λ for the rectangle $[0, M] \times [0, 1]$. This is achieved by the *extremal length* of the family Γ of all rectifiable curves γ joining two opposite "sides" $[p1, p2]$ and $[p3, p4]$ of Ω. The extremal length of a curve family Γ is defined as

$$\lambda(\Gamma) = \sup_{\rho} \frac{(\inf_\gamma \int_\gamma \rho |dz|)^2}{\int_{\mathbb{C}} \rho^2 dxdy}, \tag{1}$$

where the supremum is over all "metrics" (measurable functions) $\rho : \mathbb{C} \to [0, \infty)$. For the family of curves joining the horizontal sides in the rectangle $[0, M] \times [0, 1]$, it is not hard to show $\lambda(\Gamma) = M$. This simple idea is actually one of the most powerful tools of geometric function theory. See e.g. [Po2] or [GM] for references, properties and applications.

Discrete versions of extremal length (or the "conformal modulus" $1/\lambda$) have been around since the work of Duffin [Duf]. In conformal geometry, they have been very succesfully employed beginning with the groundbreaking paper [Can]. Cannon's extremal length on a graph $G = (V, E)$ is obtained from (1) by viewing non-negative functions $\rho : V \to [0, \infty)$ as metrics on G, defining the length of a "curve" $\gamma \subset V$ as the sum $\sum_{v \in \gamma} \rho(v)$, and the "area" of the graph as $\sum \rho(v)^2$. See [CFP1] for an account of Cannon's discrete Riemann mapping theorem, and for instance the papers [HK] and [BK] concerning applications to quasiconformal geometry. Oded's applications to square packings and transboundary extremal length are briefly discussed in Section 2.7 below.

A third and very different method is topological in nature and is one of the key ideas in [HS1]. Suppose we are given two rectangles Ω, Ω' with different aspect ratio and overlapping as in Fig. 5, and a conformal map f between them mapping corners to corners. Then the difference $f(z) - z$ is $\neq 0$ on the boundary $\partial\Omega$. Traversing $\partial\Omega$ in the positive direction, inspection of Fig. 5 shows that the image curve under $f(z) - z$ winds around 0 in the negative direction. But a negative winding is impossible for analytic functions (by the argument principle, the winding number counts the number of preimages of 0).

2.4.2 The Incompatibility Theorem

Again consider the overlapping rectangles Ω, Ω' of Fig. 5, and two combinatorially equivalent packings P, P' whose nerves triangulate the rectangles, as in Fig. 6. Assume for simplicity that the sets D_v and D'_v of the packings are closed topological discs (except for the four sides $D_1, ...D_4$, $D'_1, ..., D'_4$ of the rectangles, which are considered to be sets of the packing). Intuitively, two topological discs D and D' are called *incompatible* if they intersect as in Fig. 5. More formally, say that D *cuts* D' if there are two points in $D' \setminus \text{interior}(D)$ that cannot be connected by a curve in $\text{interior}(D' \setminus D)$. Then Oded calls D and D' incompatible if D cuts D' or D' cuts D. As he notes, *the motivation for the definition comes from the simple but very important observation that the possible patterns of intersection of two circles are very special, topologically.* Indeed, any two circles are compatible.

Theorem 2.10 ([S2], Theorem 3.1)**.** *There is a vertex v for which D_v and D'_v are incompatible.*

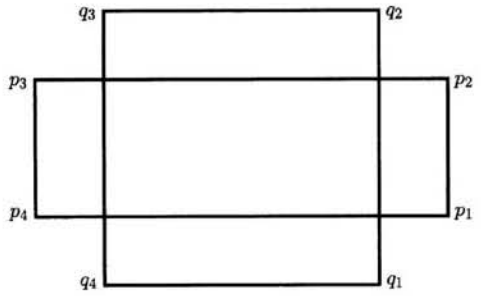


Figure 5: Conformally inequivalent rectangles; from [HS1].

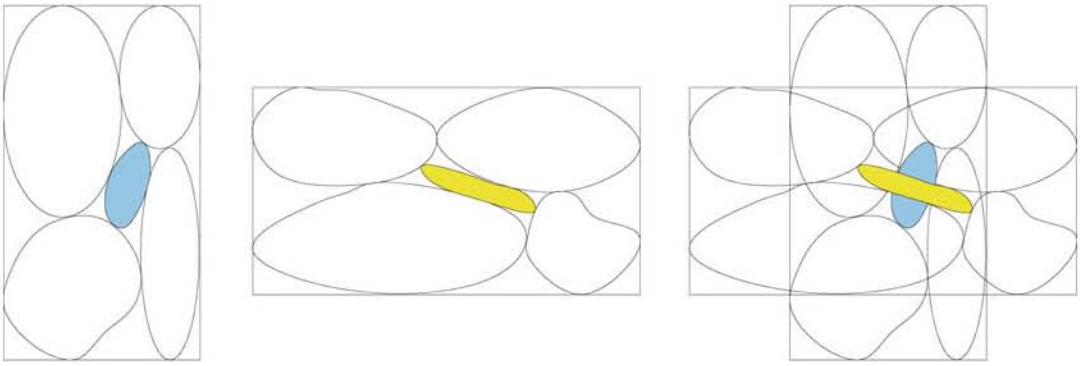


Figure 6: An incompatibility at the center.

Oded calls this result a combinatorial version of the modulus. However, it has rather little in common with the above notion of discrete modulus, except for the setup.

Oded's clever proof by induction on the number of sets in the packing uses arguments from plane topology. An immediate consequence is that two rectangles cannot be packed by the same circle pattern, unless they have the same modulus M and hence are similar: if they could, just place the two packings on top of each other as in Fig. 6 and obtain two incompatible circles, a contradiction. In the same vein, it is not difficult to reduce the proof of the rigidity Theorem 2.9 to an application of the incompatibility theorem.

2.4.3 A simple uniqueness proof

To end this section, here is a beautifully simple proof of the rigidity of finite circle packings whose nerve triangulates S^2. I copied it from the wikipedia (search for circle packing theorem), and believe it is due to Oded. As before, stereographically project the packing to obtain a packing of discs in the plane. This time, assume that the north pole belongs to the complement of the discs, so that the planar packing will consist of three "outer" circles and the remaining circles contained in the interstice between them.

"There is also a more elementary proof based on the maximum principle, which we now sketch. The key observation here is that if you look at the triangle formed by connecting the centers of three mutually tangent circles, then the angle formed at the center of one of the circles is monotone decreasing in its radius and monotone increasing in the two other radii. Consider two packings corresponding to G. First apply reflections and Möbius transformations to make the outer circles in these two packings correspond to each other and have the same radii. Next, consider a vertex v where the ratio between the corresponding radius in the one packing and the corresponding radius in the other packing is maximized. Since the angle sum formed at the center of the corresponding circles is the same (360 degrees) in both packings, it follows from the above observation that the radius ratio is the same at all the neighbors of v as well. Since G is connected, we conclude the radii in the two packings are the same, which proves uniqueness."

2.5 Koebe's Kreisnormierungsproblem

Koebe's 1908 conjecture [Ko1] that every planar domain can be mapped conformally onto a circle domain is still open, despite considerable effort by Koebe and others. Important contributions were made by Grötzsch, Strebel, Sibner and others. One difficulty is the aforementioned lack of uniqueness. Another problem is that Theorem 2.5 is not true in the infinitely connected case, as the following example from [S6] illustrates: If $K = \{x + iy : x = 0, \pm 1, \pm\frac{1}{2}, \pm\frac{1}{3}, ..., y \in [-1, 1]\}$, and if $D = \hat{\mathbb{C}} \setminus K$, then there is no conformal map f of D, normalized by $f(z) - z \to 0$ as $z \to \infty$, such that the component $\{iy : y \in [-1, 1]\}$ of ∂D corresponds to a horizontal line segment (or a point) while the other complementary components of $f(D)$ are vertical line segments. The same example also illustrates the fundamental continuity problem: There is a circle domain D' conformally equivalent to D, but the boundary component corresponding to $\{iy : y \in [-1, 1]\}$ is just a point, so that the conformal map from D' to D cannot be extended to the boundary.

The first joint paper of He and Schramm provided a breakthrough:

Theorem 2.11 ([HS1]). *If Ω has at most countably many boundary components, then Ω is conformally equivalent to a circle domain Ω', and Ω' is unique up to Möbius transformation.*

Essentially, this result is still the strongest to date. Oded later [S6] gave a conceptually different and simpler proof based on his transboundary extremal length, which also applies to certain classes of domains with uncountably many boundary components.

The proof in [HS1] used transfinite induction and was based on the topological concept of the *fixed-point index*. I will illustrate the beautiful idea by sketching their proof of uniqueness. As it turned out, this argument for uniqueness had been given earlier by Strebel [Str]. The simple but crucial idea is to use the following (see [HS1], Lemma 2.2): If f is a fixed-point free orientation preserving homeomorphism between two circles C' and C'', then the winding number of the curve $f(z) - z, z \in C'$, around 0 is non-negative (recall Fig. 5 for a situation where the winding number is negative). Let $f : \Omega' \to \Omega''$ be a conformal map and assume for simplicity that f extends continuously to the boundary (in case of finitely many boundary components this is immediate from the reflection principle, but in the countable case this step is non-trivial), and that f has no fixed points on the boundary. Composing with Möbius transformations, we may assume that $\infty \in \Omega'$ and that $f(z) = z + a_1/z + a_2/z^2 + \cdots$. We want to show that f is the identity. If not, denote a_j the first non-zero Taylor coefficient, then $f(z) - z$ has winding number $-j$ as z traverses a large circle $|z| = R$, because $f(z) - z$ behaves like $a_j z^{-j}$. Moreover, each circular boundary component maps to a circular component. These boundary components are oriented negatively (to keep the domain to the left) and thus, by the above crucial idea, contribute a non-positive number to the winding of $f(z) - z, z \in \partial(\Omega \cap \{|z| \leq R\}$ around 0. Hence the total winding number is negative, contradicting their generalization of the argument principle (the winding number counts the number of zeroes of $f(z) - z$.) Of course, I have swept most details under the rug, most notably the proof of continuity based on a powerful generalization of Schwarz' Lemma to circle domains (Theorem 0.6 in [HS1]).

Combining the fixed-point index method of [HS1] with an analysis of quasiconformal deformations using the reflection group approach and Sullivan's rigidity theorems, He and Schramm [HS4] improve Theorem 2.11 to domains Ω for which all boundary components are circles or points except those in a countable and closed family. They also obtain the following generalization of the Riemann mapping theorem. Let $A \subseteq \mathbb{C}$ be simply connected.

Theorem 2.12 ([HS4],[HS5]). *If $\Omega \subset A$ is a relative circle domain (each connected component of $A \setminus \Omega$ is a point or a closed disc), then there is a relative circle domain Ω^* in $\mathbb{D}$ conformally equivalent to Ω, and so that ∂A corresponds to $\partial\mathbb{D}$. Conversely, if Ω^* is a relative circle domain in $\mathbb{D}$, there is such $\Omega \subset A$.*

The converse direction is the main result of [HS5].

2.6 Convergence to conformal maps

Let us return to the setting of the Rodin-Sullivan Theorem 2.3 about convergence of the discrete map f_ε to the conformal map f. Consider the piecewise linear extension of f_ε from the carrier of P to the carrier of P' that maps equilateral triangles to the corresponding triangles (formed by the centers of P'). By the elementary "Ring Lemma" of [RSu], the angles of these triangles are

bounded away from 0 and π (so that f_ε is quasiconformal with dilation uniformly bounded above). At the heart of the Rodin-Sullivan proof is the *uniqueness* of the hexagonal packing as the only packing in the plane with nerve the triangular lattice (see the discussion in Section 2.4). It rather easily implies that tangent circles centered in a compact set of Ω correspond to tangent circles in $\mathbb{D}$ whose radii are asymptotically equal as $\varepsilon \to 0$. Hence the triangles in P' are nearly equilaterals when ε is small (the angles tend to $\pi/3$), so that f_ε is nearly angle preserving in each triangle. Now the theory of quasiconformal maps readily yields equicontinuity of the family of maps f_ε, and shows that every subsequential limit $\lim f_{\varepsilon_j}$ is a conformal map. The theorem follows from uniqueness of normalized conformal maps.

He's thesis [He] provided a quantitative estimate for the rate of convergence of the angles (the difference to $\pi/3$ is $O(\varepsilon)$). This estimate was known to imply convergence of the ratio of corresponding radii $\mathrm{rad}(\mathrm{D}')/\mathrm{rad}(\mathrm{D})$ to the absolute value $|f'|$ of the derivative. A probabilistic proof of C^0 (locally uniform) convergence of circle packings was given by Stephenson [Ste1]. Convergence of f_ε to f for packings other than the hexagonal was proved in [HR], under the assumption of bounded valency of the graph. In [DHR], the quality of convergence was improved to convergence in C^2 (that is, convergence of first and second derivatives; strictly speaking, instead of f_ε they considered the "piecewise Möbius" map that sends interstices between triples of mutually tangent circles to the corresponding interstices). He and Schramm [HS6] found an elementary new convergence proof, based on the topological ideas discussed above and thus avoiding quasiconformal maps. Their proof also gave convergence up to C^2, and worked in a more general setting. In particular, it does not need the assumption of uniformly bounded degree of [HR].

In the remarkable paper [HS8], He and Schramm proved C^∞-convergence of hexagonal disk packings to the Riemann map:

Theorem 2.13 ([HS8], Theorem 1.1)**.** *The discrete functions $f_\varepsilon : V_\varepsilon \to \mathbb{D}$ converge in C^∞ to the Riemann mapping $f : \Omega \to \mathbb{D}$, in the sense that the discrete partial derivatives of f_ε of any order converge locally uniformly to the corresponding partial derivatives of f.*

The discrete first-order derivatives for $v \in V_\varepsilon$ are

$$\partial_{\varepsilon,k} f_\varepsilon(v) = \varepsilon^{-1}(f_\varepsilon(v + \varepsilon\, \omega^k) - f_\varepsilon(v)),$$

where $k \in 0, 1, ..., 5$ and $\omega = (1 + i\sqrt{3})/2$ is a $6-$th root of unity. In particular, it follows that $(\partial_{\varepsilon,0})^k f_\varepsilon$ converges to the $k-$th derivative $f^{(k)}$ locally uniformly on G.

The Schwarzian derivative

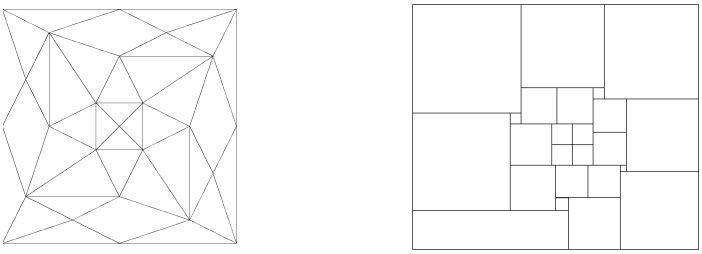
$$S(f)(z) = \frac{f'''(z)}{f'(z)} - \frac{3}{2}\frac{f''(z)^2}{f'(z)^2} \tag{2}$$

of a locally univalent analytic function measures the deviation of f from a Möbius transformation, in particular $S(f) \equiv 0$ if f is Möbius. A key idea in the proof is to define a discrete analog of the Schwarzian derivative, to compute the (discrete) Laplacian of this Schwarzian, and to employ a regularity theorem for discrete elliptic equations to obtain boundedness of all partials of the Schwarzian. The definition of the discrete Schwarzian is the circle packing analog of an invariant that Oded so masterfully employed in his earlier work [S8] on circle patterns with the combinatorics of the square grid.

2.7 Other topics

Oded's approach to both mathematics and to life was extraordinarily innovative and unaccepting of conventions. Notions that most people take for granted without even thinking about, he would open-mindedly question, often coming up with amazing alternative solutions. For example, I would not even think about camping on the foot of a glacier without a sleeping bag. Climbing little Tahoma peak with Oded, he proved to me that even this idea can be pursued. It was perhaps one of his less successful innovations, though.

In the lovely paper [S5], Oded shows that for each triangulation G of a quadrilateral, there is a packing of a rectangle R by (horizontal) squares with the combinatorics of G (a square might degenerate to a point, as in Figure 7).



(a) (b)

Figure 7: A triangulation and the associated square packing. Thanks to David Wilson for providing this figure from [S5]

The packing is actually a tiling: Indeed, Oded points out the following simple observation.

Let P_a, P_b, P_c be three rectangles whose edges are parallel to the coordinate axis. Suppose that the intersection of every two of these rectangles is nonempty. Then $P_a \cap P_b \cap P_c \neq \emptyset$.

The same tiling theorem was obtained independently by Cannon, Floyd and Parry in [CFP1]. Both employ Cannon's discrete extremal length (see Section 2.4.1) and obtain the side lengths $s(v)$ of the squares as the weights $\rho(v)$ of the extremal metric (corresponding to the family of "combinatorial curves" joining two opposite sides of the quadrilateral). It is quite different from the classical square packings of Brooks, Smith, Stone and Tutte [BSST], in particular, since the metrics considered here live on the vertices rather than the edges of the graph.

A very similar idea is exploited in the important paper [S6]. The classical setting of extremal length (recall (1)) is a family Γ of curves contained in a domain Ω. Invariance

$$\lambda(f(\Gamma)) = \lambda(\Gamma)$$

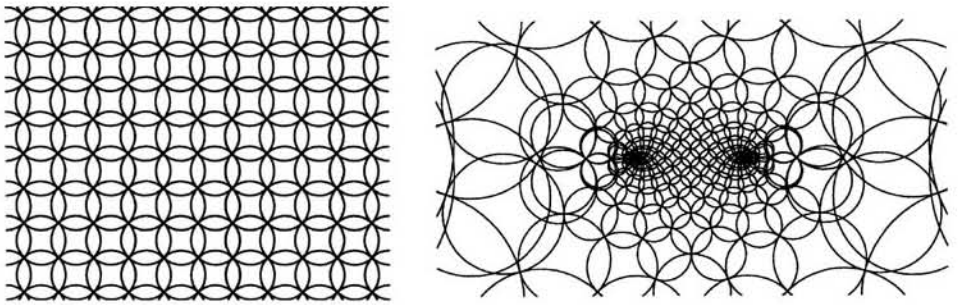
under conformal maps of Ω is almost trivial (just pull back metrics from $f(\Omega)$). Oded's notion of *transboundary extremal length* $\lambda_\Omega(\Gamma)$ applies to curve families Γ that are not necessarily contained in Ω. The metrics are now replaced by generalized metrics ρ that, roughly speaking, also assign length to complementary components. The length $\int_\gamma \rho|dz|$ is replaced by $\int_{\gamma\cap\Omega} \rho|dz| + \sum_p \rho(p)$ if γ is not contained in Ω, where the sum is over all boundary components of Ω that γ meets. Then the definition is $\lambda_\Omega(\Gamma) = \sup_\rho(\inf_\gamma \int_{\gamma\cap\Omega} \rho|dz| + \sum_p \rho(p))/\int_{\mathbb{C}} \rho^2 dxdy$, and conformal invariance is again immediate. Using this innocent looking extension, Oded provides an elegant self-contained proof of the countable Koebe conjecture, and moreover is able to deal with the case of domains for which the complementary component satisfy a certain fatness condition ($\text{area}(A \cap B(x,r)) \geq cr^2$ for each component A, each $x \in A$ and each disc $B(x,r)$ that does not contain A).

Circle packings corresponding to infinite graphs G can be obtained by taking Hausdorff limits of packings corresponding to finite subgraphs, but where do they "live"? Beardon and Stephenson [BSt1],[BSt2] have shown, under the assumption that the degrees of the vertices are uniformly bounded, that the carrier of such a packing is either the plane (call this case *parabolic*), or that it can be chosen to be the disc (*hyperbolic*). They also showed that both cases are mutually exclusive, and that the packing is hyperbolic if each degree is at least seven. The uniform boundedness assumption was later removed by He and Schramm [HS1], and they proved in general that the *type* of a packing is unique (that is, there is no infinite graph that packs both the disc and the plane). In the impressive paper [HS3], they characterize the type in terms of the discrete extremal length, and use it to show that the packing is parabolic if simple random walk on G is recurrent. They conclude (Theorem 10.1) that a packing is parabolic if at most finitely many vertices have degree greater than 6 (notice that every vertex of the hexagonal packing has degree 6). This paper contains their earlier result [HS3b] that a packing is hyperbolic if the lower average degree is greater than 6. By definition, the lower average degree is

$$\text{lav}(G) = \sup_{W_0} \inf_{W \supset W_0} \frac{1}{|W|} \sum_{v \in W} \deg(v).$$

In the case that the degrees of the vertices are uniformly bounded, they also show that transience implies hyperbolicity. Jointly with Itai Benjamini, this line of investigation was carried further in [BS1] and [BS2], by applying circle- and square packings to constructions of harmonic functions on graphs. Another nice application of circle packings is the recurrence of (weak) limits of random planar graphs with bounded degree, [BS2].

I have always admired Oded's ability to find a good modification of a difficult problem that turns it into a tractable problem while keeping its essential features. One of the many examples is his work on discrete analytic function [S8]. Since circle packings can be viewed as discrete analogs of conformal maps, it is natural to ask for the analogs of analytic functions, thus giving up injectivity (disjointness of the discs). See [Ste2] for the state of the art and beautiful illustrations. Peter Doyle described collections of discs that are tangent according to the hexagonal pattern that are analogs of the exponential function. He conjectured that these would be the only "entire" circle packing immersions. While Oded was not able to resolve this conjecture, he did find that collections of overlapping discs based on the square grid seem better suited for the problem, and constructed the analog of the error function $\int e^{-z^2} dz$ in this setting. Along the way, he introduced Möbius invariants that are discrete analogs of the Schwarzian derivative and became instrumental in his later work [HS8].



(a) (b)

Figure 8: The square grid and the $\sqrt{i}$SG erf pattern, from [S8]

3 The Schramm-Loewner Evolution

There are several excellent lecture notes, overview articles, and a textbook on SLE [La3], mostly by and for probabilists or theoretical physicists, see [BB3], [Car2], [Dup], [GK], [KN], [S11], [W1], [W2] and the references therein. It is not my intention to provide another streamlined introduction to the area. Instead, I would like to give a somewhat historic account with an emphasis on Oded's contributions, highlighting some of the mathematical challenges he faced.

3.1 Pre-history

It is perhaps appropriate to very briefly describe the state of knowledge related to conformally invariant scaling limits prior to Oded's discovery of SLE, and to describe some of the results that were instrumental in his work. Oded's own historical narrative is Section 1.2 in [S11].

Two-dimensional lattice processes such as the Self-Avoiding Walk (SAW), the Ising model, percolation, and diffusion limited aggregation (DLA), to name just a few, have been intensively studied by physicists and by probabilists for a long time. See Figure 21 for some pictures, and the aforementioned articles for descriptions of the models. In the physics community, many problems such as finding the Hausdorff dimension of scaling limits of these sets were considered well-understood. The implicit assumption of conformal invariance of the scaling limit allowed the use of the powerful machinery of conformal field theory and led to results such as Cardy's formula for the crossing probability of critical percolation [Car1]. On the mathematical side, progress was much slower, one of the hurdles being that in most cases the existence of a scaling limit was unknown. Even finding suitable definitions of the concept of scaling limit was a nontrivial task.

In the late 1980's, Christian Pommerenke told me how compositions of (random) conformal maps onto slitted discs could be viewed as a variant of the Witten-Sander model for DLA [WS]. At the same time, Richard Rochberg and his son David were working on this setup. It seems that the only trace of this is a talk given by Rochberg at the March 1990 AMS Regional meeting in Manhattan, Kansas, titled "Stochastic Loewner Equation". Their model is similar to an approach

to Laplacian growth proposed by Hastings and Levitov [HL], and is quite different from what is now called Stochastic Loewner Evolution or Schramm-Loewner Evolution SLE. At that time, other analysts such as Lennart Carleson, Peter Jones and Nick Makarov worked with similar ideas, see e.g. [CM]. Oded was at best dimly aware of these activities, and was not really interested in stochastic processes such as DLA until much later.

Greg Lawler's invention [La] of the Loop Erased Random Walk (LERW) provided the mathematics community with a process that shared some features with the Self-Avoiding Walk, but at the same time was more tractable, partly due to its Markovian property. Pemantle's work [**?**] and Wilson's algorithm provided a link between Uniform Spanning Trees (UST) and LERW's. Intensive research on the UST [Ly] culminated in the paper [BLPS] by Benjamini, Lyons, Peres and Schramm. The deep work of Rick Kenyon [Ke1] combined powerful combinatorics and discrete complex analysis and exhibited conformal invariance properties of the LERW. He was also able to determine its expected length.

3.2 Definition of SLE

Oded told me in 1997 about his idea to exploit conformal invariance in order to study the LERW. The streamlined way to present SLE in courses or texts, beginning with a crashcourse on the Loewner equation followed by a crash course on stochastic calculus (or the other way round) is, of course, not quite representative of its emergence. In a 2006 email exchange with Yuval Peres and myself about the history of SLE, Oded wrote:

Up to the time when I started thinking about SLE, I did not really know what Loewner's equation was, or what was the idea behind it, though I did know that it was a tool which was important for the coefficients problem and that it involved slit mappings and a differentiation in the space of conformal maps. I kind of rediscovered it in the context of SLE and then made the connection.

3.2.1 The (radial) Loewner equation

Loewner ([Lo]; see also [Dur],[Po2] or [La3]) introduced his differential equation as a tool in his attempt to prove the Bieberbach conjecture $|a_n| \leq n$ concerning the Taylor coefficients of normalized conformal maps $f(z) = z + \sum_{n=2}^{\infty} a_n z^n$ of the unit disc. It was also instrumental in the final solution by de Branges in 1984.

Let γ be a simple path that is contained in $\mathbb{D}$ except for one endpoint on $\partial\mathbb{D}$. More precisely, let $\gamma : [0, \tau] \to \overline{\mathbb{D}}$ be continuous and injective with $\gamma(0) \in \partial\mathbb{D}$ and $\gamma(\tau) = 0$, such that $\gamma(0, \tau] \subset \mathbb{D}$. Denote $G_t = \mathbb{D} \setminus \gamma[0, t]$ so that $G_0 = \mathbb{D}$. Then, for each $0 \leq t < \tau$, there is a unique conformal map $g_t : G_t \to \mathbb{D}$ that is normalized by $g_t(0) = 0$ and $g_t'(0) > 0$. By Schwarz' Lemma, $g_t'(0)$ strictly increases, $g_0'(0) = 1$, and it is not hard to see that $g_t'(0) \to \infty$ as $t \to \tau$. Hence we can reparametrize γ so that $\tau = \infty$ and $g_t'(0) = e^t$. Loewner's theorem says that

$$\frac{\partial}{\partial t} g_t(z) = g_t(z) \frac{\zeta_t + g_t(z)}{\zeta_t - g_t(z)} \tag{3}$$

for all $t \geq 0$ and all $z \in G_t$, where the "driving term"

$$\zeta_t = g_t(\gamma(t)) \in \partial\mathbb{D}$$

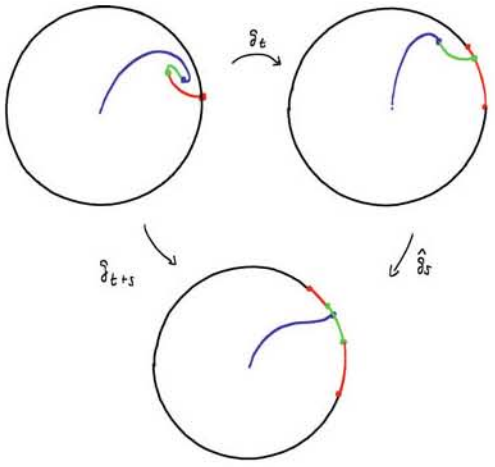


Figure 9: Conformal maps from slit discs onto discs.

is continuous (a priori, g_t is only defined in G_t, but it can be shown that g_t extends to $\gamma(t)$).

A simple but crucial observation is that the driving term ζ^T of the curve $\gamma^T = g_T(\gamma)$ (more precisely, the parametrized curve $\gamma^T(s) := g_T(\gamma(T+s))$) is given by $\zeta^T_s = \zeta_{T+s}$. Thus "conformally pulling down" a portion of γ corresponds to shifting the driving term. Intuitively, one can think of the Loewner equation as describing a conformal map to a slitted disc as a composition of conformal maps onto infinitesimally slitted discs with slit at ζ_t, plus the statement that the conformal map onto such a disc is $z \mapsto z + z\frac{\zeta_t+z}{\zeta_t-z}\Delta t$ up to first order in Δt.

Thus the Loewner equation associates with each simple curve $\gamma \subset \mathbb{D}$ a continuous function ζ_t with values in $\partial\mathbb{D}$. Conversely, it is not hard to show that the solution $g_t(z)$ to the initial value problem (3), $g_0(z) = z$, forms a family of conformal maps of simply connected domains G_t onto $\mathbb{D}$. In fact, G_t is the set of those points $\zeta \in \mathbb{D}$ for which the solution is well-defined on the interval $[0, t]$. It easily follows that G_t increases in t, and that $z \in G_t$ unless $g_s(z) = \zeta_s$ for some $s \leq t$. The complement

$$K_t = \mathbb{D} \setminus G_t$$

is called the *hull* of ζ. In our original setup of a slit disc, we simply recover the curve, $K_t = \gamma[0, t]$. It has been known since Kufarev [Ku] that smooth functions ζ generate smooth curves γ, but that there also exist continuous functions ζ for which the associated hull is not a simple arc in $\mathbb{D}$. Kufarev's example simply is the computation that a circular chord γ of the unit circle has continuous driving term. In fact, it can be topologically wild (not locally connected), see [MR]. My own interest in the Loewner equation originated when Oded asked me which driving terms generate curves.

3.2.2 The scaling limit of LERW

The LERW is obtained from simple random walk by erasing loops chronologically. The main result

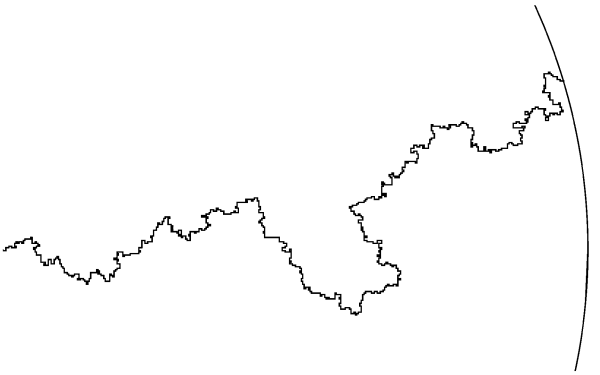


Figure 10: A loop erased random walk in the disc, from [S9].

of Oded's celebrated paper [S9] was a conditional theorem: Assuming the existence and conformal invariance of the scaling limit of LERW, he showed that the Loewner driving term of the resulting (random) limiting curve is a Brownian motion on the unit circle, $\zeta_t = e^{iB_{2t}}$. To make this rigorous, he first gave the following definition of the notion of scaling limit: Let $D \subsetneq \mathbb{C}$ be a domain, fix $a \in D$, and for $\delta > 0$, consider the LERW on the graph $\delta\mathbb{Z}^2 \cap D$, started at a point closest to a and stopped when reaching ∂D. Viewing the path of the LERW as a random subset of the sphere $S^2 = \mathbb{C} \cup \{\infty\}$, its distribution is a discrete measure μ_δ on the space of compact subsets of S^2. Equipped with the Hausdorff distance, the space of compact subsets of S^2 is a compact metric space, and so is the space of its Borel measures. The existence of subsequential weak limits $\mu = \lim_j \mu_{\delta_j}$ follows at once. If the limit measure $\mu = \lim_{\delta \to 0} \mu_\delta$ exists, it is called the scaling limit of LERW from a to ∂D.

Theorem 3.1 ([S9], Theorem 1.1)**.** *If each connected component of ∂D has positive diameter, then every subsequential scaling limit measure μ of the LERW from a to ∂D is supported on simple paths.*

In other words, the measure of the set of non-simple curves is zero. This theorem is interesting in its own right. It has been known previously that, loosely speaking and under mild assumptions, random curves have uniform continuity properties that imply their (subsequential) scaling limits to be supported on continuous curves [AB]. However, the fact that the loop erased paths are *simple* curves does not directly imply that the limiting objects have no loops. Indeed, the limits of other discrete random simple curves such as the critical percolation interface or the uniform spanning tree Peano path are not simple. The proof uses estimates for the probability distribution of "bottlenecks", based on harmonic measure estimates and Wilson's algorithm, and a topological characterization of simple curves.

Next, Oded formulated the conjecture of existence and conformal invariance of the scaling limit as follows.

Conjecture *([S9],1.2) Let $D \subsetneq \mathbb{C}$ be a simply connected domain in $\mathbb{C}$, and let $a \in D$. Then the scaling limit of LERW from a to ∂D exists. Moreover, suppose that $f : D \to D'$ is a conformal homeomorphism onto a domain $D' \subset \mathbb{C}$. Then $f_* \mu_{a,D} = \mu_{f(a),D'}$, where $\mu_{a,D}$ is the scaling limit measure of LERW from a to ∂D, and $\mu_{f(a),D'}$ is the scaling limit measure of LERW from $f(a)$ to $\partial D'$.*

The most important and exciting result of [S9] was the insight that this conjecture implied an explicit construction of the limit in terms of the Loewner equation. By Theorem 3.1, the conjectural scaling limit μ induces a measure on the space of continuous real-valued functions $\hat{\zeta}_t$ via the correspondence $\gamma \mapsto \zeta = e^{i\hat{\zeta}}$ of the Loewner equation. Oded showed that the law of $\hat{\zeta}$ is that of a time-changed Brownian motion, B_{2t}:

Theorem 3.2 ([S9], Theorem 1.3)**.** *Assuming the above conjecture, the scaling limit μ is equal to the law of the hulls K associated with the driving term $\zeta = e^{iB_{2t}}$, where $B_t, t \geq 0$ is a Brownian motion started at a uniform random point in $[0, 2\pi)$.*

In his characteristic way, Oded pointed out the simple idea behind the theorem. From his paper:

At the heart of the proof of Theorem 3.2 lies the following simple combinatorial fact about LERW. Conditioned on a subarc β' of the LERW β from 0 to ∂D, which extends from some point $q \in \beta$ to ∂D, the distribution of $\beta \setminus \beta'$ is the same as that of LERW from 0 to $\partial(D - \beta')$, conditioned to hit q. When we take the scaling limit of this property, and apply the conformal map from $D - \beta'$ to $\mathbb{D}$, this translates into the Markov property and stationarity of the associated Löwner parameter ζ.

He also notes that "the passage to the scaling limit is quite delicate". The translation into the Markov property and stationarity is by means of the aforementioned principle that "conformally pulling down" a portion γ' of γ corresponds to shifting the driving term. Thus $\hat{\zeta}$ is a continuous process with stationary and independent increments. Now the theory of Levy processes (and the symmetry of LERW under reflection) implies that $\hat{\zeta}_t$ has the law of $\sqrt{\kappa} B_t$ for some $\kappa > 0$ and a standard Brownian motion B. It remained to determine the constant κ. To this end, Oded gives the following

Definition. *The (radial) stochastic Loewner evolution SLE_κ with parameter $\kappa > 0$ is the random process of conformal maps g_t generated by the Loewner equation driven by $\zeta_t = e^{i\sqrt{\kappa} B_t}$.*

In Section 7 of [S9] he actually defined SLE as a process of random paths generated by the Loewner equation, and therefore had to restrict to those values of κ for which the resulting hulls are simple curves; he conjectured that this is the interval $[0, 4]$.

Next, he analyzed the winding number of the SLE-path around 0, and of the LERW: If $\theta_\kappa(t) = \arg \gamma(t)$, $t \geq 0$, denotes the continuously defined argument along the curve, then he computed the variance

$$E[\theta_\kappa(t)^2] = (\kappa + o(1)) \log t.$$

It follows that the winding number of the portion of the SLE path until its first hitting of the circle of radius ε centered at 0 has variance $(\kappa + o(1)) \log(1/\varepsilon)$. On the other hand, Kenyon's work [Ke2] implies that the variance of the winding number of LERW in $\mathbb{D} \cap \varepsilon\mathbb{Z}^2$ is $(2 + o(1)) \log(1/\varepsilon)$, and after some work the conclusion $\kappa = 2$ follows.

Oded went on to compute what he called the critical value for SLE: He proved that for $\kappa > 4$, almost surely SLE will *not* generate simple paths, and conjectured that it will for $\kappa \in [0, 4]$. See the discussion in Section 3.3.3 for the simple proof using Ito's formula (in the chordal case). However, Oded was a self-taught newcomer to stochastic calculus, discovered some of the basics himself, and commented on his proof in an email in January 1999:

This must all be quite standard, to people with the right background. But not for me.

3.2.3 The chordal Loewner equation, percolation, and the UST

The classical (radial) Loewner equation is well-suited for curves that join an *interior point* to a *boundary point*, such as curves generated by the LERW. Other processes generate curves joining *two boundary points*. Oded realized how important the "correct" normalizations are in dealing with conformal maps and in particular the Loewner equation, and found the appropriate variant of the Loewner equation (it turned out later that this version has been described earlier, beginning with N.V. Popova [Pop1],[Pop2]; I would like to thank Alexander Vasiliev for this reference). He describes this in another email in January 1999:

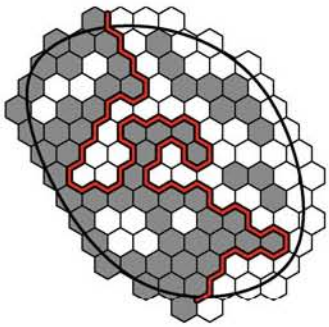


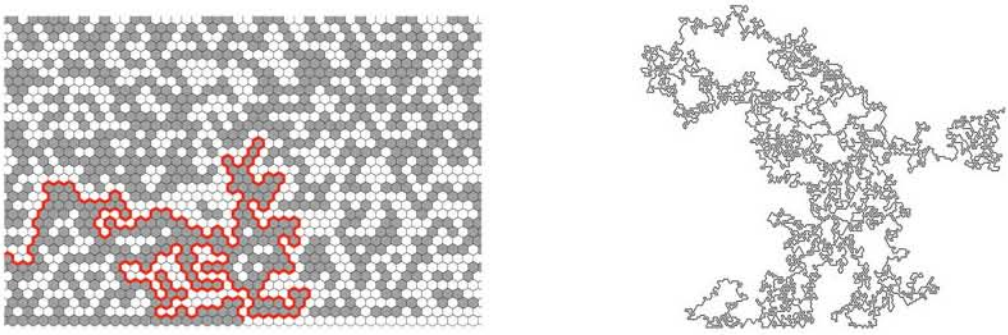
Figure 11: Percolation in a domain, from [S11].

I have a mathematical querry. Before the question itself, here's the motivation. For the LERW scaling limit, the natural object is a probability measure on the set of paths from a point in the domain to the boundary. In other settings, the natural object is a probability measure joining two points on the boundary of a domain. Consider, for example, percolation in the unit disk. Let (γ, β) be a partition of the boundary of the disk to two arcs, disjoint except for the endpoints. Let K be the union of all percolation clusters inside the disk that are connected to γ. Then the outer boundary of K is a path, α, joining the two endpoints of γ. The scaling limit of α is conjecturally conformally invariant (but not a simple path). Assuming conformal invariance, I'm optimistic that the scaling limit can be represented by a Loewner-like Brownian evolution. The first step for this seems to be the following variation on Loewner's theorem:

Thm: *Let $\alpha : [0, \infty) \to \mathbb{C}$ be a continuous simple path such that $\alpha(0) = 0, \lim_{t\to\infty} \alpha(t) = \infty$ and $\mathrm{Im}(\alpha(t)) > 0$ when $t > 0$. Let f_t be the conformal map from the upper half plane to the upper half*

plane minus $\{\alpha(s) : 0 \le s \le t\}$*, which is normalized by* $f_t(z) = z + O(1/z)$ *near infinity. By change of parameterization of* α *(and perhaps changing its interval of definition), we may assume that* $f_t(z) = z + t/z + O(1)/z^2$ *near infinity. Set* $g(w,t) = f_t^{-1}(w)$*. Then* $\{\partial g/\partial t\} = 1/(g - k(t))$*, where* $k(t) = g(\alpha(t), t)$*.*

In the situation of percolation and related conformal invariance models, one should expect $k(t) = cBM(t)$*, where BM is on the real line. Have you seen this theorem? The proof should not be difficult. It can either be derived from Loewner's theorem, or by adapting the proof.*



(a) (b)

Figure 12: The percolation exploration path

In order to coincide with the normalization of the radial Loewner equation, he later slightly adjusted the parametrization of the path α so that

$$f_t(z) = z + 2t/z + O(1)/z^2$$

and

$$\frac{\partial}{\partial t} g_t(z) = \frac{2}{g_t(z) - W_t}. \tag{4}$$

With this normalization and assuming conformal invariance of the percolation scaling limit, he showed that the (non-simple) limit curves would satisfy the chordal Loewner equation (4) with $W_t = \sqrt{6}B_t$. The value 6 can be found as the only κ such that the random sets generated by $\sqrt{\kappa}B_t$ satisfy Cardy's formula, or the locality property discussed below. This led to the definition of chordal SLE_κ as the random process of conformal maps

$$g_t : \mathbb{H} \setminus K_t \to \mathbb{H}$$

generated by the Loewner equation (4) with driving function $W(t) = \sqrt{\kappa}B_t$, where B is a standard Brownian motion.

The *hull* K_t is the set of those points z for which $g_s(z) = W_s$ for some $s \le t$ so that (4) becomes undefined. Since g_t is determined by K_t, one has the equivalent

Definition: Chordal SLE_κ is the process of random hulls $(K_t, t \geq 0)$ generated by the Loewner equation (4) with $W_t = \sqrt{\kappa} B_t$.

In the same paper, Oded also defines and analyzes subsequential scaling limits of the uniform spanning tree. He ends the paper by speculating (that is, stating without giving detailed proofs) about the Loewner driving term of the UST Peano curve in the upper half plane $\mathbb{H}$. He finds that, again assuming existence and conformal invariance of the limit, this random space filling curve is SLE_8. At the end of the introduction, he summarizes the findings of the paper as follows:

The emerging picture is that different values of κ in the differential equation (3) *or* (4) *produce paths which are scaling limits of naturally defined processes, and that these paths can be space-filling, or simple paths, or neither, depending on the parameter κ.*

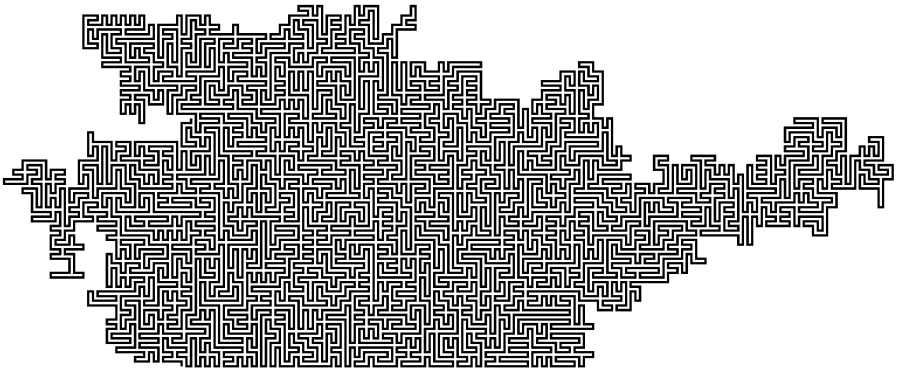


Figure 13: The UST Peano curve, from [RoS].

3.3 Properties and applications of SLE

Two exciting developments took place shortly after the introduction of SLE, namely the very productive collaboration of Greg Lawler, Oded Schramm and Wendelin Werner, and the surprising proof of existence and conformal invariance of the percolation scaling limit by Stas Smirnov. I will begin describing the former, and defer the latter to Section 3.3.4.

3.3.1 Locality

In the important papers [LW1] and [LW2], Lawler and Werner discovered that Brownian excursions have a certain restriction property (explained below), and that intersection exponents of conformally invariant processes with this property are closely related to those of Brownian motion. What was missing was a way to compute exponents of *some* conformally invariant process. Lawler, Schramm and Werner discovered that SLE provided such a process to which the universality arguments of Lawler and Werner [LW2] applied. The following email from Oded describes the crucial property:

I don't remember if I've mentioned to you the restriction property for SLE(6) that Greg, Wendelin and I have proved. It says that (up to time parameterization) the law of SLE(6) in an arbitrary

domain D starting from a point p on the boundary and stopped when it exits a small ball B around p, does not depend on the shape of the domain outside B (provided that D-B is connected, say). Thus, SLE(6) is purely a local process, like BM. This is not true for $\kappa \neq 6$. For example, SLE(6) in the disk (with Loewner's original equation) is the same as SLE(6) in the half plane, with my variation on Loewner's equation. This seems to say that SLE(6) is a very special process.

The above property is trivial for the discrete critical percolation exploration path, since the path can be grown "dynamically" by deciding the color of a hexagon only when the path meets it and needs to decide whether to turn "right" or "left". Hence, in light of the conjectured scaling limit, locality for chordal SLE_6 was not unexpected. The coincidence of chordal and radial SLE_6, discussed below, was more surprising.

Here is a precise statement of the locality property. Let $D = \mathbb{H} \setminus A$ be a simply connected subdomain of $\mathbb{H}$ such that A is bounded and also bounded away from 0. Denote g_A the conformal map from D onto $\mathbb{H}$ with the hydrodynamic normalization ($g_A(z) - z \to 0$ as $z \to \infty$), and set $\Phi_A = g_A - g_A(0)$. Then SLE in D from 0 to ∞ is defined as the preimage of SLE in $\mathbb{H}$ from 0 to ∞ under Φ_A. The following expresses the fact that SLE in D is a time-change of SLE in $\mathbb{H}$, up to the time that the process hits A. Let $T = \inf\{t : K_t \cap A \neq \emptyset\}$ and $\tilde{T} = \inf\{t : K_t \cap \Phi_A(\partial A) \neq \emptyset\}$.

Theorem 3.3 ([LSW2], Theorem 2.2)**.** *For $\kappa = 6$, the processes $(\Phi_A(K_t), t < T)$ and $(K_t, t < \tilde{T})$ have the same law, up to re-parametrization of time.*

The equivalence of chordal and radial SLE_6 was established in [LSW3], Theorem 4.1: Define the hulls K_t of chordal SLE_κ in $\mathbb{D}$ from 1 to -1 as the image of chordal SLE_κ in $\mathbb{H}$, under the conformal map $f(z) = (i - z)/(i + z)$. Thus K_t are hulls growing from 1 towards -1 in $\overline{\mathbb{D}}$. Denote $T \leq \infty$ the first time when K_t contains 0. Similarly, denote $\widetilde{K}_t$ the chordal SLE_κ hulls in $\mathbb{D}$, started at 1, and denote $\widetilde{T}$ the first time when $\widetilde{K}_t$ contains -1. Then, for $\kappa = 6$, the laws of $(K_t, t < T)$ and $(\widetilde{K}_t, t < \widetilde{T})$ are the same, up to a random time change $s = s(t)$. The proof and Girsanov's theorem also show that for all values of κ, the laws of K_t and $\widetilde{K}_{s(t)}$ are equivalent (in the sense of absolute continuity of measures), if t and $\widetilde{t}$ are bounded away from T and $\widetilde{T}$. See Proposition 4.2 in [LSW3].

The first proof of the locality Theorem 3.3 was rather long and technical, based on an analysis of the Loewner driving function of a curve under continuous deformation of the surrounding domain. A different and simpler proof was found later (see Proposition 5.1 in [LSW8]), by analyzing $\tilde{W}_t := h_t(W_t)$, where $h_t := \tilde{g}_t \circ g_A \circ g_t^{-1}$ and $\tilde{g}_t = g_{g_t(D \setminus K_t)}$ is the normalized conformal map of $g_t(D \setminus K_t)$ to $\mathbb{H}$. Writing

$$\tilde{g}_t(z) = z + \frac{a_t}{z} + O(\frac{1}{z^2}),$$

computation shows that

$$\partial_t \tilde{g}_t(z) = \frac{\partial_t a_t}{\tilde{g}_t(z) - \tilde{W}_t} = \frac{2h_t'(W_t)^2}{\tilde{g}_t(z) - \tilde{W}_t}.$$

With $W_t = \sqrt{\kappa} B_t$, computation using Ito's formula shows

$$d\tilde{W}_t = h_t'(W_t) dW_t + \big((\kappa/2) - 3\big) h_t''(W_t) dt. \tag{5}$$

Thus $\tilde{W}_t$ is a local martingale if (and only if) $\kappa = 6$, and a time change shows that $(\tilde{g}_t, t \geq 0)$ is SLE_6.

3.3.2 Intersection exponents and dimensions

The locality of SLE_6 has been used to determine the so-called *intersection exponents* of 2-dimensional Brownian motion, and to compute the Hausdorff dimensions of various sets associated with its trace. These results established Schramm's SLE and the Lawler-Werner universality arguments as a fundamental and powerful new tool.

If B_t^1 and B_t^2 are two independent planar Brownian motions started at two different points $B_0^1 \neq B_0^2$, it easily follows from the subadditivity of $t \mapsto \log \mathbb{P}[B^1[0,t] \cap B^2[0,t] = \emptyset]$ that there is a number $\zeta > 0$ such that

$$\mathbb{P}[B^1[0,t] \cap B^2[0,t] = \emptyset] = \left(\frac{1}{t}\right)^{\zeta + o(1)}.$$

Similarly, the half-plane exponent $\tilde{\zeta}$ of the event that two independent motions do not intersect and stay in a halfplane is given by

$$\mathbb{P}[B^1[0,t] \cap B^2[0,t] = \emptyset \text{ and } B^j[0,t] \subset \mathbb{H}, j = 1,2] = \left(\frac{1}{t}\right)^{\tilde{\zeta} + o(1)}.$$

More generally, one considers exponents ζ_p for the probability of the event that p independent motions are mutually disjoint, $\zeta(j,k)$ for the event that two packs of Brownian motions $B^1 \cup \cdots \cup B^j$ and $B^{j+1} \cup \cdots \cup B^{j+k}$ are disjoint, and the corresponding half-plane exponents $\tilde{\zeta}_p$ and $\tilde{\zeta}(j,k)$. So $\zeta = \zeta(1,1)$. Also relevant is the *disconnection exponent* $2\eta_j$ for the event that the union of j Brownian motions, started at 1, does not disconnect 0 from ∞ before time t.

These and other intersection exponents have been studied intensively, and values such as $\zeta = 5/8$ had been obtained by Duplantier and Kwon [DK] using the mathematically non-rigorous method of conformal field theory.

An extension of $\zeta(j,k)$ for positive real $k > 0$ was given in [LW1], and some fundamental properties (in particular the "cascade relations") were established. In the series of papers [LSW2],[LSW3],[LSW5],[LSW7] (see [LSW1] for a guide and sketches of proofs), Lawler, Schramm and Werner were able to confirm the predictions, and they proved

Theorem 3.4. *For all integers $j \geq 1$ and all real numbers $k \geq 0$,*

$$\zeta(j,k) = \frac{\sqrt{24j+1} + \sqrt{24k+1} - 2)^2 - 4}{96}, \quad \zeta_n = \frac{4n^2 - 1}{48},$$

$$\tilde{\zeta}(j,k) = \frac{\sqrt{24j+1} + \sqrt{24k+1} - 1)^2 - 1}{48}, \quad \tilde{\zeta}_n = \frac{2n^2 + n}{6},$$

and

$$\eta_k = \zeta(k,0) = \frac{(\sqrt{24k+1} - 1)^2 - 4}{48}.$$

In particular,

$$\zeta = \frac{5}{8}, \tilde{\zeta} = \frac{5}{3}, \eta_1 = \frac{1}{4}, \eta_2 = \frac{2}{3}.$$

The proofs are technical masterpieces combining a variety of different methods. A very rough description is as follows: First, half-plane intersection exponents of SLE_6 are computed, based on estimates for the crossing probability of (long) rectangles. This is done by establishing a version of Cardy's formula. Then, the universality ideas of [LW2] are employed to pass from SLE_6 to Brownian motion. Finally, to cover the case $k < 1$, real analyticity of the exponent is shown by recognizing $e^{-2\zeta(j,k)}$ as the leading eigenvalue of an operator T_k on a space of functions on pairs of paths.

For a fixed time t, the *Brownian frontier* is the boundary of the unbounded connected component of the complement of $B[0,t]$, and the set of *cut points* is the set of those points p for which $B[0,t]\backslash\{p\}$ is disconnected. The set of *pioneer points* is the union of the frontiers over all $t > 0$. Mandelbrot [M] observed that the Brownian frontier looks like a long self-avoiding walk. Since the Hausdorff dimension of the self-avoiding walk was predicted by physicists to have Hausdorff dimension 4/3, he conjectured that the Hausdorff dimension of the Brownian frontier is 4/3. Greg Lawler had shown in a series of papers (see [La2]) how the intersection exponents are related to the Hausdorff dimension of subsets of the Brownian path. He found the values $2-2\zeta$, $2-\eta_2$, and $2-\eta_1$ for the dimension of the Brownian frontier, the set of cut points, and the set of pioneer points. This actually required his stronger estimates of the intersection probabilites up to constant factors, rather than up to $(1/t)^{o(1)}$. Simpler proofs of those estimates are the content of [LSW6]. In combination with Theorem 3.4, this proved Mandelbrot's conjecture.

Theorem 3.5 ([LSW3], [LSW7]). *The Hausdorff dimension of the frontier, the set of cut points, and the set of pioneer points of 2-dimensional Brownian motion is* $4/3, 3/4$ *and* $7/4$ *almost surely.*

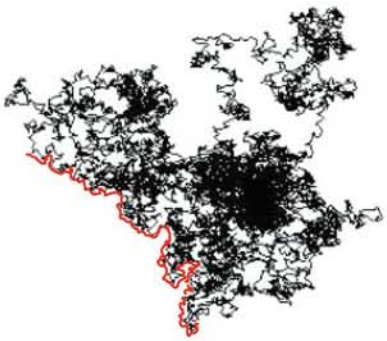


Figure 14: A Brownian path, with part of the frontier highlighted

To put this result in perspective, notice that it is rather difficult to show even that the dimension of the Brownian frontier is more than one [BJPP], and that the set of cutpoints is non-empty, [Bu]. It should also be mentioned that by work of Lawler, the intersection exponents for simple random walk are the same as for Brownian motion, so that Theorem 3.4 also shows, for instance, that

$$\mathbb{P}[S^1[0,n] \cap S^2[0,n] = \emptyset] = \left(\frac{1}{n}\right)^{\frac{5}{8}+o(1)}$$

if S_1 and S_2 are two independent planar simple random walks started at different points.

Meanwhile, there is a more elegant approach to these dimension results, also due to Lawler, Schramm and Werner, see Section 3.3.5

3.3.3 Path properties

I have been fortunate to collaborate with Oded on several projects over the past two decades. Sometimes this meant just trying to catch up with his fast output, and watching in awe how one clever idea replaced another. But perhaps even more impressively, Oded had an amazing ability and willingness to listen, and to think along. Sometimes, when I failed in an attempt to articulate a vague idea and was about to give up a faint line of thought, he surprised me by completely understanding what I tried to express, and by continuing the thought, almost like mind reading.

The definition of SLE_κ as a family of conformal maps g_t through a stochastic differential equation does not shed much light upon the structure of the hulls $K_t = g_t^{-1}(\mathbb{H})$. By Section 3.3.1, it is enough to consider chordal SLE. We say that the hull $(K_t)_{t\geq 0}$ is *generated by a curve* γ if $\gamma : [0,\infty) \to \overline{\mathbb{H}}$ is continuous and if K_t is obtained from γ by "filling in the holes" of $\gamma[0,t]$ (more precisely, K_t is the complement in $\mathbb{H}$ of the unbounded connected component of $\mathbb{H} \setminus \gamma[0,t]$). Since continuity of the driving term W_t is equivalent to the requirement that the increments $K_{t+\varepsilon} \setminus K_t$ have "small diameter within D_t" (more precisely, there is a set $S \subset D_t$ of small diameter that disconnects $K_{t+\varepsilon} \setminus K_t$ from ∞ within D_t, see [LSW2], Theorem 2.6), such a curve cannot cross itself, but it can have double points and "bounce off" itself ([Po1]). There are examples of continuous W for which K_t is not locally connected, and such sets cannot be generated by curves [MR]. Fortunately, this does not happen for SLE:

Theorem 3.6 ([RoS], Theorem 5.1; [LSW9], Theorem 4.7)**.** *For each $\kappa > 0$, the hulls K_t are generated by a curve, almost surely.*

It follows that, a.s., the conformal maps $f_t = g_t^{-1}$ extend continuously to the closed half space $\overline{\mathbb{H}}$, and $\gamma(t) = f_t(W_t)$. For $\kappa \neq 8$, the proof hinges on estimates for the derivative expectations $\mathbb{E}[|f_t'(z)|^p]$. For $\kappa = 8$, the only known proof is by exploiting the fact that SLE_8 is the scaling limit of UST, and that the UST scaling limit is a continuous curve a.s., [LSW9].

As Oded already noticed in [S9], the SLE_κ trace has different phases, depending on the value of κ.

Theorem 3.7 ([RoS])**.** *For $\kappa \leq 4$, the SLE trace γ is a simple curve in $\mathbb{H} \cup \{0\}$, almost surely. It "swallows" points (for fixed $z \in \overline{\mathbb{H}} \setminus \{0\}$, a.s. $z \in K_t$ for large t, but $z \notin \gamma[0,\infty)$) if $4 < \kappa < 8$, and it is space-filling ($\gamma[0,\infty) = \overline{\mathbb{H}}$) if $\kappa \geq 8$. For all κ, the trace is transient a.s.: $|\gamma(t)| \to \infty$ as $t \to \infty$.*

Let us explain the phase transition at $\kappa = 4$, already observed and conjectured in [S9] (in the radial case). Let $x > 0$ and denote

$$X_t = g_t(x) - W_t,$$

where $W_t = \sqrt{\kappa} B_t$. Then

$$dX_t = \frac{2}{X_t} dt - \sqrt{\kappa} dB_t$$

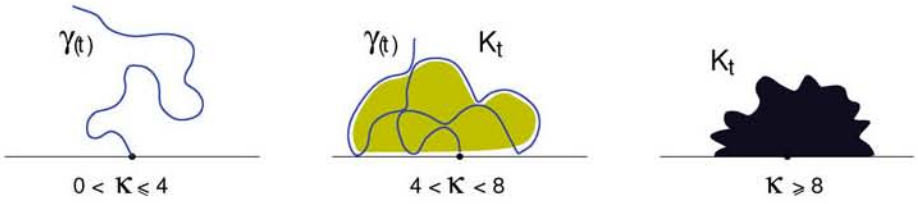


Figure 15: The three phases of SLE; picture courtesy of Michel Bauer and Denis Bernard [BB1]

is an Ito diffusion and can be easily analyzed using stochastic calculus. In fact, X_t is a Bessel process of dimension $1 + 4/\kappa$. Thus $X_t > 0$ for all t if and only if $\kappa \leq 4$. In this range, we obtain $K_t \cap \mathbb{R} = \{0\}$ for all t, and it is an easy consequence of Theorem 3.6 that γ is simple (if $r < s < t$ are such that $\gamma(r) = \gamma(t) \neq \gamma(s)$, then the curve $g_s(\gamma[s,t])$ has the law of SLE_κ shifted by $g_s(\gamma(s))$, but has two points on $\mathbb{R}$).

The other phase transition can be seen by examining the SLE-version of Cardy's formula: If $X = \inf\big([1,\infty) \cap \gamma[0,\infty)\big)$ denotes the first intersection of the SLE trace with the interval $[1,\infty)$, then a.s. $X = 1$ if $\kappa \geq 8$, whereas for $\kappa \in (4,8)$

$$\mathbb{P}[X \geq s] = \frac{4^{(\kappa-4)/\kappa}\,\sqrt{\pi}\,{}_2F_1(1-4/\kappa, 2-8/\kappa, 2-4/\kappa, 1/s)\,s^{(4-\kappa)/\kappa}}{\Gamma(2-4/\kappa)\,\Gamma(4/\kappa-1/2)}, \tag{6}$$

where ${}_2F_1$ denotes the hypergeometric function. At the corresponding time where $\gamma(t) = X$, the nontrivial interval $[1, X]$ gets "swallowed" by K at once. The proof of Cardy's formula in [RoS] is similar to the more elaborate Theorem 3.2 in [LSW2] and based on computing exit probabilities of a renormalized version of g_t,

$$Y_t = \frac{g_t(1) - W_t}{g_t(s) - W_t} \in (0,1).$$

At the exit time T, we have $Y_T = 0$ or 1 according to whether $X < s$ or $> s$. Now Cardy's formula can be obtained using standard methods of stochastic calculus.

For a simply connected domain $D \neq \mathbb{C}$ and boundary points p, q, chordal SLE from p to q in D is defined as the image of SLE in $\mathbb{H}$ under a conformal map of $\mathbb{H}$ onto D that takes 0 and ∞ to p and q. Since the conformal map between $\mathbb{H}$ and D generally does not extend to $\overline{\mathbb{H}}$, the continuity of the SLE trace in D does not follow from Theorem 3.6. However, using Theorem 3.9 below and general properties of conformal maps, it can be shown to still hold true, [GRS]. Another natural question is whether SLE is *reversible*, namely if SLE in D from p to q has the same law as SLE from q to p. This question was recently answered positively for $\kappa \leq 4$ by Dapeng Zhang [Z1]. It is known to be false for $\kappa \geq 8$ [RoS], and unknown for $4 < \kappa < 8$.

Theorem 3.8 ([Z1]). *For each $\kappa < 4$, SLE_κ is reversible, and for $\kappa \geq 8$ it is not reversible.*

The aforementioned derivative expectations $\mathbb{E}[|f_t'(z)|^p]$ also led to upper bounds for the dimensions of the trace and the frontier. The technically more difficult lower bounds were proved by Vincent Beffara [Be] for the trace.

For $\kappa > 4$, notice that the outer boundary of K_t is a simple curve joining two points on the real line. There is a relation between SLE_κ and $SLE_{16/\kappa}$, first derived by Duplantier with mathematically non-rigorous methods, and recently proved in the papers of Zhang [Z2] and Dubedat [Dub3]. Roughly speaking, Duplantier duality says that this curve is $SLE_{16/\kappa}$ between the two points. A precise formulation is based on a generalization of SLE, the so-called $SLE(\kappa, \rho)$ introduced in [LSW8]. As a consequence, the dimension of the frontier can thus be obtained from the dimension of the dual SLE.

Based on a clever construction of a certain martingale, in [SZ] Oded and Wang Zhou determined the size of the intersection of the trace with the real line. The same result was found independently and with a different method by Alberts and Sheffield [AlSh]. Summarizing:

Theorem 3.9. *For* $\kappa \leq 8$,

$$\dim\, \gamma[0, \mathrm{t}] = 1 + \frac{\kappa}{8}.$$

For $\kappa > 4$,

$$\dim\, \partial \mathrm{K_t} = 1 + \frac{2}{\kappa}.$$

For $4 < \kappa < 8$,

$$\dim\, \gamma[0, \mathrm{t}] \cap \mathbb{R} = 2 - \frac{8}{\kappa}.$$

The paper [SZ] also examined the question how the SLE trace tends to infinity. Oded and Zhou showed that for $\kappa < 4$, almost surely γ eventually stays above the graph of the function $x \mapsto x(\log x)^{-\beta}$, where $\beta = 1/(8/\kappa - 2)$.

3.3.4 Discrete processes converging to SLE

In [LSW2], Lawler, Schramm and Werner wrote that

... at present, a proof of the conjecture that SLE_6 is the scaling limit of critical percolation cluster boundaries seems out of reach...

Smirnov's proof [Sm1] of this conjecture came as a surprise. More precisely, he proved convergence of the critical site percolation exploration path on the triangular lattice (see Figure 12(b)) to SLE_6. See also [CN] and [Sm2]. This result was the first instance of a statistical physics model proved to converge to an SLE. The key to Smirnov's theorem is a version of Cardy's formula. Lennart Carleson realized that Cardy's formula assumes a very simple form when viewed in the appropriate geometry: When $\kappa = 6$, the right hand side $f(s)$ of (6) is a conformal map of the upper half plane onto an equilateral triangle ABC such that $0, 1$ and ∞ correspond to A, B and C. Since SLE_6 in ABC from A to B has the same law as the image of SLE_6 in $\mathbb{H}$ from 0 to ∞, the first point X' of intersection with BC has the law of $f(X)$. It follows that X' is uniformly distributed on BC. (A similar statement is true for all $4 < \kappa < 8$, where "equilateral" is replaced by "isosceles", and the angle of the triangle depends on κ, [Dub1]). Smirnov proved that the law of a corresponding observable on the lattice converges to a harmonic function, as the lattice size tends to zero. And he was able to identify the limit, through its boundary values. The proof makes use of the symmetries of the triangular lattice, and does not work on other lattices such as the square grid, where convergence is still unknown.

The next result concerning convergence to SLE was obtained by the usual suspects Lawler, Schramm and Werner [LSW9]. They proved Oded's original Conjecture 3.2.2 about convergence of

LERW to SLE_2, and the dual result (also conjectured in [S9]) that the UST converges to SLE_8, see Figures 10 and 13.

The *harmonic explorer* is a (random) interface defined as follows: Given a planar simply connected domain with two marked boundary points that partition the boundary into black and white hexagons, color all hexagons in the interior of the domain grey, see Figure 16 (a). The (growing)

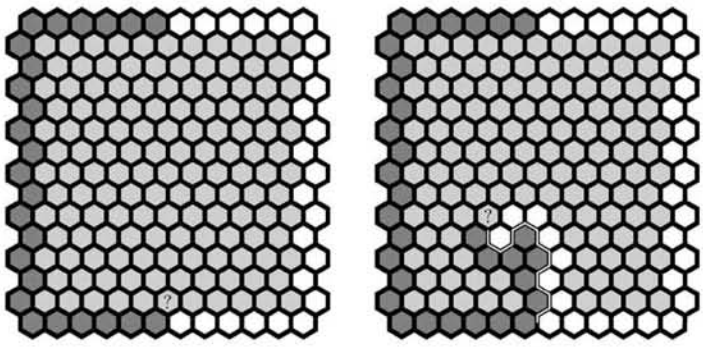


Figure 16: Definition of the Harmonic explorer path, from [SS1].

interface γ starts at one of the marked boundary points and keeps the black hexagons on its left and the white hexagons on its right. It is (uniquely) determined (by turning left at white hegagons and right at black) until a grey hexagon is met. When it meets a grey hexagon h (marked by ? in Figure 16) the (random) color of h is determined as follows. A random walk on the set of hexagons is started, beginning with the hexagon h. The walk stops as soon as it meets a white or black hexagon, and h assumes that color. Continuing in this fashion, γ will eventually reach the other boundary point. In [SS1], Oded and Scott Sheffield showed (distributional) convergence of γ to SLE_4. The overall strategy is again to directly analyze the Loewner driving term of the discrete path. The crucial property of SLE_4 is that, conditioned on the SLE trace $\gamma[0,t]$, the probability that a point $z \in \mathbb{H}$ will end up on the left of $\gamma[0,\infty)$ is a harmonic function of z (it is equal to the argument of $g_t(z) - W_t$, divided by π).

Other processes are believed to converge to SLE_4, too, in particular Rick Kenyon's double domino path, and the $q-$state Pott's model with $q = 4$.

The self-avoiding walk, first proposed in 1949 as a simple model for the structure of polymers, has played an important role in the development of SLE, in several ways: First, Lawler's invention of the LERW was partly motivated by the desire to create a model that is simpler than SAW. Second, the apparent similarity to the Brownian frontier motivated Mandelbrot's conjecture. Third, and most significantly, the SAW is conjectured to converge to $SLE_{8/3}$. See [LSW10] for precise formulations, and a proof of this conjecture assuming existence and conformal invariance of the scaling limit, and [K] for strong numerical evidence. However, still very little is known rigorously about the SAW.

Another famous classical model is the Ising model for ferromagnetism. Stas Smirnov [Sm3] has recently obtained another breakthrough concerning convergence of lattice models to SLE. He found

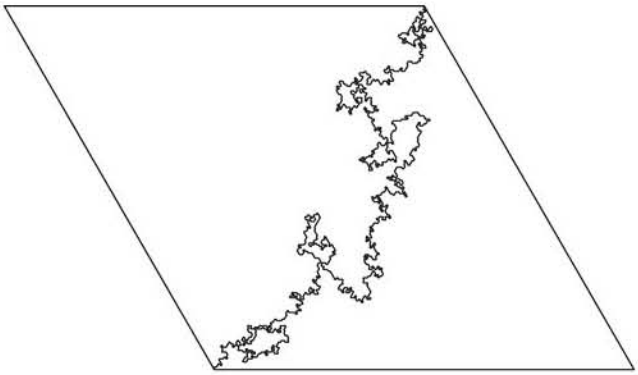


Figure 17: Harmonic explorer path, from [S11].

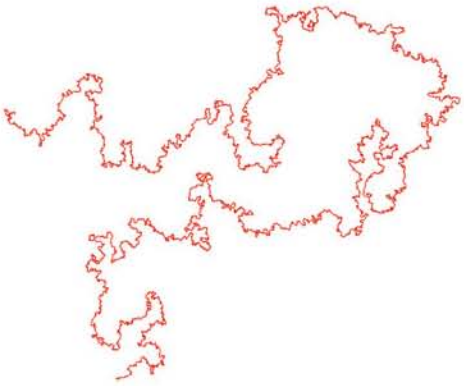


Figure 18: half-plane SAW, picture courtesy of Tom Kennedy.

observables for the Ising model at criticality and was able to prove their conformal invariance in the scaling limit. As a consequence, he obtained SLE_3 in the limit. Quoting from [Sm2]:

Theorem. *As the lattice step goes to zero, interfaces in Ising and Ising random cluster models on the square lattice at critical temperature converge to SLE(3) and SLE(16/3) correspondingly.*

3.3.5 Restriction measures

The elegant and important paper [LSW8] is a culmination of the universality arguments that have been initiated in [LW2] and developed in the subsequent collaboration of Lawler, Schramm and Werner. In the setting of random sets joining two boundary points of a simply connected domain, [LSW8] gives a complete characterization of laws satisfying the conformal restriction property, and various constructions of them.

Roughly speaking, a family of random sets K joining 0 and ∞ in $\mathbb{H}$ satisfies conformal restriction,

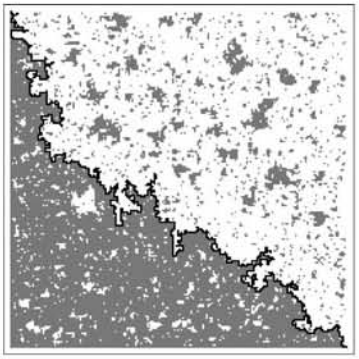


Figure 19: Critical Ising interface, picture courtesy of Stas Smirnov [Sm2].

if for every reasonable subdomain $D = \mathbb{H} \setminus A$ of $\mathbb{H}$, the law of K conditioned on $K \subset D$ is the same as the law of $g(K)$, where g is a conformal map from D to $\mathbb{H}$ fixing 0 and ∞. More precisely, the sets K are supposed to be connected, have connected complement, and are such that $\mathbb{H} \setminus K$ has two connected components. The subdomain D is reasonable if it is simply connected and contains (relative) neighborhoods of 0 and ∞.

An equivalent definition is to consider, for each simply connected domain D and each pair of boundary points a, b, a law $P_{D,a,b}$ on subsets of D joining a and b. Then the two required properties are conformal invariance, namely

$$g_* P_{D,a,b} = P_{g(D),g(a),g(b)}$$

for conformal maps g of D, and "restriction": For reasonable $D' \subset D$, the law $P_{D,a,b}$ of K, when restricted to $K \subset D'$, equals $P_{D',a,b}$. The remarkable main result is that there is a unique one-parameter family of such measures.

Theorem 3.10. *$\mathbb{P} = P_{\mathbb{H},0,\infty}$ is a conformal restriction measure if and only if there is $\alpha > 0$ such that*

$$\mathbb{P}[K \subset D] = g'_D(0)^\alpha \tag{7}$$

for every reasonable $D \subset \mathbb{H}$. For each $\alpha \geq \frac{5}{8}$ there is a conformal restriction measure P_α. Furthermore, $\alpha_0 = \frac{5}{8}$ is the smallest α for which there is a restriction measure, $P_{5/8}$ is the only restriction measure supported on simple curves, and $P_{5/8}$ is $SLE_{8/3}$.

An important observation, due to Balint Virag [V], is that P_1 is the law of Brownian excursions from 0 to ∞ in $\mathbb{H}$ (roughly, Brownian motion started at 0 and conditioned to "stay in $\mathbb{H}$" for all time). An elegant application goes as follows. If K_1 and K_2 are independent samples from P_{α_1} and P_{α_2}, then (7) implies that $K_1 \cup K_2$ has the law of $P_{\alpha_1+\alpha_2}$ (after the "loops" of the union have been filled in). By uniqueness, it follows that the law of the union of 5 independent Brownian excursions in $\mathbb{H}$ (plus loops) is the same as that of 8 copies of $SLE_{8/3}$ (with loops added). In particular, the frontiers are the same and thus have Hausdorff dimension 4/3 by Theorem 3.9. A similar result is [LSW8], Theorem 9.1: The law of the hull of whole-plane SLE_6, stopped when

reaching the boundary of a disc D, is the same as the law of a planar Brownian motion (with the bounded complementary components added), stopped upon leaving D.

The proof that $SLE_{8/3}$ is $P_{5/8}$ is based on the following computation. Using the same notation as (5), one can show

$$d\ h'_t(W_t) = h''_t(W_t)dW_t + \Big(\frac{h''_t(W_t)^2}{2h'_t(W_t)} + (\frac{\kappa}{2} - \frac{4}{3})h'''_t(W_t)\Big)dt. \tag{8}$$

For $\kappa = \frac{8}{3}$, it follows that

$$d\ h'_t(W_t)^{5/8} = \frac{5}{8}\frac{h''_t(W_t)}{h'_t(W_t)^{3/8}}dW_t$$

so that $h'_t(W_t)^{5/8}$ is a local martingale. Writing as before $T = \inf\{t : K_t \cap A \neq \emptyset\}$, it is not hard to show that $h'_t(W_t)$ tends to 0 as $t \to T$ if $K \cap A \neq \emptyset$ (the case $T < \infty$), and $\lim_{t\to T} h'_t(W_t) \to 1$ otherwise. Thus

$$\mathbb{P}[K \subset D] = \mathbb{P}[T = \infty] = \mathbb{E}[h'_T(W_T)^{5/8}] = \mathbb{E}[h'_0(W_0)^{5/8}] = g'_D(0)^{5/8}.$$

For values greater than 5/8, there are several constructions of the restriction measures described in [LSW8]. One is by adding "Brownian bubbles" to SLE-traces.

3.3.6 Other results

There are other versions of the Loewner equation. The "whole plane" equation was developed and used in [LSW3] to deal with hulls K_t that are growing in the plane rather than a disc or half-plane. "Di-polar SLE" was introduced in [BB2], see also [BBH]. An important generalization of SLE are the $SLE(\kappa, \rho)$ and variations, first introduced in [LSW8]. An elegant and unified treatment of all these variants is in [SWi]. Also, defining SLE in multiply connected domains creates a new difficulty that is not present in the simply connected case, since a slit multiply connected domain is not conformally equivalent to the unslit domain. See [Z],[BF1],[BF2].

Since SLE is amenable to computations, the convergence of discrete processes to SLE can be used to obtain results about the original process. In this fashion, Oded [S10] obtained the limiting probability, as the lattice size tends to zero in critical site percolation on the triangular lattice in the disc $\mathbb{D}$, that the union of a given arc $A \subset \partial\mathbb{D}$ and a percolation cluster surrounds 0. In [LSW4], Lawler, Schramm and Werner showed that the probability of the event $0 \leftrightarrow C_R$ that the percolation cluster containing the origin reaches the circle of radius R behaves like $R^{-5/48}$,

$$P[0 \leftrightarrow C_R] = R^{-5/48+o(1)}$$

as $R \to \infty$. See also [SW] for related exponents.

Because of space, in this note we have ignored the mathematically nutritious "Brownian loop soup" [LW3] and its relation to restriction measures, as well as the growing literature around the important Conformal Loop Ensemble CLE_κ introduced by Scott Sheffield [Sh2]. See [W3] and [SSW].

There are deep and exciting connections between the Gaussian Free Field and SLE, as explored by Oded and Scott Sheffield. The GFF has made its first appearance in this area in Rick Kenyon's work on the height of domino tilings [Ke3]. See [Sh1] for definitions and properties. Here is a very brief description of their work. Let $D \subset \mathbb{C}$ be a domain bounded by a simple closed curve that is partitioned into two arcs by two marked boundary points. Approximate D by a portion $G = (V, E)$ of the triangular grid as before (see Figure 20, where again vertices are represented by hexagons), and denote ∂V the boundary vertices. Fix a constant λ, and let $h = h_\varepsilon$ be an instance of the Discrete Gaussian Free Field, with boundary values $\pm\lambda$ on the two boundary arcs. This means that $h(v), v \in V \setminus \partial V$, is a $\sharp(V \setminus \partial V)$-dimensional Gaussian random variable whose density is proportional to $\exp(-\sum_{(u,v)\in E}(h(v) - h(u))^2/2)$. Extend h in a piecewise linear fashion from the vertices to the triangles. The main result of the deep and very long paper [SS2] is, roughly speaking, the following. If $\lambda = 3^{-1/4}\sqrt{\pi/8}$, then the level curve γ_ϵ of level $h = 0$, joining the two marked boundary points, converges to SLE_4 as $\varepsilon \to 0$. Other values of λ lead to variants of SLE_4. See also [SS3] and [Dub4].

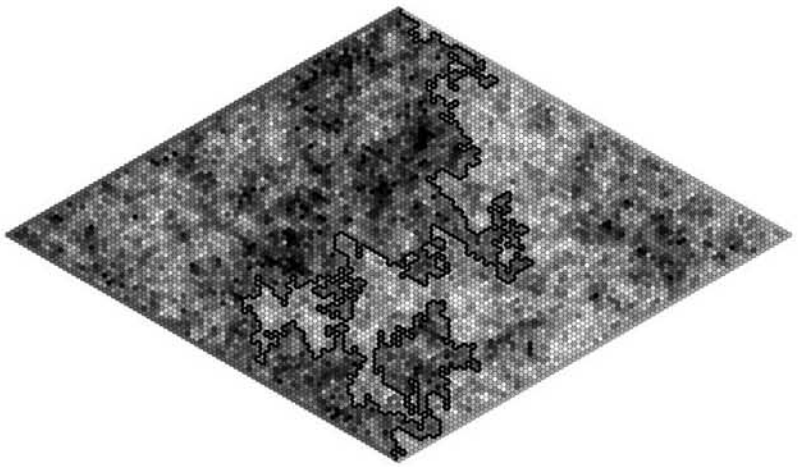


Figure 20: Level set of the Discrete GFF, from [SS2].

Finally, there are several collaborations of Oded being written at the moment. For instance, there are deep results of Christophe Garban, Gabor Pete and Oded concerning near-critical percolation and its scaling limit, which is different from SLE_6, see [GPS]. The nice paper [ShWi] describes Oded's (unpublished) proof of Watt's formula for double crossings in critical percolation, and provides insight into Oded's masterful use of Mathematica. Watt's formula was first proved rigorously by Dubedat [Dub2].

3.4 Problems

Many of Oded's papers contain open problems, some (such as [S6]) even propose a direction to tackle them. His ICM talk [S11] contains a large number of problems around SLE, and has provided the

field with a sense of direction. Some additional SLE-related problems are in [RoS]. Several of his problems have been solved since their publication.

As already mentioned, the convergence of the Ising interface to SLE_3, Problem 2.5 in [S11], was proved by Smirnov. The reversability of the chordal SLE path, Problem 7.3 of [S11], has been established for $\kappa \leq 4$ by Zhang [Z1], and the method has even led to a proof of a version of Duplantier's duality by Zhang and by Dubedat, [Z2], [Dub3]. Near-critical percolation and dynamical percolation (Sections 2.6 and 5 of [S11]) are now well-understood by [GPS] and [NW].

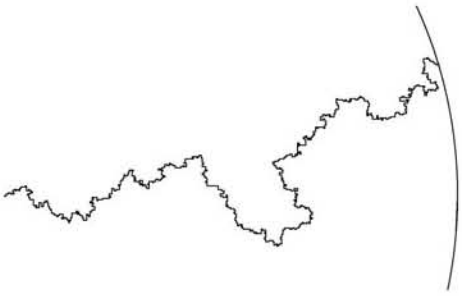
Spectacular progress has been made concerning the mathematical foundations of quantum gravity. Whereas the existence of an (unscaled) limit of "random triangulations of the sphere" was already established in [AnS], the important Knizhnik-Polyakov-Zamolodchikov formula (relating exponents of statistical physics models in "random geometries" to corresponding exponents in plane geometry) seemed out of reach until recently. In Section 4 of [S11], Oded wrote:

However, there is still no mathematical understanding of the KPZ formula. In fact, the author's understanding of KPZ is too weak to even state a concrete problem.

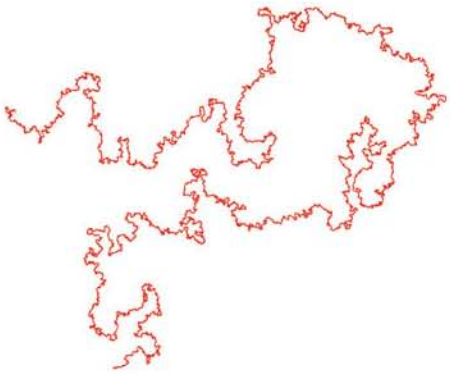
The Problem 4.1 in [S11], to show the existence of the (weak) Gromov-Hausdorff scaling limit of the graph metric on random triangulations of the sphere, was solved in the impressive work of Le Gall [LG]. Le Gall and Paulin showed [LGP] that the limiting space is a topological sphere, almost surely. Duplantier and Sheffield [DS] described a random measure (a scaling limit of the measure $e^{ch_\varepsilon} dxdy$ where h_ε is the Gaussian free field, averaged over circles of radius ε) which exhibits a KPZ-like relation. They conjecture a precise relation between a scaling limit of [AnS] and their random continuous space, and discuss connections to SLE. Following Duplantier and Sheffield, simpler random metric spaces exhibiting KPZ were considered in [BS4] and [RV].

3.5 Conclusion

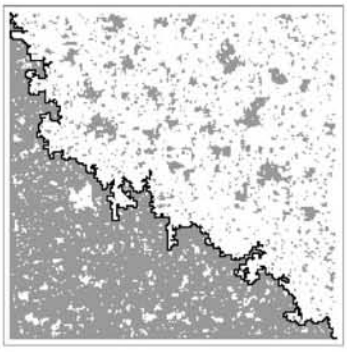
We have seen how Oded shaped the field of circle packings, and how he developed a deep understanding of discrete approximations to conformal maps. His results on the Koebe conjecture are still the best to date. We have also seen how Oded's discovery of SLE led to a powerful new tool in probability theory and in mathematical physics. In fact, it has changed the way physicists and mathematicians think about critical lattice interfaces, and has led to very fruitful interactions across disciplines. The number of mathematicians and physicists working with SLE is increasing fast, and the last few years have seen a number of exciting developments. Oded has already established his place in the history of mathematics. I have no doubt that we will see many more wonderful developments directly or indirectly related to Oded's work, thus keeping his spirit alive through the work of his fellow mathematicians, coauthors, and friends.



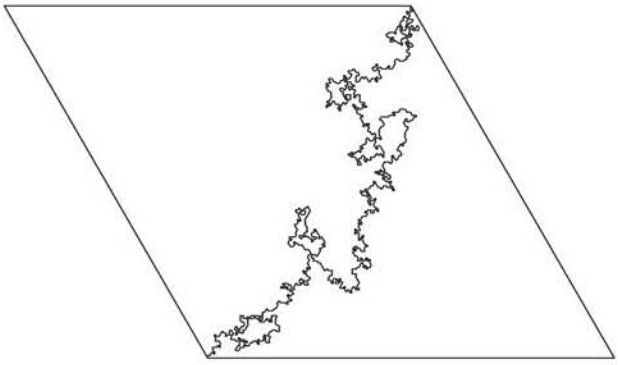
(a) LERW, $\kappa = 2$



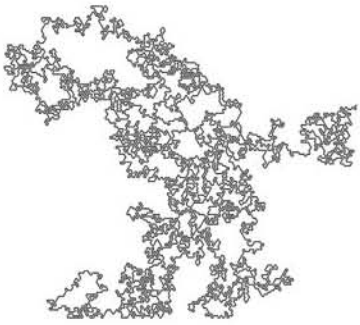
(b) SAW, $\kappa = \frac{8}{3}$?



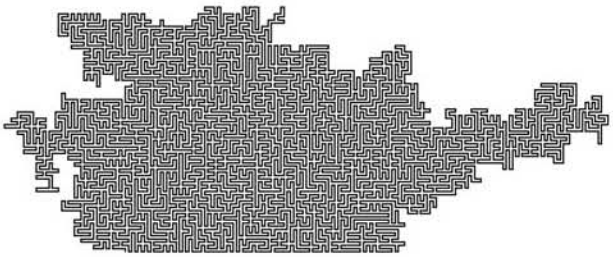
(c) Ising, $\kappa = 3$



(d) Harmonic explorer, $\kappa = 4$



(e) Percolation, $\kappa = 6$



(f) UST, $\kappa = 8$

Figure 21: Various random curves converging to SLE's

References

[AB] M. Aizenman, A. Burchard, Hölder regularity and dimension bounds for random curves, *Duke Math. J.* **99** (1999), 419–453.

[AlSh] T. Alberts, S. Sheffield, Hausdorff Dimension of the SLE curve intersected with the real line, *Electron. J. Probab.* **13** (2008), 1166–1188.

[A1] E. Andreev, Convex polyhedra in Lobacevskii space, *Math. USSR Sbornik* **10** (1970), 413–440.

[A2] E. Andreev, Convex polyhedra of finite volume in Lobacevskii space, *Math. USSR Sbornik* **12** (1970), 255-259.

[AnS] O. Angel, O. Schramm, Uniform Infinite Planar Triangulations, *Comm.Math.Phys.* **241** (2003), 191–213.

[BB1] M. Bauer, D. Bernard, Conformal Field Theories of Stochastic Loewner Evolutions, *Commun. Math. Phys.* **239** (2003), 493–521.

[BB2] M. Bauer, D. Bernard, SLE, CFT and zig-zag probabilities, arXiv:math-ph/0401019v1.

[BB3] M. Bauer, D. Bernard, 2D growth processes: SLE and Loewner chains, *Phys. Rept.* **432** (2006), 115–221.

[BBH] M. Bauer, D. Bernard, J. Houdayer, Dipolar stochastic Loewner evolutions, *J. Stat. Mech. Theory Exp.* **3** (2005).

[BF1] R. Bauer, R. Friedrich, Stochastic Loewner evolution in multiply connected domains, *C. R. Math. Acad. Sci. Paris* **339** (2004), 579–584.

[BF2] R. Bauer, R. Friedrich, Chordal and bilateral SLE in multiply connected domains, *Math. Z.* **258** (2008), 241–265.

[BSt1] A. Beardon, K. Stephenson, Circle packings in different geometries, *Tohoku Math. J.* **43** (1991), 27–36.

[BSt2] A. Beardon, K. Stephenson, The Schwarz-Pick lemma for circle packings, *Illinois J. Math.* **35** (1991), 577–606.

[Be] V. Beffara, The dimension of the SLE curves, *Ann. Probab.* **36** (2008) 1421-1452.

[BS1] I. Benjamini, O. Schramm, Random walks and harmonic functions on infinite planar graphs using square tilings, *Ann. Probab.* **24** (1996), 1219–1238.

[BS2] I. Benjamini, O. Schramm, Harmonic functions on planar and almost planar graphs and manifolds, via circle packings, *Invent. Math.* **126** (1996), 565–587.

[BS3] I. Benjamini, O. Schramm, Recurrence of Distributional Limits of Finite Planar Graphs, *Electron. J. Probab.* **6** (2001), 1–13.

[BS4] I. Benjamini, O. Schramm, KPZ in one dimensional random geometry of multiplicative cascades, *Comm. Math. Phys.* **289** (2009), 653–662.

[BLPS] I. Benjamini, R. Lyons, Y. Peres, O. Schramm, Uniform spanning forests, *Ann. Probab.* **29** (2001), 1–65.

[BJPP] C. Bishop, P. Jones, R. Pemantle, Y. Peres, The dimension of the Brownian frontier is greater than 1, *J. Funct. Anal.* **143**, (1997),309336.

[BHS] A. Bobenko, T. Hoffmann, B. Springborn, Minimal surfaces from circle patterns: geometry from combinatorics, *Ann. of Math.* **164** (2006), 231–264.

[BK] M. Bonk, B. Kleiner, Quasisymmetric parametrizations of two-dimensional metric spheres, *Invent. Math.* **150** (2002), 127–183.

[Bo] M. Bonk, Quasiconformal geometry of fractals, International Congress of Mathematicians. Vol. II, Eur. Math. Soc., Z"urich (2006), 1349–1373.

[Br] M. Brandt, Ein Abbildungssatz fuer endlich-vielfach zusammenhängende Gebiete, *Bull. de la Soc. des Sc. et des Lettr. de Lodz XXX,* **4** (1980).

[BSST] R. Brooks, C. Smith, A. Stone, W. Tutte, The dissection of squares into squares, *Duke Math. J.* **7** (1940), 312–340.

[Bu] K. Burdzy, Cut points on Brownian paths, *Ann. Probab.* **17** (1989), 1012–1036.

[CN] F. Camia, C. Newman, Critical percolation exploration path and SLE_6: a proof of convergence, *Probab. Theory Related Fields* **139** (2007), 473-519.

[Can] J. Cannon, The combinatorial Riemann mapping theorem, *Acta Math.* **173** (1994), 155–234.

[CFP1] J. Cannon, W. Floyd, W. Parry, Squaring rectangles: The finite Riemann mapping theorem, *Contemp. Math.* **169** (1994), Amer. Math. Soc., Providence, RI, 133–212.

[CFP2] J. Cannon, W. Floyd, W. Parry, The Length-Area method and discrete Riemann mappings, manuscript (1999)

[Car1] J. Cardy, Critical percolation in finite geometries, *J. Phys. A* **25***(4)* (1992), 201–206.

[Car2] J. Cardy, SLE for theoretical physicists, *Annals Phys.* **318** (2005), 81-118.

[CM] L. Carleson, N. Makarov, Aggregation in the plane and Löwner's equation, *Comm. Math. Phys.* **216** (2001) 583–607.

[CMS] R. Courant, B. Manel, M. Shiffman, A general theorem on conformal mapping of multiply connected domains, *Proc. Nat. Acad. Sci. U. S. A.* **26** (1940), 503–507.

[dV1] C. de Verdière, Empilements de cercles: Convergence d'une méthode de point fixe, *Forum Math.* **I** (1989), 395–402.

[dV2] C. de Verdière, Une principe variationnel pour les empilements de cercles, *Invent. Math.* **104** (1991), 655–669.

[DHR] P. Doyle, Z.-X. He, B. Rodin, Second derivatives of circle packings and conformal mappings, *Discrete Comput. Geom.* **11** (1994), 35–49.

[Dub1] J. Dubedat, SLE and triangles, *Elect. Comm. in Probab.* **8** (2003), 28–42.

[Dub2] J. Dubedat, Excursion decompositions for SLE and Watts crossing formula, *Probab. Theory Related Fields* **134** (2006), 453-488.

[Dub3] J. Dubedat, Duality of Schramm-Loewner Evolutions, *Ann. Sci. c. Norm. Supr.* **42** (2009), 697–724.

[Dub4] J. Dubedat, SLE and the free field: partition functions and couplings, *J. Amer. Math. Soc.* **22** (2009), 995–1054.

[Duf] R. Duffin, The extremal length of a network, *J. Math. Anal. Appl.* **5** (1962), 200–215.

[Dup] B. Duplantier, Conformal Random Geometry, arXiv:math-ph/0608053v1.

[DK] B. Duplantier, K. Kwon, Conformal invariance and intersection of random walks, *Phys. Rev. Let.* **61** (1988), 2514–2517.

[DS] B. Duplantier, S. Sheffield, Liouville Quantum Gravity and KPZ, arXiv:0808.1560v1.

[Dur] P. Duren, Univalent functions, Springer-Verlag, New York, 1983.

[GPS] C. Garban, G. Pete, O. Schramm, The scaling limit of the Minimal Spanning Tree - a preliminary report, arXiv:0909.3138.

[GRS] C. Garban, S. Rohde, O. Schramm, Continuity of the SLE trace in simply connected domains, arXiv:0810.4327, *Israel J. Math* to appear.

[GM] J. Garnett, D. Marshall, Harmonic measure, Cambridge University Press, New York, 2005.

[GK] I. Gruzberg, L. Kadanoff, The Loewner equation: maps and shapes, *J. Stat. Phys.* **114** (2004), 1183–1198.

[Ha] A. Harrington, Conformal mappings onto domains with arbitrarily specified boundary shapes, J. Analyse Math. 41 (1982), 39–53.

[HL] M. Hastings, L. Levitov, Laplacian growth as one-dimensional turbulence, *Physica D* **116** (1998) 244–252.

[He] Z.-X. He, An estimate for hexagonal circle packings, *J. Differential Geom.* **33** (1991), 395–412.

[HR] Z.-X. He, B. Rodin, Convergence of circle packings of finite valence to Riemann mappings, *Comm. Anal. Geom.* **1** (1993), 31–41.

[HS1] Z.-X. He, O. Schramm, Fixed points, Koebe uniformization and circle packings, *Ann. of Math. (2)* **137** (1993), 369–406.

[HS2] Z.-X. He, O. Schramm, Rigidity of circle domains whose boundary has σ-finite linear measure, *Invent. Math.* **115** (1994), 297–310.

[HS3] Z.-X. He, O. Schramm, Hyperbolic and parabolic packings, *Discrete Comput. Geom.* **14** (1995), 123–149.

[HS3b] Z.-X. He, O. Schramm, A hyperbolicity criterion for circle packings, preprint, 1991.

[HS4] Z.-X. He, O. Schramm, Koebe uniformization for "almost circle domains", *Amer. J. Math.* **117** (1995), 653–667.

[HS5] Z.-X. He, O. Schramm, The inverse Riemann mapping theorem for relative circle domains, *Pacific J. Math.* **171** (1995), 157–165.

[HS6] Z.-X. He, O. Schramm, On the convergence of circle packings to the Riemann map, *Invent. Math.* **125** (1996), no. 2, 285–305.

[HS7] Z.-X. He, O. Schramm, On the distortion of relative circle domain isomorphisms, *J. Anal. Math.* **73** (1997), 115–131.

[HS8] Z.-X. He, O. Schramm, The C^∞-convergence of hexagonal disk packings to the Riemann map, *Acta Math.* **180** (1998), 219–245.

[HK] J. Heinonen, P. Koskela, Definitions of quasiconformality, *Invent. Math.* **120** (1995), 61–79.

[KN] W. Kager, B. Nienhuis, A Guide to Stochastic Loewner Evolution and its Applications, *J. Stat. Phys.* **115** (2004), 1149–1229.

[K] T. Kennedy, Conformal invariance and stochastic Loewner evolution predictions for the 2D self-avoiding walk–Monte Carlo tests, *J. Statist. Phys.* **114** (2004), 51–78.

[Ke1] R. Kenyon, Conformal invariance of domino tiling, *Ann. Probab.* **28** (2000), 759–795.

[Ke2] R. Kenyon, The asymptotic determinant of the discrete Laplacian, *Acta Math.* **185** (2000), 239–286.

[Ke3] R. Kenyon, Dominos and the Gaussian free field, *Ann. Probab.* **29** (2001), 1128–1137.

[Ko1] P. Koebe, Ueber die Uniformisierung beliebiger analytischer Kurven (Dritte Mitteilung), *Nachr. Ges. Wiss. Gött.* (1908), 337-358.

[Ko2] P. Koebe, Kontaktprobleme der Konformen Abbildung, *Ber. Sächs. Akad. Wiss. Leipzig, Math.-Phys. Kl.* **88** (1936), 141–164.

[Ku] P. Kufarev, A remark on integrals of Löwner's equation, *Doklady Akad. Nauk SSSR* **57** (1947), 655–656, in Russian.

[La] G. Lawler, A self avoiding walk, *Duke Math. J.* **47** (1980), 655–694.

[La2] G. Lawler, Geometric and fractal properties of Brownian motion and random walk paths in two and three dimensions, *Bolyai Soc. Math. Stud.* **9** (1998), 219–258.

[La3] G. Lawler, Conformally Invariant Processes in the Plane, *Mathematical Surveys and Monographs,* **114**. American Mathematical Society, Providence, RI, 2005.

[LSW1] G. Lawler, O. Schramm, W. Werner, The dimension of the planar Brownian frontier is 4/3, *Math. Res. Lett.* **8** (2001), 401–411.

[LSW2] G. Lawler, O. Schramm, W. Werner, Values of Brownian intersection exponents I, Half-plane exponents, *Acta Math.* **187** (2001), 237–273.

[LSW3] G. Lawler, O. Schramm, W. Werner, Values of Brownian intersection exponents II, Plane exponents, *Acta Math.* **187** (2001), 275–308.

[LSW4] G. Lawler, O. Schramm, W. Werner, One-arm exponent for critical 2D percolation, *Electron. J. Probab.* **7** (2002), 1–13.

[LSW5] G. Lawler, O. Schramm, W. Werner, Values of Brownian intersection exponents III, Two-sided exponents, *Ann. Inst. H. Poincar Probab. Statist.* **38** (2002), 109–123.

[LSW6] G. Lawler, O. Schramm, W. Werner, Sharp estimates for Brownian non-intersection probabilities, *In and out of equilibrium* (Mambucaba, 2000), 113–131, Progr. Probab., 51, Birkhuser Boston, Boston, MA, 2002.

[LSW7] G. Lawler, O. Schramm, W. Werner, Analyticity of intersection exponents for planar Brownian motion, *Acta Math.* **189** (2002), 179–201.

[LSW8] G. Lawler, O. Schramm, W. Werner, Conformal restriction: the chordal case, *J. Amer. Math. Soc.* **16** (2003), 917–955.

[LSW9] G. Lawler, O. Schramm, W. Werner, Conformal invariance of planar loop-erased random walks and uniform spanning trees, *Ann. Probab.* **32** (2004), 939–995.

[LSW10] G. Lawler, O. Schramm, W. Werner, On the scaling limit of planar self-avoiding walk, *Fractal geometry and applications: a jubilee of Benoit Mandelbrot, Part 2*, 339–364, Proc. Sympos. Pure Math., 72, Part 2, Amer. Math. Soc., Providence, RI, 2004.

[LW1] G. Lawler, W. Werner, Intersection exponents for planar Brownian motion, *Ann. Probab.* **27** (1999) 1601–1642.

[LW2] G. Lawler, W. Werner, Universality for conformally invariant intersection exponents, *J. Europ. Math. Soc.* **2** (2000) 291–328.

[LW3] G. Lawler, W. Werner, The Brownian loop soup, *Probab. Theory Related Fields* **128** (2004), 565–588.

[LG] J. Le Gall, The topological structure of scaling limits of large planar maps, *Invent. Math.* **169** (2007), 621–670.

[LGP] J. Le Gall, F. Paulin, Scaling limits of bipartite planar maps are homeomorphic to the 2-sphere, *Geom. Funct. Anal.* **18** (2008), 893–918.

[Lo] C. Loewner, Untersuchungen über schlichte konforme Abbildungen des Einheitskreises, I, *Math. Ann.* **89** (1923), 103–121.

[Ly] R. Lyons, A bird's-eye view of uniform spanning trees and forests, Microsurveys in discrete probability (Princeton, NJ, 1997), 135–162.

[M] B. Mandelbrot, The beauty of fractals, Freeman 1982.

[MR] D. Marshall, S. Rohde, The Loewner differential equation and slit mappings, *J. Amer. Math. Soc.* **18** (2005), 763–778.

[MS] D. Marshall, C. Sundberg, Harmonic measure and radial projection, *Trans. Amer. Math. Soc.* **316** (1989), 81–95.

[MT] G. Miller, W. Thurston, Separators in two and three dimensions, In Proceedings of the 22nd Annual ACM Symposium on Theory of Computing, pages 300309. ACM, Baltimore, 1990.

[NW] P. Nolin, W. Werner, Asymmetry of near-critical percolation interfaces, *J. Amer. Math. Soc.* **22** (2009), 797–819.

[Pe] R. Pemantle, Choosing a spanning tree for the integer lattice uniformly, *Ann. Probab.* **19** (1991), 1559 – 1574.

[Po1] C. Pommerenke, On the Loewner differential equation, *Michigan Math. J.* **13** (1966), 435–443.

[Po2] C. Pommerenke, Boundary behaviour of conformal maps, Springer Verlag, Berlin Heidelberg 1992

[Pop1] N.V. Popova, Investigation of some integrals of the equation $\frac{dw}{dt} = \frac{A}{w-\lambda(t)}$, *Novosibirsk Gos. Ped Inst Uchon. Zapiski* **8** (1949), 13–26.

[Pop2] N.V. Popova, Relations between Löwner's equation and the equation $\frac{dw}{dt} = \frac{1}{w-\lambda(t)}$, *Izv. Akad. Nauk BSSR, Ser. Fiz.-Mat. Nauk* **6** (1954), 97–98.

[RV] R. Rhodes, V. Vargas, KPZ formula for log-infinitely divisible multifractal random measures, 2008, arXiv:math.PR/0807.1036.

[RSu] B. Rodin, D. Sullivan, The convergence of circle packings to the Riemann mapping, *J. Differential Geometry* **26** (1987), 349–360.

[RoS] S. Rohde, O. Schramm, Basic properties of SLE, *Ann. Math.* **161** (2005), 879–920.

[Sa] H. Sachs, Coin graphs, polyhedra, and conformal mapping, *Discrete Math.* **134** (1994), 133-138.

[S1] O. Schramm, Combinatorially prescribed packings and applications to conformal and quasiconformal maps, Ph. D. thesis, Princeton (1990).

[S2] O. Schramm, Rigidity of infinite (circle) packings, *J. Amer. Math. Soc.* **4** (1991), 127–149.

[S3] O. Schramm, How to cage an egg, *Invent. Math.* **107** (1993), 543–560.

[S4] O. Schramm, Existence and uniqueness of packings with specified combinatorics, *Israel J. Math.* **73** (1991), 321–341.

[S5] O. Schramm, Square tilings with prescribed combinatorics, *Israel J. Math.* **84** (1993), 97–118.

[S6] O. Schramm, Transboundary extremal length, *J. Anal. Math.* **66** (1995), 307–329.

[S7] O. Schramm, Conformal uniformization and packings, *Israel J. Math.* **93** (1996), 399–428.

[S8] O. Schramm, Circle patterns with the combinatorics of the square grid, *Duke Math. J.* **86** (1997), 347–389.

[S9] O. Schramm, Scaling limits of loop-erased random walks and uniform spanning trees, *Israel J. Math.* **118** (2000), 221–288.

[S10] O. Schramm, A percolation formula, *Electron. Comm. Probab.* **6** (2001), 115–120.

[S11] O. Schramm, Conformally invariant scaling limits: an overview and a collection of problems. In *Proceedings of ICM Madrid 2006,* **1**, 513–543. European Math. Soc. 2007.

[SS1] O. Schramm, S. Sheffield, Harmonic explorer and its convergence to SLE_4, *Ann. Probab.* **33** (2005), 2127–2148.

[SS2] O. Schramm, S. Sheffield, Contour lines of the two-dimensional discrete Gaussian free field, *Acta Math.* **202** (2009), 21–137.

[SS3] O. Schramm, S. Sheffield, A contour line of the continuum Gaussian free field, preprint (2010).

[SSW] O. Schramm, S. Sheffield, D. Wilson, Conformal radii for conformal loop ensembles, *Comm. Math. Phys.* **288** (2009), 43–53.

[SWi] O. Schramm, D. Wilson, SLE coordinate changes, *New York J. Math.* **11** (2005), 659–669.

[SZ] O. Schramm, W. Zhou, Boundary proximity of SLE, *Probab. Theory Related Fields* **146** (2010), 435–450.

[Sh1] S. Sheffield, Gaussian free fields for mathematicians, *Probab. Theory Related Fields* **139** (2007), 521–541.

[Sh2] S. Sheffield, Exploration trees and conformal loop ensembles, *Duke Math. J.* **147** (2009), 79–129.

[ShWi] S. Sheffield, D. Wilson, Schramm's proof of Watts' formula, rXiv:1003.3271.

[Sm1] S. Smirnov, Critical percolation in the plane: conformal invariance, Cardy's formula, scaling limits, *C. R. Acad. Sci. Paris Sr. I Math.* **333** (2001), 239–244.

[Sm2] S. Smirnov, Towards conformal invariance of 2D lattice models, In *Proceedings of ICM Madrid 2006,* **2**, 1421–1451. European Math. Soc. 2007.

[Sm3] S. Smirnov, Conformal invariance in random cluster models I, Holomorphic fermions in the Ising model, arXiv:0708.0039, *Ann. Math.*, to appear.

[SW] S. Smirnov, W. Werner: Critical exponents for two-dimensional percolation, *Math. Rev. Lett.* **8** (2001), 729744.

[Ste1] K. Stephenson, A probabilistic proof of Thurston's conjecture on circle packings, *Rend. Sem. Mat. Fis. Milano* **66** (1996), 201–291.

[Ste2] K. Stephenson, Introduction to circle packing. The theory of discrete analytic functions, Cambridge University Press, Cambridge, 2005.

[Str] K. Strebel, Über das Kreisnormierungsproblem der konformen Abbildung, *Ann. Acad. Sci. Fennicae. Ser. A* **101** (1951), 1–22.

[T] W. Thurston, The geometry and Topology of 3-manifolds, Chap. 13 (Math. Notes, Princeton) Princeton University Press 1980.

[V] B. Virag, Brownian Beads, *Probab. Theory Related Fields* **127** (2003), 367–387.

[W1] W. Werner, Random planar curves and Schramm-Loewner evolutions, *Lecture Notes in Math.* **1840**, Springer, Berlin, 2004.

[W2] W. Werner, Conformal restriction and related questions, *Probab. Surv.* **2** (2005), 145–190.

[W3] W. Werner, The conformally invariant measure on self-avoiding loops, *J. Amer. Math. Soc.* **21** (2008), 137–169.

[WS] T. Witten, L. Sander, Diffusion-limited aggregation, a kinetic phenomenon, *Phys. Rev. Lett.* **47** (1981), 1400-1403.

[Z] D. Zhan, Loewner evolution in doubly connected domains, *Probab. Theory Related Fields* **129** (2004), 340–380.

[Z1] D. Zhan, Reversibility of chordal SLE, *Ann. Probab.* **36** (2008), 1472–1494.

[Z2] D. Zhan, Duality of chordal SLE, *Invent. Math.* **174** (2008), 309–353.

ILLUMINATING SETS OF CONSTANT WIDTH

ODED SCHRAMM

Abstract. The problem of illuminating the boundary of sets having constant width is considered and a bound for the number of directions needed is given. As a corollary, an estimate for Borsuk's partition problem is inferred. Also, the illumination number of sufficiently symmetric strictly convex bodies is determined.

§1. *Introduction.* Let $\mathbf{x}$ be a point on the boundary ∂K of a convex body* K in Euclidean space, $\mathbf{R}^n$. A direction $\mathbf{u} \in S^{n-1}$ is said to *illuminate* K at $\mathbf{x}$ if the line $\{x+t\mathbf{u} \mid t \in \mathbf{R}\}$ "enters" $\mathbf{K}$ in $\mathbf{x}$. More precisely, $\mathbf{u}$ illuminates K at $\mathbf{x}$ if $\mathbf{x}+t\mathbf{u}$ is an interior point of K, for some positive t. Instead of saying "$\mathbf{u}$ illuminates K at $\mathbf{x}$" we will just say "$\mathbf{u}$ illuminates $\mathbf{x}$". This will cause no confusion, because the convex body K should be clear from the context.

Directions $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m \in S^{n-1}$ are said to illuminate K if every point on the boundary of K is illuminated by at least one of these directions. We denote by $I(K)$ the minimal number of directions sufficient to illuminate K and call it the *illumination number* of K. This concept of illumination was introduced by V. G. Boltjansky in [2]. There he proved that for convex bodies K, $I(K)$ is equal to $H(K)$, the number of smaller, positively homothetic copies of K required to cover K (see [3]).

The maximum values of $I(K)$ for convex bodies K in $\mathbf{R}^n$, are unknown when $n > 2$. By the above result of Boltjansky, Hadwiger's conjecture, about the covering of a set by homothetic copies of it, is equivalent to

$$I(\text{convex body in } \mathbf{R}^n) \leq I(n\text{-dimensional parallelogram}) = 2^n.$$

See [3] for a discussion of this conjecture.

A set of *constant width d* is a convex body such that the distance between any two distinct parallel supporting hyperplanes of it is d (see [5, pp. 122–131], [4]). In this note we prove

THEOREM 1. *If W is a set of constant width in $\mathbf{R}^n$ then*

$$I(W) < 5n\sqrt{n}(4+\log n)\left(\frac{3}{2}\right)^{n/2}.$$

The research exposed in this note was done while I was at the Hebrew University in Jerusalem, as a student of Professor Gil Kalai. I am grateful to Kalai for his interest, encouragement and advice, and for valuable improvements of the original manuscript.

* A convex body is a compact convex set that has interior points.

[MATHEMATIKA, 35 (1988), 180–189]

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_2, © Springer Science+Business Media, LLC 2011

Using this theorem, and the equivalence $I(W) = H(W)$, we get

COROLLARY 2. *Every set of constant width $W \subset \mathbf{R}^n$ can be covered by less than $5n\sqrt{n}(4+\log n)(3/2)^{n/2}$ homothetic copies of itself, having some homothety coefficient α, $0 < \alpha < 1$.*

Since every bounded set of positive diameter is contained in a set of constant width having the same diameter (see [**5**, p. 126]), we have the following estimate for Borsuk's partition problem.

COROLLARY 3. *Every set of diameter d $(0 < d < \infty)$ in $\mathbf{R}^n$ can be covered by less than $5n\sqrt{n}(4+\log n)(3/2)^{n/2}$ sets having smaller diameters.*

To the best of our knowledge, this is, asymptotically, the best bound known for Borsuk's partition problem (see [**7**] for references and results concerning this problem). The approach to Borsuk's problem through the the illumination problem is suggested in [**3**] and [**8**, p. 420].

Theorem 1 is proved using a probabilistic argument: The probability that a "small" region in ∂W will be illuminated by a random, uniformly distributed, direction is estimated from below. A straightforward computation then shows that if "enough" directions are chosen randomly, uniformly, and independently, the probability that they will completely illuminate W is nonzero. An essential part of the proof relies on finding lower bounds for volumes of spherical sets of a certain type. In [**11**] analogous results for sets in $\mathbf{R}^n$ give lower bounds for the volumes of sets having constant width.

The only known result in the direction opposite to Theorem 1 is that $I(K) \geq n+1$ for every convex body $K \subset \mathbf{R}^n$ ([**3**]). It is not known if there is a set of constant width $W \subset \mathbf{R}^n$, that cannot be illuminated by $n+1$ directions.

Rogers [**10**] has shown that every $K \subset \mathbf{R}^n$, having diameter $0 < d < \infty$ and invariant under the group of congruences that leave invariant an n-dimensional regular simplex, can be partitioned into $n+1$ subsets, each with diameter $< d$. Inspired by this, we prove in Section 4 that if K is a strictly convex body† invariant under a group of orthogonal transformations that is generated by reflections through hyperplanes and acts irreducibly‡ on $\mathbf{R}^n$, then $I(K) = n+1$. This can give an alternate proof of Rogers' result.

§2. First we will introduce some notation and give a condition for the illumination of a single boundary point. Throughout this note, K will denote an arbitrary convex body, and W will denote a set of constant width in $\mathbf{R}^n$. For a set $A \subset S^{n-1}$ we define

$$A^+ = \{\mathbf{u} \in S^{n-1} \mid \mathbf{u} \cdot \mathbf{v} > 0 \text{ for all } \mathbf{v} \in A\}.$$

† A strictly convex body is a convex body whose boundary contains no line segment.

‡ "G acts irreducibly on $\mathbf{R}^n$" means that there is no subspace of $\mathbf{R}^n$, other than 0 and $\mathbf{R}^n$, which is invariant under all elements of G.

When $\mathbf{x}$ is a boundary point of a convex body K, we use $N_K(\mathbf{x})$ to denote the set of inward normal unit vectors of K at $\mathbf{x}$:

$$N_K(\mathbf{x}) = \{\mathbf{u} \in S^{n-1} \mid \mathbf{u} \cdot \mathbf{p} \geq \mathbf{u} \cdot \mathbf{x} \text{ for all } \mathbf{p} \in K\}.$$

Notice that $N_K(\mathbf{x})$ is nonempty for $\mathbf{x} \in \partial K$.

LEMMA 4. *Let K be a convex body in $\mathbf{R}^n$, let $\mathbf{x}$ be a boundary point of K, and let $\mathbf{u} \in S^{n-1}$. Then $\mathbf{x}$ is illuminated by the direction $\mathbf{u}$, if, and only if, $\mathbf{u} \in N_K(\mathbf{x})^+$.*

Proof. Suppose $\mathbf{u} \in N_K(\mathbf{x})^+$. If the line $L = \{\mathbf{x} + t\mathbf{u} \mid t \in \mathbf{R}\}$ contains no interior points of K, then it is a supporting line of K. Therefore there is a hyperplane, $H = \{\mathbf{p} \in \mathbf{R}^n \mid \mathbf{p} \cdot \mathbf{w} = \mathbf{x} \cdot \mathbf{w}\}$, supporting K and containing L. Thus one of the vectors $\pm\mathbf{w}/\|\mathbf{w}\|$ is in $N_K(\mathbf{x})$. That is a contradiction to $\mathbf{u} \in N_K(\mathbf{x})^+$, because $H \supset L$ implies that $\mathbf{w} \cdot \mathbf{u} = 0$. Therefore L contains interior points of K. Pick any $\mathbf{v} \in N_K(\mathbf{x})$. K is contained in the half space $\{\mathbf{p} \in \mathbf{R}^n \mid \mathbf{p} \cdot \mathbf{v} \geq \mathbf{x} \cdot \mathbf{v}\}$. Since $\mathbf{u} \cdot \mathbf{v} > 0$, this means that the points $\mathbf{x} + t\mathbf{u}$ with $t < 0$ are not in K. Thus the interior points of L correspond to positive values of t, and $\mathbf{u}$ illuminates $\mathbf{x}$.

Now suppose that $\mathbf{u}$ illuminates $\mathbf{x}$. Let t be a positive number such that $\mathbf{p} = \mathbf{x} + t\mathbf{u}$ is an interior point of K. Because $\mathbf{p}$ is an interior point, for every $\mathbf{v} \in N_K(\mathbf{x})$ we have $\mathbf{p} \cdot \mathbf{v} > \mathbf{x} \cdot \mathbf{v}$, which implies $\mathbf{u} \cdot \mathbf{v} > 0$, so $\mathbf{u} \in N_K(\mathbf{x})^+$.

From now on, we work with an arbitrary, but fixed, set of constant width $W \subset \mathbf{R}^n$. The lemma above shows that if E is a subsets of the boundary of W, then one direction can illuminate E, if, and only if,

$$\bigcap_{\mathbf{x} \in E} N_W(\mathbf{x})^+ = \left(\bigcup_{\mathbf{x} \in E} N_W(\mathbf{x}) \right)^+$$

is nonempty. The following proposition will help us find subsets of ∂W that are "easily" illuminated. For a subset $A \subset S^{n-1}$ define $U_W(A)$ to be the union of the sets $N_W(\mathbf{x})$, $\mathbf{x} \in \partial W$, that intersect A:

$$U_W(A) = \bigcup_{N_W(\mathbf{x}) \cap A \neq \emptyset} N_W(\mathbf{x}).$$

A direction in $U_W(A)^+$ illuminates every point $\mathbf{x} \in W$ that satisfies

$$N_W(\mathbf{x}) \cap A \neq \emptyset.$$

In order to show that when A is chosen properly these points are "easily" illuminated, we want to prove that $U_W(A)^+$ is "large". Our means of doing so is by estimating the diameter of $U_W(A)$. (We view S^{n-1} with the metric induced by the Euclidean metric in $\mathbf{R}^n$. The diameters of subsets of S^{n-1} refer to this metric.)

PROPOSITION 5. *Let A be a nonempty subset of S^{n-1}. Then*

$$\text{diameter } U_W(A) \leq 1 + \text{diameter } A.$$

Proof. Since $U_W(A)$ does not change when we replace W with a positively homothetic copy of itself, we may, and will, assume that W has constant width

1, and therefore also diameter 1. Let $\mathbf{v}_1, \mathbf{v}_2$ be unit vectors in $U_W(A)$. By the definition of $U_W(A)$, there are points $\mathbf{x}_1, \mathbf{x}_2 \in \partial W$ such that $\mathbf{v}_i \in N_W(\mathbf{x}_i)$ and $N_W(\mathbf{x}_i) \cap A \neq \varnothing$ for $i = 1, 2$. Suppose $\mathbf{u}_i$ is in $N_W(\mathbf{x}_i) \cap A$, $i = 1, 2$. Since $\mathbf{u}_i$ is an inward normal of W at $\mathbf{x}_i$, and since W has constant width 1, the hyperplane $\{\mathbf{p} \in \mathbf{R}^n \mid \mathbf{p} \cdot \mathbf{u}_i = \mathbf{x}_i \cdot \mathbf{u}_i + 1\}$ is a supporting hyperplane of W. The only point on this hyperplane whose distiance from $\mathbf{x}_i$ is not greater than 1 is $\mathbf{x}_i + \mathbf{u}_i$. Since diameter $W = 1$, we conclude that $\mathbf{x}_i + \mathbf{u}_i \in \partial W$ for $i = 1, 2$. Therefore

$$1 = (\text{diameter } W)^2 \geq \|(\mathbf{x}_1 + \mathbf{u}_1) - \mathbf{x}_2\|^2 = \|\mathbf{x}_1 - \mathbf{x}_2\|^2 + 2\mathbf{u}_1 \cdot (\mathbf{x}_1 - \mathbf{x}_2) + 1$$

and

$$1 = (\text{diameter } W)^2 \geq \|(\mathbf{x}_2 + \mathbf{u}_2) - \mathbf{x}_1\|^2 = \|\mathbf{x}_2 - \mathbf{x}_1\|^2 + 2\mathbf{u}_2 \cdot (\mathbf{x}_2 - \mathbf{x}_1) + 1.$$

Summing these inequalities and rearranging we get

$$(\mathbf{u}_1 - \mathbf{u}_2) \cdot (\mathbf{x}_2 - \mathbf{x}_1) \geq \|\mathbf{x}_2 - \mathbf{x}_1\|^2.$$

Because $(\mathbf{u}_1 - \mathbf{u}_2) \cdot (\mathbf{x}_2 - \mathbf{x}_1) \leq \|\mathbf{u}_1 - \mathbf{u}_2\| \, \|\mathbf{x}_2 - \mathbf{x}_1\|$, this implies $\|\mathbf{u}_1 - \mathbf{u}_2\| \geq \|\mathbf{x}_2 - \mathbf{x}_1\|$. So

$$\|\mathbf{x}_2 - \mathbf{x}_1\| \leq \text{diameter } A. \tag{2.1}$$

As with $\mathbf{x}_i + \mathbf{u}_i$, the points $\mathbf{x}_i + \mathbf{v}_i$ lie in ∂W, and therefore

$$1 \geq \|(\mathbf{x}_1 + \mathbf{v}_1) - (\mathbf{x}_2 + \mathbf{v}_2)\| \geq \|\mathbf{v}_1 - \mathbf{v}_2\| - \|\mathbf{x}_2 - \mathbf{x}_1\|.$$

Using (2.1), this implies

$$1 + \text{diameter } A \geq \|\mathbf{v}_1 - \mathbf{v}_2\|.$$

We use μ to denote the standard probability measure on S^{n-1}. Define

$$g(n, d) = \inf \{\mu(A^+) \mid A \subset S^{n-1}, \text{ diameter } A \leq d\}.$$

Let $N(n, \varepsilon)$ be the number of sets having diameter ε that is required to cover S^{n-1}. The core of this note is:

PROPOSITION 6. *For $0 < \varepsilon < \sqrt{2} - 1$ we have*

$$I(W) \leq 1 + \frac{\log N(n, \varepsilon)}{-\log (1 - g(n, 1 + \varepsilon))}.$$

Proof. It is easily verified that $0 < g(n, 1 + \varepsilon) < 1$ (if $\varnothing \neq A \subset S^{n-1}$ then A^+ is contained in a hemisphere, so that $\mu(A^+) \leq \frac{1}{2}$. If also diameter $A = d < \sqrt{2}$ then A^+ contains a spherical cap of radius $\sqrt{2} - d$ around any point of A), so that the right-hand side is well defined. Let M be a natural number satisfying

$$M > \frac{\log N(n, \varepsilon)}{-\log (1 - g(n, 1 + \varepsilon))}.$$

It is sufficient to show that M directions can illuminate W. Set $N = N(n, \varepsilon)$, and let $A_1, \ldots, A_N$ be a covering of S^{n-1} with sets of diameter ε. By Proposition 5, we have diameter $U_W(A_i) \leq 1 + \varepsilon$, and therefore

$$\mu(U_W(A_i)^+) \geq g(n, 1 + \varepsilon), \qquad i = 1, 2, \ldots, N.$$

Pick M directions $\mathbf{u}_1, \ldots, \mathbf{u}_M$ at random, uniformly and independently distributed on S^{n-1}. Take any $i, j, 1 \leq i \leq N$, $1 \leq j \leq M$. The probability that $\mathbf{u}_j$ will be in $U_W(A_i)^+$ is $\mu(U_W(A_i)^+)$, which is at least $g(n, 1+\varepsilon)$. Therefore the probability that $U_W(A_i)^+$ will contain none of the points $\mathbf{u}_1, \ldots, \mathbf{u}_M$ is at most $(1-g(n, 1+\varepsilon))^M$. Thus the probability p that at least one $U_W(A_l)^+$, $1 \leq l \leq N$ will contain no points of $\mathbf{u}_1, \ldots, \mathbf{u}_M$ satisfies

$$p \leq \sum_{l=1}^{N} (1-g(n, 1+\varepsilon))^M < N(1-g(n, 1+\varepsilon))^{\log N/-\log(1-g(n,1+\varepsilon))} = 1.$$

This shows that one can choose M directions, so that each set $U_W(A_l)^+$, $l = 1, \ldots, N$, contains at least one of them. Let $\mathbf{v}_1, \ldots, \mathbf{v}_M$ be such directions, and let $\mathbf{x}$ be a point of ∂W. We claim that one of these directions illuminates $\mathbf{x}$. Since $N_W(\mathbf{x})$ is nonempty, and the sets $A_1, \ldots, A_N$ cover S^{n-1}, one of them, say A_i, intersects $N_W(\mathbf{x})$. By the definition of $U_W(A_i)$, we have $N_W(\mathbf{x}) \subset U_W(A_i)$. So that

$$N_W(\mathbf{x})^+ \supset U_W(A_i)^+.$$

$U_W(A_i)^+$ contains at least one of $\mathbf{v}_1, \ldots, \mathbf{v}_M$, say $\mathbf{v}_k$. We have

$$\mathbf{v}_k \in U_W(A_i)^+ \subset N_W(\mathbf{x})^+$$

and therefore, by Lemma 4, $\mathbf{v}_k$ illuminates $\mathbf{x}$. This shows that the directions $\mathbf{v}_1, \ldots, \mathbf{v}_M$ illuminate W.

In order to deduce Theorem 1 from Proposition 6, we only have to estimate $g(n, 1+\varepsilon)$ from below, and $N(n, \varepsilon)$ from above. The former will be done in the next section, and the latter is dealt with by the following well known fact.

LEMMA 7. $N(n, \varepsilon) \leq (1+4/\varepsilon)^n$.

Proof. Let E be a maximal subset of S^{n-1} having the property that $\|\mathbf{u}-\mathbf{v}\| > \frac{1}{2}\varepsilon$ for $\mathbf{u} \neq \mathbf{v}$ in E. The maximality of E shows that the balls with radius $\frac{1}{2}\varepsilon$ and centers in E cover S^{n-1}, therefore

$$|E| \geq N(n, \varepsilon).$$

All the balls $B(\mathbf{u}, \varepsilon/4)$, $\mathbf{u} \in E$, are disjoint and are contained in the ball $B(0, 1+\varepsilon/4)$. Comparing volumes gives

$$|E|(\varepsilon/4)^n \leq (1+\varepsilon/4)^n,$$

or

$$|E| \leq \left(\frac{4}{\varepsilon}+1\right)^n.$$

Remark. Better estimates for $N(n, \varepsilon)$ are known (see [9]), but do not seem to contribute any significant improvement to Theorem 1.

§3. In this section we give a lower bound for $g(n, d)$, and prove Theorem 1.

PROPOSITION 8. *Let $d>0$ and let A be a nonempty subset of S^{n-1} having diameter $\leq d$. Suppose $\mathbf{u}\in S^{n-1}$, $a>0$ and A is contained in the half-space $\{\mathbf{p}\in\mathbf{R}^n \mid \mathbf{p}\cdot\mathbf{u}\geq a\}$, then*

$$A^+\cup TA^+\supset D_0(\mathbf{u},\arctan(2a/d)),$$

where $T:\mathbf{R}^n\to\mathbf{R}^n$ is the reflection through the line determined by $\mathbf{u}$, $-\mathbf{u}$:

$$T\mathbf{p}=2(\mathbf{p}\cdot\mathbf{u})\mathbf{u}-\mathbf{p},$$

and $D_0(\mathbf{u},\psi)$ is the open spherical cap consisting of all unit vectors having an angle with $\mathbf{u}$, which is smaller than ψ.

Proof. Suppose $\mathbf{x}$ is a point in S^{n-1} but not in $A^+\cup TA^+$, and let θ be the angular distance between $\mathbf{x}$ and $\mathbf{u}$, $0\leq\theta\leq\pi$. Write

$$\mathbf{x}=(\cos\theta)\mathbf{u}+(\sin\theta)\mathbf{v}, \tag{3.1}$$

where $\mathbf{v}$ is a unit vector orthogonal to $\mathbf{u}$ (we ignore the trivial case $n=1$). Since $\mathbf{x}\notin A^+$, there is a point $\mathbf{y}\in A$ with

$$0\geq\mathbf{y}\cdot\mathbf{x}=\mathbf{y}\cdot\mathbf{u}\cos\theta+\mathbf{y}\cdot\mathbf{v}\sin\theta. \tag{3.2}$$

Since $T^{-1}=T$ and $\mathbf{x}\notin TA^+$ we have $T\mathbf{x}\notin A^+$. Thus there is a point $\mathbf{z}\in A$ with

$$0\geq\mathbf{z}\cdot T\mathbf{x}=\mathbf{z}\cdot\mathbf{u}\cos\theta-\mathbf{z}\cdot\mathbf{v}\sin\theta. \tag{3.3}$$

Summing (3.2) and (3.3), and using $\|\mathbf{y}-\mathbf{z}\|\leq d$, $\sin\theta\geq 0$, we have

$$0\geq(\mathbf{y}\cdot\mathbf{u}+\mathbf{z}\cdot\mathbf{u})\cos\theta+(\mathbf{y}-\mathbf{z})\cdot\mathbf{v}\sin\theta\geq(\mathbf{y}\cdot\mathbf{u}+\mathbf{z}\cdot\mathbf{u})\cos\theta-d\sin\theta. \tag{3.4}$$

Temporarily suppose that $\theta<\frac{1}{2}\pi$. Then $\cos\theta>0$, so by (3.4)

$$\tan\theta\geq\frac{\mathbf{y}\cdot\mathbf{u}+\mathbf{z}\cdot\mathbf{u}}{d}\geq\frac{2a}{d}.$$

The last inequality is justified by the hypothesis that $A\subset\{\mathbf{p}\in\mathbf{R}^n \mid \mathbf{p}\cdot\mathbf{u}\geq a\}$. Whether or not $\theta<\frac{1}{2}\pi$, we have $\theta\geq\arctan(2a/d)$. This shows that $A^+\cup TA^+\supset D_0(\mathbf{u},\arctan(2a/d))$.

PROPOSITION 9.

$$g(n,d)\geq\frac{1}{\sqrt{8\pi n}}\left(\frac{3}{2}+\frac{(2n+1)d^2-(2n+2)}{4n+4-2d^2n}\right)^{-(n-1)/2}\qquad \text{for } 0<d\leq\sqrt{2}.$$

Proof. Let $0<d\leq\sqrt{2}$ and let A be a nonempty subset of S^{n-1} having diameter $\leq d$. By Jung's theorem [**5**, p. 111], there is, in $\mathbf{R}^n$, a ball with radius $d\sqrt{n/(2n+2)}$ containing A. Let $\mathbf{q}$ be the center of this ball. Write $\mathbf{q}=t\mathbf{u}$ with $t\geq 0$ and $\mathbf{u}\in S^{n-1}$. For every $\mathbf{x}\in A$ we have

$$d^2\frac{n}{2n+2}\geq\|\mathbf{x}-\mathbf{q}\|^2=\|\mathbf{x}-t\mathbf{u}\|^2=1-2t\mathbf{x}\cdot\mathbf{u}+t^2. \tag{3.5}$$

Since $1-2t\mathbf{x}\cdot\mathbf{u}+t^2=1-(\mathbf{x}\cdot\mathbf{u})^2+(\mathbf{x}\cdot\mathbf{u}-t)^2\geq 1-(\mathbf{x}\cdot\mathbf{u})^2$, inequality (3.5)

implies

$$d^2 \frac{n}{2n+2} \geq 1-(\mathbf{x} \,.\, \mathbf{u})^2. \tag{3.6}$$

From (3.5), $t \geq 0$ and $d \leq \sqrt{2}$ it can be seen that $\mathbf{x} \,.\, \mathbf{u} \geq 0$, so, using (3.6), we obtain

$$\mathbf{x} \,.\, \mathbf{u} \geq \sqrt{1-d^2 \frac{n}{2n+2}}.$$

Set

$$a = \sqrt{1-d^2 \frac{n}{2n+2}}. \tag{3.7}$$

The above argument shows that A is contained in the half-space

$$\{\mathbf{p} \in \mathbf{R}^n \,|\, \mathbf{p} \,.\, \mathbf{u} \geq a\}.$$

Proposition 8 can be applied, yielding

$$A^+ \cup TA^+ \supset D_0(\mathbf{u}, \arctan(2a/d)).$$

Now, since T is an orthogonal transformation, we have $\mu(A^+) = \mu(TA^+)$, and

$$\begin{aligned} \mu(A^+) &= \tfrac{1}{2}(\mu(A^+) + \mu(TA^+)) \geq \tfrac{1}{2}\mu(A^+ \cup TA^+) \\ &\geq \tfrac{1}{2}\mu\left(D_0\left(\mathbf{u}, \arctan\frac{2a}{d}\right)\right) = \frac{1}{2}\,\frac{\mathrm{Vol}_{n-1}\, D_0(\mathbf{u}, \arctan 2a/d)}{\mathrm{Vol}_{n-1}\, S^{n-1}} \\ &= \frac{\mathrm{Vol}_{n-1}\, D_0(\mathbf{u}, \arctan 2a/d)}{2n\Omega_n}. \end{aligned} \tag{3.8}$$

Here Ω_n denotes the volume of the n-dimensional unit ball, and Vol_{n-1} is the $(n-1)$-dimensional volume.

Let D' be the orthogonal projection of $D_0(\mathbf{u}, \arctan 2a/d)$ to the hyperplane $\{\mathbf{p} \in \mathbf{R}^n \,|\, \mathbf{p} \,.\, \mathbf{u} = 0\}$. Obviously we have

$$\mathrm{Vol}_{n-1}\, D_0\left(\mathbf{u}, \arctan\frac{2a}{d}\right) \geq \mathrm{Vol}_{n-1}\, D'. \tag{3.9}$$

D' is an $(n-1)$-dimensional ball having radius

$$\sin\left(\arctan\frac{2a}{d}\right) = \left(1+\frac{d^2}{4a^2}\right)^{-1/2},$$

so

$$\mathrm{Vol}_{n-1}\, D' = \Omega_{n-1}\left(1+\frac{d^2}{4a^2}\right)^{-(n-1)/2}. \tag{3.10}$$

Using (3.8), (3.9), (3.10), we get

$$\mu(A^+) \geq \frac{\Omega_{n-1}}{2n\Omega_n}\left(1+\frac{d^2}{4a^2}\right)^{-(n-1)/2}. \tag{3.11}$$

Now

$$\frac{\Omega_{n-1}}{\Omega_n}=\frac{\pi^{(n-1)/2}/\Gamma((1+n)/2)}{\pi^{n/2}/\Gamma(1+n/2)}=\frac{\Gamma(1+n/2)}{\sqrt{\pi}\Gamma((1+n)/2)} \tag{3.12}$$

where Γ is the Gamma function. Since $\log\Gamma$ is convex ([1, p. 12]), we have

$$\Gamma(1+n/2)\Gamma(n/2)\geqslant\Gamma((1+n)/2)^2,$$

and therefore

$$\begin{aligned}\frac{\Gamma(1+n/2)}{\Gamma((1+n)/2)}&\geqslant\frac{\Gamma(1+n/2)}{\Gamma((1+n)/2)}\sqrt{\frac{\Gamma((1+n)/2)^2}{\Gamma(1+n/2)\Gamma(n/2)}}\\&=\sqrt{\frac{\Gamma(1+n/2)}{\Gamma(n/2)}}=\sqrt{\frac{n}{2}}.\end{aligned} \tag{3.13}$$

Using (3.11), (3.12), (3.13), we get

$$\mu(A^+)\geqslant\frac{1}{2n}\frac{1}{\sqrt{\pi}}\sqrt{\frac{n}{2}}\left(1+\frac{d^2}{4a^2}\right)^{-(n-1)/2}.$$

And after substituting the value of a,

$$\begin{aligned}\mu(A^+)&\geqslant\frac{1}{\sqrt{8\pi n}}\left(1+\frac{d^2}{4-2d^2n/(n+1)}\right)^{-(n-1)/2}\\&=\frac{1}{\sqrt{8\pi n}}\left(\frac{3}{2}+\frac{(2n+1)d^2-2n-2}{4n+4-2d^2n}\right)^{-(n-1)/2}.\end{aligned}$$

Now proving Theorem 1 is just a matter of putting the pieces together.

Proof of Theorem 1. Since $t<-\log(1-t)$ for $0<t<1$, Proposition 6 implies

$$I(W)<1+\frac{\log N(n,\varepsilon)}{g(n,1+\varepsilon)},\qquad 0<\varepsilon<\sqrt{2}-1.$$

Choose

$$\varepsilon=\sqrt{1+\frac{1}{2n+1}}-1.$$

From Lemma 7 and Proposition 9 we get

$$\begin{aligned}I(W)<1+\frac{\log N(n,\varepsilon)}{g(n,1+\varepsilon)}&\leqslant 1+\frac{\log\left(1+\frac{4}{\varepsilon}\right)^n}{g\left(n,\sqrt{\frac{2n+2}{2n+1}}\right)}\\&\leqslant 1+\sqrt{8\pi n}\left(\frac{3}{2}\right)^{(n-1)/2}n\log\left(1+\frac{4}{\varepsilon}\right)\\&=1+4n\sqrt{\pi n/3}\left(\frac{3}{2}\right)^{n/2}\log\left(1+\frac{4}{\varepsilon}\right).\end{aligned}$$

Since one easily sees that $\varepsilon > 1/(4n+3)$, we have

$$I(W) < 1+4n\sqrt{\pi n/3}\log(13+16n)\left(\frac{3}{2}\right)^{n/2} \leq 5n\sqrt{n}(4+\log n)\left(\frac{3}{2}\right)^{n/2}$$

Remarks. 1. In [**11**] we give a lower bound for the volumes of sets of constant width in $\mathbf{R}^n$, using results analogous to Propositions 8 and 9.

2. The factor $5n\sqrt{n}(4+\log n)$ in Theorem 1 should not be taken seriously. It can be improved with some more careful estimates. However, any improvement of the exponential factor, $(3/2)^{n/2}$, would be interesting. A possible way to do this may be to try to get a better lower bound for $g(n, d)$. An advance in this direction may lead to better estimates for the minimal volume of a set having constant width 1 in $\mathbf{R}^n$.

§4.

THEOREM 10. *Let $K \subset \mathbf{R}^n$ be a strictly convex body, invariant under a group of orthogonal transformations G that is generated by reflections through hyperplanes and acts irreducibly on $\mathbf{R}^n$. Then $I(K) = n+1$.*

We preface the proof with a few definitions and a lemma. A unit vector $\mathbf{r}$ is called a *root* of G if the orthogonal reflection through the subspace orthogonal to $\mathbf{r}$ is an element of G. We denote this reflection by $S_{\mathbf{r}}$:

$$S_{\mathbf{r}}\mathbf{x} = \mathbf{x} - 2(\mathbf{x}\,.\,\mathbf{r})\mathbf{r}, \qquad \mathbf{x} \in \mathbf{R}^n.$$

If $\mathbf{v}, \mathbf{r}$ are roots of G then $S_{\mathbf{r}}\mathbf{v}$ is also a root of G, because $S_{S_{\mathbf{r}}\mathbf{v}} = S_{\mathbf{r}}S_{\mathbf{v}}S_{\mathbf{r}}$. In particular $-\mathbf{r} = S_{\mathbf{r}}\mathbf{r}$ is a root.

LEMMA 11. *Let G be as in Theorem 10. If $n > 1$ then there are $n+1$ roots of G, $\mathbf{r}_0, \ldots, \mathbf{r}_n$, such that every nonzero $\mathbf{x} \in \mathbf{R}^n$ has a negative inner product with at least one of them.*

At least when G is finite this follows from known results (see [**6**]).

Proof of the Lemma. We will say that a set of vectors $\{\mathbf{v}_0, \ldots, \mathbf{v}_m\}$ is *almost independent* if every proper subset of it is linearly independent but $\{\mathbf{v}_0, \ldots, \mathbf{v}_m\}$ is linearly dependent. It is easily checked that $\{\mathbf{v}_0, \ldots, \mathbf{v}_m\}$ is almost independent, if, and only if, $\mathbf{v}_1, \ldots, \mathbf{v}_m$ are linearly independent and $\mathbf{v}_0$ is a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_m$ with *nonzero* coefficients.

Because of the hypotheses on G and because $n \geq 2$, G has at least one root, $\mathbf{r}$. $\{\mathbf{r}, -\mathbf{r}\}$ is an almost independent set. Suppose $\{\mathbf{r}_0, \ldots, \mathbf{r}_m\}$ is the largest almost independent set of roots of G. Let U be the subspace generated by $\mathbf{r}_0, \ldots, \mathbf{r}_m$. We claim that $U = \mathbf{R}^n$ and therefore $n = m$. Let $\mathbf{r}$ be a root of G and suppose that U is not invariant under $S_{\mathbf{r}}$. This implies $\mathbf{r} \notin U$ and $S_{\mathbf{r}}\mathbf{r}_i \neq \mathbf{r}_i$ for some $i = 0, 1, \ldots, m$. Assume, without loss of generality, that $S_{\mathbf{r}}\mathbf{r}_0 \neq \mathbf{r}_0$. $\mathbf{r}_0$ is a linear combination of $\mathbf{r}_1, \ldots, \mathbf{r}_m$ with nonzero coefficients. Since $S_{\mathbf{r}}\mathbf{r}_0 - \mathbf{r}_0$ is a nonzero multiple of $\mathbf{r}$, $S_{\mathbf{r}}\mathbf{r}_0$ is a linear combination of $\mathbf{r}, \mathbf{r}_1, \ldots, \mathbf{r}_m$ with nonzero coefficients. Because $\mathbf{r}, \mathbf{r}_1, \ldots, \mathbf{r}_m$ are linearly independent, this means

that the set $\{S_r\mathbf{r}_0, \mathbf{r}, \mathbf{r}_1, \ldots, \mathbf{r}_m\}$ is almost independent, contradicting the choice of $\{\mathbf{r}_0, \ldots, \mathbf{r}_m\}$.

This forces us to conclude that U is invariant under the reflections generating G, and therefore under every transformation in G. G acts irreducibly on $\mathbf{R}^n$ and $U \neq 0$. This implies $U = \mathbf{R}^n$ and $n = m$.

Because $\{\mathbf{r}_0, \ldots, \mathbf{r}_n\}$ is almost independent, it is possible to write

$$0 = \sum_{i=0}^{n} a_i \mathbf{r}_i$$

with all coefficients nonzero. We replace some of the roots $\mathbf{r}_i$ by their negatives, to have all the coefficients a_i positive. If $\mathbf{x}$ is any nonzero vector, we must have $\mathbf{x} \cdot \mathbf{r}_i \neq 0$ for some $i = 0, 1, \ldots, n$, because the $\mathbf{r}_i$ span $\mathbf{R}^n$. Since $0 = \mathbf{x} \cdot 0 = \sum_{i=0}^{n} a_i(\mathbf{x} \cdot \mathbf{r}_i)$ and $a_i > 0$, we must have $\mathbf{x} \cdot \mathbf{r}_i < 0$ for some i.

Proof of Theorem 10. As mentioned in Section 1, $I(K) \geqslant n+1$ holds for every convex body in $\mathbf{R}^n$, thus we only need to show that $I(K) \leqslant n+1$. The Theorem is obviously true when $n = 1$, so we assume $n > 1$.

Let $\mathbf{r}_0, \ldots, \mathbf{r}_n$ be the roots of G guaranteed by the lemma and let $\mathbf{x} \in \partial K$. First observe that the origin is necessarily an interior point of K so $\mathbf{x} \neq 0$. For some $\mathbf{r}_i$, $\mathbf{x} \cdot \mathbf{r}_i < 0$. We claim that this $\mathbf{r}_i$ illuminates $\mathbf{x}$. $\mathbf{x} - 2(\mathbf{x} \cdot \mathbf{r}_i)\mathbf{r}_i = S_{\mathbf{r}_i}\mathbf{x} \in K$, since $S_{\mathbf{r}_i} \in G$. For some positive t, $\mathbf{x} + t\mathbf{r}_i$ is an interior point of K, because $\mathbf{x}$, $\mathbf{x} - 2(\mathbf{x} \cdot \mathbf{r}_i)\mathbf{r}_i \in K$, K is strictly convex and $-2(\mathbf{x} \cdot \mathbf{r}_i) > 0$. This verifies the claim and we see that $\mathbf{r}_0, \ldots, \mathbf{r}_n$ illuminate K.

Remark. A set of constant width is strictly convex; therefore Theorem 10 applies to (sufficiently symmetric) sets of constant width.

References

1. E. Artin. *The Gamma Function*, translated from German (Holt, Rinehart and Winston, 1964).
2. V. G. Boltjansky. The problem of the illumination of the boundary of a convex body. In Russian. *Izvestiya Moldavskogo filiala Akademii Nauk SSSR*, 10 (1960), 77-84.
3. V. G. Boltjansky and I. Ts. Gohberg. *Results and problems in combinatorial geometry*, translated from Russian (Cambridge Univ. Press, Cambridge, 1985).
4. G. D. Chakerian and H. Groemer. Convex bodies of constant width, in "*Convexity and its applications*," ed. by P. M. Gruber and J. M. Wills (Birkhäuser, Basel, 1983).
5. H. G. Eggleston. *Convexity* (Cambridge Univ. Press, Cambridge, 1958).
6. L. C. Grove and C. T. Benson. *Finite Reflection Groups* (Springer, New York, 1985).
7. B. Grünbaum. Borsuk's problem and related questions. *Proc. Symp. Pure Math.*, 7 (1963), 271-284.
8. B. Grünbaum. *Convex Polytopes* (Wiley, New York, 1967).
9. C. A. Rogers. Covering a sphere with spheres. *Mathematika*, 10 (1963), 157-164.
10. C. A. Rogers. Symmetrical sets of constant width and their partitions. *Mathematika*, 18 (1971), 105-111.
11. O. Schramm. On the volume of sets having constant width. To appear, *Israel J. Math.*

Dr. O. Schramm,
Mathematics Department,
Fine Hall,
Princeton University,
Princeton NJ 08544,
USA.

52A20. *CONVEX SETS AND RELATED GEOMETRIC TOPICS; Convex sets in n dimensions.*

Received on the 15th of February, 1988.

ISRAEL JOURNAL OF MATHEMATICS, Vol. 63, No. 2, 1988

ON THE VOLUME OF SETS HAVING CONSTANT WIDTH†

BY

ODED SCHRAMM

Mathematics Department, Fine Hall, Princeton University, Princeton, NJ 08544, USA

ABSTRACT

A lower bound is given for the volume of sets of constant width.

1. Introduction

A set of constant width d in Euclidean space $\mathbf{R}^n$ is a compact, convex set, such that the distance between any distinct, parallel supporting hyperplanes of it is d (see [3, pp. 122–131], [2]).

The Blaschke–Lebesgue theorem states that of all planar sets having constant width d the Reuleaux triangle has the least area, $\frac{1}{2}(\pi - \sqrt{3})d^2$. The problem of determining the minimal volume of sets having constant width d in $\mathbf{R}^n$, $n > 2$, seems considerably more difficult. Lower bounds for it have been given by Firey [4] and Chakerian [1].

Let W be a set of constant width d and circumradius r in $\mathbf{R}^n$. In this note we prove the lower bound

$$\text{Vol } W \geqq \left(\sqrt{5 - 4\frac{r^2}{d^2}} - 1 \right)^n \text{Vol } B(0, d/2), \tag{1.1}$$

which implies

$$\text{Vol } W \geqq \left(\sqrt{3 + \frac{2}{n+1}} - 1 \right)^n \text{Vol } B(0, d/2). \tag{1.2}$$

† The research exposed in this note was done while I was at the Hebrew University of Jerusalem, as a student of Professor Gil Kalai. I would like to thank Prof. Kalai for his interest, encouragement and advice.

Received February 22, 1988

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_3, © Springer Science+Business Media, LLC 2011

Here Vol denotes the n-dimensional volume in $\mathbf{R}^n$ and $B(\mathbf{x}, \rho)$ is the ball having center $\mathbf{x}$ and radius ρ. This bound is, for $n > 4$, an improvement over those previously known.

We also prove

THEOREM 1. *Let K be a set of constant width d and circumradius r in $\mathbf{R}^n$ having the origin 0 as the center of its circumsphere, then $K \cup -K$ contains the ball of radius $\sqrt{5(d/2)^2 - r^2} - (d/2)$ around the origin.*

This result can be seen as a relative to the well known theorem stating that the insphere of a set of constant width d is concentric to the circumsphere and its radius is $d - r$, where r is the circumradius (see [3, p. 125]).

Arguments analogous to those below, but dealing with subsets of the unit sphere, are used in [5], where an upper bound is given for the number of directions sufficient to illuminate the boundary of sets having constant width.

2. For a set $A \subset \mathbf{R}^n$ and for $\lambda > 0$ we denote by A^λ the intersection of all the balls of radius λ, having centers in A:

$$A^\lambda = \bigcap_{\mathbf{x} \in A} B(\mathbf{x}, \lambda) = \{\mathbf{p} \in \mathbf{R}^n \mid B(\mathbf{p}, \lambda) \supset A\}.$$

We also use

$$h(A, \mathbf{x}) = \sup_{\mathbf{y} \in A} \mathbf{y} \cdot \mathbf{x} \qquad \text{(the support function of } A\text{)},$$

$$\rho(A, \mathbf{x}) = \inf\{t > 0 \mid t\mathbf{x} \notin A\}.$$

Define

$$g(\lambda, r, t) = \sqrt{\lambda^2 - r^2 + t^2} - t.$$

Notice that $g(\lambda, r, t)$ is monotonic decreasing, positive and strictly convex as a function of t when $\lambda > r$.

LEMMA 1. *Let K be a nonempty set contained in the ball of radius r around the origin in $\mathbf{R}^n$, then the relation*

$$\rho(K^\lambda, \mathbf{u}) \geqq g(\lambda, r, h(K, -\mathbf{u})) \tag{2.1}$$

is satisfied for every $\lambda \geqq r$ and every $\mathbf{u} \in S^{n-1}$.

PROOF. Let $\mathbf{u}$ be any unit vector, let λ satisfy $\lambda \geqq r$ and let a be the right hand side of (2.1). We first show that $a\mathbf{u} \in K^\lambda$. Let $\mathbf{x}$ be any point of K. We have

$$\| \mathbf{x} \| \leqq r, \qquad -\mathbf{x} \cdot \mathbf{u} \leqq h(K, -\mathbf{u}).$$

Using this and $a \geqq 0$ we obtain

$$\| \mathbf{x} - a\mathbf{u} \|^2 = \| \mathbf{x} \|^2 - 2a\mathbf{x} \cdot \mathbf{u} + a^2 \leqq r^2 + 2ah(K, -\mathbf{u}) + a^2 = \lambda^2.$$

This means that $a\mathbf{u} \in B(\mathbf{x}, \lambda)$ and since $\mathbf{x}$ is an arbitrary point of K, we have

$$a\mathbf{u} \in \bigcap_{\mathbf{x} \in K} B(\mathbf{x}, \lambda) = K^\lambda.$$

The origin is also a point of K^λ, because $K \subset B(0, r)$ and $\lambda \geqq r$. K^λ is obviously convex, so we have

$$\{t\mathbf{u} \mid 0 \leqq t \leqq a\} \subset K^\lambda.$$

This shows that $\rho(K^\lambda, \mathbf{u}) \geqq a$, as needed. ∎

In some contexts, a good way to present the volume of a set $K \subset \mathbf{R}^n$ is to specify the radius of the ball having the same volume as K. We will call it the *effective radius* of the set K and denote it by $\operatorname{er} K$:

$$\operatorname{Vol} K = \operatorname{Vol} B(0, \operatorname{er} K).$$

By μ we denote the $n-1$ dimensional surface area measure on S^{n-1}, the boundary of the unit ball.

THEOREM 2. *Let K be a set of diameter d and circumradius r. Let λ satisfy $\lambda > r$, then*

$$\operatorname{er} K^\lambda \geqq g(\lambda, r, d/2).$$

PROOF. As we know, K^λ contains the origin and is convex. We can therefore rewrite its volume thus:

$$\operatorname{Vol} K^\lambda = \frac{1}{n} \int_{S^{n-1}} \rho(K^\lambda, \mathbf{u})^n d\mu(\mathbf{u}).$$

Using the lemma we have

$$\operatorname{Vol} K^\lambda \geqq \frac{1}{n} \int_{S^{n-1}} g(\lambda, r, h(K, -\mathbf{u}))^n d\mu(\mathbf{u})$$

$$= \frac{1}{n} \int_{S^{n-1}} (\tfrac{1}{2} g(\lambda, r, h(K, -\mathbf{u}))^n + \tfrac{1}{2} g(\lambda, r, h(K, \mathbf{u}))^n) d\mu(\mathbf{u}).$$

Since $g(\lambda, r, t)$ is positive and convex in t, so is $g(\lambda, r, t)^n$. Therefore the above inequality implies

$$\text{(2.2)} \qquad \operatorname{Vol} K^\lambda \geqq \frac{1}{n} \int_{S^{n-1}} g(\lambda, r, \tfrac{1}{2}h(K, -\mathbf{u}) + \tfrac{1}{2}h(K, \mathbf{u}))^n d\mu(\mathbf{u}).$$

Since K has diameter d, we have

$$\tfrac{1}{2}h(K, -\mathbf{u}) + \tfrac{1}{2}h(K, \mathbf{u}) \leqq \tfrac{1}{2}d.$$

From (2.2) and the decreasing monotonicity of $g(\lambda, r, t)$ in t, we can therefore conclude that

$$\operatorname{Vol} K^\lambda \geqq \frac{1}{n} \int_{S^{n-1}} g(\lambda, r, \tfrac{1}{2}d)^n d\mu(\mathbf{u}) = \operatorname{Vol} B(0, g(\lambda, r, d/2)). \qquad \blacksquare$$

PROOF OF (1.1), (1.2). Since $K^d = K$ for sets of constant width d (see [3, p. 123]), (1.1) can be derived easily from Theorem 2 using $\lambda = d$, $W = K$. (1.2) is a consequence of (1.1) and Jung's Theorem $r \leqq d\sqrt{n/(2n+2)}$ (see [3, p. 111]). $\blacksquare$

Let us denote by r_n the minimal effective radius of all sets having constant width two[†] in $\mathbf{R}^n$. (1.2) is equivalent to $r_n \geqq \sqrt{3 + 2/(n+1)} - 1$. From the proof it is evident that equality does not occur when $n > 1$. As mentioned above, the exact computation of r_n, for $n \geqq 3$, is probably very hard, however, the following problems seem to be answerable.

PROBLEM 1. Is the sequence $\{r_n\}$ monotonic decreasing?

PROBLEM 2. Show that $\lim_{n\to\infty} r_n$ exists and compute it.

Inequality (1.2) shows that $\liminf r_n \geqq \sqrt{3} - 1$. Because the unit ball has the largest volume among all sets having constant width 2 (see [3, pp. 106–107]), we have $\limsup r_n \leqq 1$. As far as we know any value between $\sqrt{3} - 1$ and 1 is a possible candidate for $\lim r_n$. (If the answer to Problem 1 is 'yes' then surely $\limsup r_n \leqq r_2 < 1$.)

3. We now prove a generalization of Theorem 1.

[†] The Blaschke selection principle implies the minimum is attained.

THEOREM 3. *Let K be a set of diameter d contained in the ball $B(0, r)$ in $\mathbf{R}^n$. Let λ satisfy $\lambda > r$, then*

$$K^\lambda \cup -K^\lambda \supset B(0, g(\lambda, r, d/2)).$$

PROOF. Let $\mathbf{u} \in S^{n-1}$. Because K has diameter d, we have $h(K, \mathbf{u}) + h(K, -\mathbf{u}) \leqq d$. Therefore

$$\min\{h(K, \mathbf{u}), h(K, -\mathbf{u})\} \leqq \tfrac{1}{2}d.$$

Using obvious properties of $\rho(\cdot, \cdot)$, Lemma 1 and the fact that $g(\lambda, r, t)$ is monotonic decreasing in t, we get

$$\begin{aligned}
\rho(K^\lambda \cup -K^\lambda, \mathbf{u}) &\geqq \max\{\rho(K^\lambda, \mathbf{u}), \rho(-K^\lambda, \mathbf{u})\} \\
&= \max\{\rho(K^\lambda, \mathbf{u}), \rho(K^\lambda, -\mathbf{u})\} \\
&\geqq \max\{g(\lambda, r, h(K, -\mathbf{u})), g(\lambda, r, h(K, \mathbf{u}))\} \\
&= g(\lambda, r, \min\{h(K, \mathbf{u}), h(K, -\mathbf{u})\}) \\
&\geqq g(\lambda, r, d/2).
\end{aligned}$$

This proves $K^\lambda \cup -K^\lambda \supset B(0, g(\lambda, r, d/2))$ as needed. ∎

PROOF OF THEOREM 1. Since $K^d = K$, using Theorem 3 with $\lambda = d$ gives Theorem 1.

REFERENCES

1. G. D. Chakerian, *Sets of constant width*, Pacific J. Math. **19** (1966), 11–21.
2. G. D. Chakerian and H. Groemer, *Convex bodies of constant width*, in *Convexity and its Applications* (P. M. Gruber and J. M. Wills, eds.), Birkhäuser, Basel, 1983, pp. 49–96.
3. H. G. Eggleston, *Convexity*, Cambridge Univ. Press, 1958.
4. W. J. Firey, *Lower bounds for volumes of convex bodies*, Arch. Math. **16** (1965), 69–74.
5. O. Schramm, *Illuminating sets of constant width*, Mathematika, to appear.

JOURNAL OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 4, Number 1, January 1991

RIGIDITY OF INFINITE (CIRCLE) PACKINGS

ODED SCHRAMM

1. Introduction

The nerve of a packing is a graph that encodes its combinatorics. The vertices of the nerve correspond to the packed sets, and an edge occurs between two vertices in the nerve precisely when the corresponding sets of the packing intersect.

The nerve of a circle packing and other well-behaved packings, on the sphere or in the plane, is a planar graph. It was an observation of Thurston [Th1, Chapter 1; 13, Th2] that Andreev's theorem [An1, An2] implies that given a finite planar graph, there exists a packing of (geometric) circles on the sphere whose nerve is the given graph. We refer to this fact as the circle packing theorem. The circle packing theorem also has a uniqueness part to it: if the graph is actually (the 1-skelaton of) a triangulation, then the circle packing is unique up to Möbius transformations.

Using the circle packing theorem, Thurston proposed a method for constructing approximate maps from a given bounded planar simply connected domain to the unit disk, and conjectured that this procedure approximates the corresponding Riemann mapping. Rodin and Sullivan proved this conjecture in [RS]. One of the crucial elements in their argument is the rigidity of the hexagonal packing, the infinite packing of circles where every circle touches six others and all have the same size. (When we say that a circle packing is rigid, we mean that any other circle packing on the sphere with the same nerve is Möbius equivalent to it. In the case of the hexagonal packing, this is the same as saying that any two *planar* circle packings with the combinatorics of the hexagonal packing are similar.) The techniques of [RS], and the rigidity of the hexagonal packing have since been used by others [CR, He2, Ro1, Ro2, Sch1] to obtain quasiconformal and conformal maps, and to study the quality of the convergence of Thurston's approximating scheme.

The existence part of the circle packing theorem is not hard to generalize to infinite, locally finite graphs, using a geometric limit. However, the uniqueness does not hold for arbitrary (locally finite) planar triangulations. In this note we

Received by the editors November 29, 1990.
1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 30C35, 52A45; Secondary 05B40, 30C60, 30C85.

©1991 American Mathematical Society
0894-0347/91 $1.00 + $.25 per page

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_4, © Springer Science+Business Media, LLC 2011

prove:

1.1. **Rigidity Theorem.** *Let $T = T(V)$ be an infinite, planar triangulation, and let $Q = (Q_v : v \in V)$ be a circle packing on the sphere S^2 with nerve T. If $S^2 - \text{carrier}(Q)$ is at most countable, then every other circle packing on the sphere with nerve T is Möbius equivalent to Q.* (The carrier of Q is the union of all the circles Q_v and their interiors, together with the curved triangular regions in the complement of the packing.)

For example, the carrier of the hexagonal packing is the whole plane, and its complement with respect to the sphere consists of a single point, the point at infinity. Therefore the rigidity of the hexagonal packing follows from Theorem 1.1.

Similar results are proved for circle packings that almost fill an open geometric disk, and for packings consisting of convex shapes. See Theorems 5.1, 5.3, 5.4.

The original proof of the rigidity of the hexagonal packing had relied on some rather heavy machinery. Later, He [He1] obtained a more direct proof with some accurate estimates. Both proofs use the theory of quasiconformal maps. The techniques of [RS] and [He1] can be used to yield rigidity results for packings with bounded valence (that is, the nerve has bounded valence) that fill the plane, but it seems unlikely that they would be sufficient to prove the rigidity theorem above. Recently, Stephenson [Ste] announced a rigidity proof that uses probabilistic arguments, but the assumptions there are even more restrictive than bounded valence.

The methods used here are mostly elementary plane topology arguments, together with simple manipulations with the Möbius group. You will find no formulae or inequalities in this paper. (But admittedly, some basic ideas of the theory of quasiconformal maps do play a role, in analogy, and in motivation. Conversely, the results here give a new perspective to the concept of the conformal modulus.) Our main tool is a theorem that we call the Incompatibility Theorem (3.1). It is a theorem about packings of sets more general than circles. In fact, the theorem is purely topological — no geometry is present in its hypotheses or its conclusions. The generality of the Incompatibility Theorem might seem a little out of place to the reader, and warrants an explanation of the background and the motivation.

In [Sch1] we generalized the existence part of the circle packing theorem to packings by more general shapes, in particular to packings of convex sets specified up to homothety, and of balls of Riemannian metrics. To give an illustration, we quote:

Convex Packing Theorem. *Let $G = G(V, E)$ be a planar graph, and for each vertex $v \in V$, let P_v be some smooth planar convex body. Then there exists a packing $(P'_v : v \in V)$ in the plane, with nerve G, and with each P'_v positively homothetic to P_v.*

The fact that P_v' is positively homothetic to P_v means that P_v' is obtained from P_v by a positive homothety; i.e., a transformation of the form $x \to ax+b$ for some $b \in \mathbb{R}^2$, $a \in \mathbb{R}$, $a > 0$.

The incompatibility theorem was conceived in an effort to understand the extent to which these more general packings are unique. In fact, the incompatibility theorem does give a rather painless proof for uniqueness, as well as existence, for packings under very general conditions; see [Sch2].

As the question of uniqueness of infinite circle packings is now answered in quite general circumstances, one is naturally led to the problem of existence. Consider an infinite planar triangulation T, and assume that T has just one end. (In the complement of any finite collection of vertices of T precisely one connected component is infinite.) It is easy to obtain a circle packing in the plane with nerve T by taking geometric limits of finite packings. But one can say more. There is such a packing whose carrier is either the plane or the open unit disk. For no such triangulation are both of these possibilities feasible and, in any case, the packing is unique up to Möbius transformations. The existence part will be proved elsewhere, and the uniqueness part follows from the results of this paper. (This proves a conjecture of Thurston [Th2].)

2. Definitions and notations

We consider the plane $\mathbb{R}^2 = \mathbb{C}$ as being contained in the sphere $\hat{\mathbb{C}} = S^2$. A circle in the plane is also a circle when thought of as a subset of S^2. When we use the term 'circle', we usually mean 'the circle together with its interior'; that is, a closed round disk.

For us, a *packing* means an indexed collection $P = (P_v : v \in V)$ of compact connected sets in the sphere S^2 with the property that the interior of each set P_v is disjoint from the other sets P_w, $w \neq v$.

The *nerve* of the packing $P = (P_v : v \in V)$ is a graph whose vertex set is V, and that is defined by the property that there is an edge between two distinct vertices, v, w, if and only if the corresponding sets, P_v, P_w, intersect. Note that there is at most a single edge between v, w in the nerve, even if P_v and P_w intersect in more than one place. If G is a graph, then we use the notation $G(V)$ to mean that V is the vertex set of G.

For any $A \subset S^2$ let $A^c = S^2 - A$ denote the complement of A.

We deal with packings of sets that are well behaved topologically. Most of the sets we pack are disklike: a set $A \subset S^2$ is *disklike* if it is the closure of its interior, its interior is connected, and the complement of any connected component of A^c is a topological disk. The obvious example for a disklike set is a topological disk. A more general example is a closed ball for a path metric on S^2 (a metric in which the infimum of the lengths of paths joining any two points is the distance between them).

Let $P = (P_v : v \in V)$ be a packing of disklike sets on the sphere, and let G be its nerve. If the intersection of any three of the sets P_v is empty, then

the nerve of the packing is a planar graph; that is, it can be embedded in the plane. To find such an embedding one picks an interior point p_v in every P_v to correspond to the vertex v, and one chooses a simple path from p_v to p_w in $P_v \cup P_w$ to correspond to an edge $v \leftrightarrow w$ in G. With a little care one makes sure that these curves intersect only at the vertices.

If P satisfies the condition that the intersection of any three sets of P is empty, then P is called a *nondegenerate* packing. Otherwise, it is called a *degenerate* packing.

An embedding of a connected locally finite graph in the sphere is called a *triangulation* if the boundary of every connected component of the complement of the graph, with respect to the sphere, either consists of three edges of the graph or does not intersect the embedded graph at all[1]. The latter situation can occur only if the graph is infinite. It is easy to see that a connected locally finite graph embedded in the sphere is a triangulation if and only if the set of neighbors of every vertex is the set of vertices of a simple closed path in the graph. (We assume that there are no multiple edges or loops, and that the graph has more than three vertices.) Because this condition does not depend on the particular embedding, it follows that a planar graph is a triangulation in every embedding if it is a triangulation in some embedding. Furthermore, the embedding is unique up to homeomorphisms of the sphere (using the fact that a cycle of an embedded graph separates the sphere). Thus we freely identify the graph with the embedded graph, and we say that the graph itself is a triangulation. A connected component of the complement of the embedded graph whose boundary consists of three edges is called a *triangle* of the graph. Generally we ignore triangulations with fewer than four vertices.

Let G be a finite planar graph, and let B be a simple closed path in G. We say that G is a triangulation with boundary B if $G - B$ is connected and if G has an embedding I in the sphere S^2 where all the connected components of $S^2 - I(G)$ are triangles (that is, bounded by three edges of $I(G)$), except possibly for one component whose boundary is $I(B)$. If B has precisely four vertices, then G is called a *triangulation of a quadrilateral.* The vertices of B are the *boundary vertices* of G.

A *quadrilateral* is a closed topological disk D in S^2 with four distinguished points p_0, p_1, p_2, p_3 on its boundary that are oriented clockwise with respect to the interior of D. D_i will be used to denote the arc of the boundary of this quadrilateral that extends clockwise from p_{i-1} to p_i, with p_4 standing for p_0. Such a quadrilateral is denoted by (D_1, D_2, D_3, D_4), $D(p_0, p_1, p_2, p_3)$, or just D, depending on convenience.

Similarly, a trilateral $D = (D_1, D_2, D_3)$ is defined as a topological disk with three distinct distinguished clockwise oriented points on its boundary.

[1] In standard terminology, the graph is the 1-skelaton of a triangulation of a spherical surface without boundary. However, as explained shortly, it is justifiable to call the graph itself a triangulation, because the triangulation is reproducible from it.

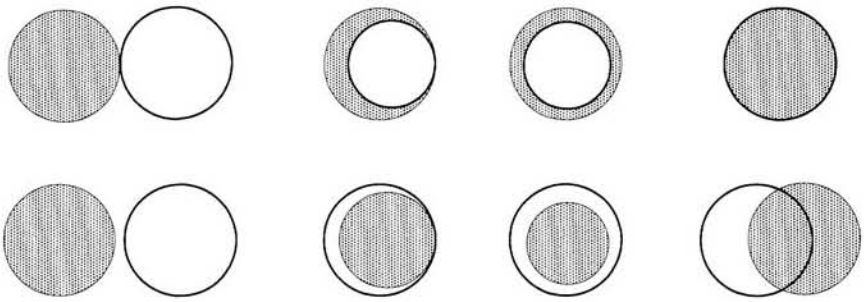


FIGURE 2.1. Topologically there are eight ways in which two circles can intersect.

Let P be a nondegenerate packing in S^2 whose nerve is a triangulation T, possibly with boundary. When we consider the sphere as oriented, the packing P naturally induces an orientation on its nerve T (that is, a 'clockwise' orientation for the triangles of T). From now on, a triangulation (with boundary) means an oriented triangulation (with boundary), and when we say that a packing has a triangulation as its nerve, it is implicit that the orientation that it induces on the nerve is compatible with the orientation of the triangulation.

The motivation for the following definitions comes from the simple but very important observation that the possible patterns of intersection of two circles are very special, topologically. See Figure 2.1.

Let D be a subset of the sphere S^2 and $p, q \in S^2$. We say that a curve γ *connects* p and q in D if the endpoints of γ are p, q, and $\text{relint}(\gamma) \subset D$. Here, and in the following, $\text{relint}(\gamma)$ means $\gamma - \{\text{its endpoints}\}$. Note that p and q do not have to be in D to be connected by a curve in D; they may be in $\overline{D} - D$.

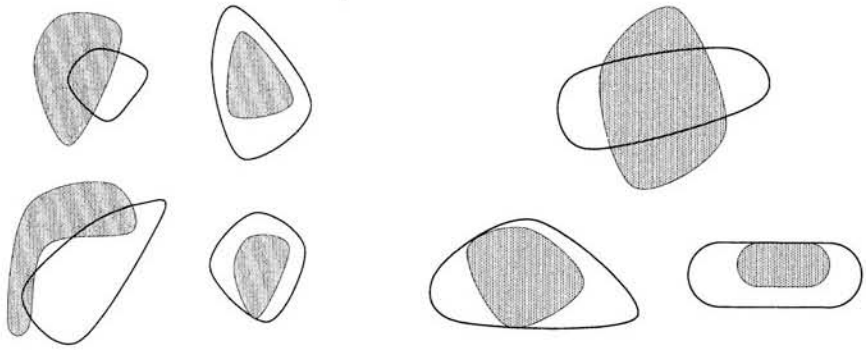
Let A, B be two topological disks in the sphere. We say that A *cuts* B if there are two points in $B - \text{interior}(A)$ that are not connected by any curve in $\text{interior}(B - A)$. A and B are *incompatible* if $A \neq B$ and A cuts B, or B cuts A. Otherwise, they are *compatible.* See Figure 2.2 on the next page.

If A, B are disklike, then they will be considered compatible if $A = B$ or A'^c and B'^c are unequal and compatible for every connected component A' of A^c and every connected component B' of B^c. In other words, either $A = B$, or whenever you adjoin to A all but one of the connected components of its complement and do the same to B, the resulting topological disks are unequal and compatible. (This definition is, in turn, compatible with the definition for the case where A, B are topological disks.)

2.1. *Examples.* (1) Any two circles are compatible.

(2) More generally, homothetic strictly convex bodies in the plane are compatible.

(3) The closures of the complements of two compatible topological disks are also compatible.



(a) Some compatible pairs. (b) Some incompatible pairs.

FIGURE 2.2

(4) Homothetic convex shapes that are not strictly convex can be incompatible.
(5) It is not hard to see that given a Riemannian metric on the sphere, any two metric balls of it are compatible.

And finally, some notation. Whenever $Q = (Q_v : v \in V)$ is a packing and $I \subset V$, the union $\bigcup_{v \in I} Q_v$ is abbreviated to Q_I.

3. The Incompatibility Theorem

This section is devoted mostly to the following theorem, which is our central tool.

3.1. **Incompatibility Theorem.** *Let $T = T(V)$ be a finite triangulation of a quadrilateral with boundary vertices a, b, c, d in clockwise order with respect to the other vertices of T (if such exist). Let $D = (D_1, D_2, D_3, D_4)$ be a quadrilateral in S^2. Suppose that $Q = (Q_v : v \in V)$ and $P = (P_v : v \in V)$ are two nondegenerate packings in D, both having nerve T. Further suppose that Q_v and P_v are disklike for $v \in V - \{a, b, c, d\}$; that $P_a \subset D_1$, $P_b \subset D_2$, $P_c \subset D_3$, $Q_d \supset D_4$; and that P_v is disjoint from Q_d for $v \in V - \{a, c, d\}$. Then there is some vertex $v \in V - \{a, b, c, d\}$ for which Q_v and P_v are incompatible.* See Figure 3.1.

3.2. *Remarks.* The quadrilateral D is not an important ingredient of the situation to which the theorem applies. It is merely a frame of reference that helps us describe the relative positions of the packings P and Q. In the situations where the theorem is applied the quadrilateral D is not mentioned.

The incompatibility theorem is the packing-theoretic analogue of the concept of the conformal modulus of a quadrilateral. Lemma 5.2 is probably the corresponding analogue for the concept of the conformal modulus of an annulus.

3.3. *Coalescing.* Before the proof of the theorem, we describe a simple procedure that will be very useful to us. Let $T = T(V)$ be some finite planar

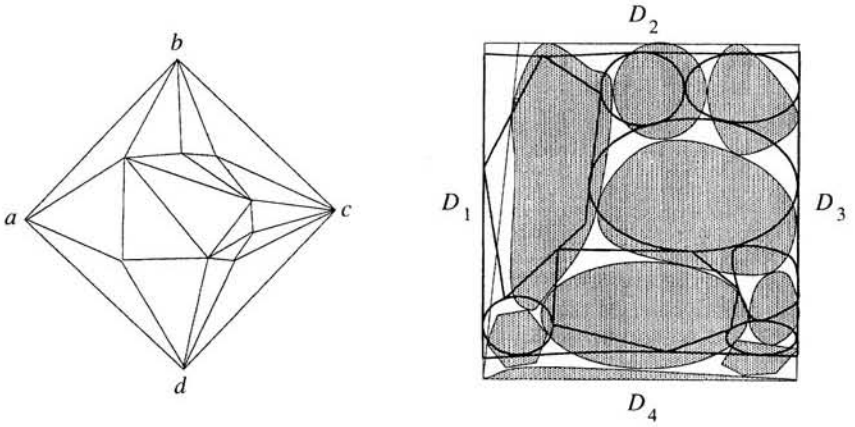


FIGURE 3.1. An incompatibility is forced. (Find it.)

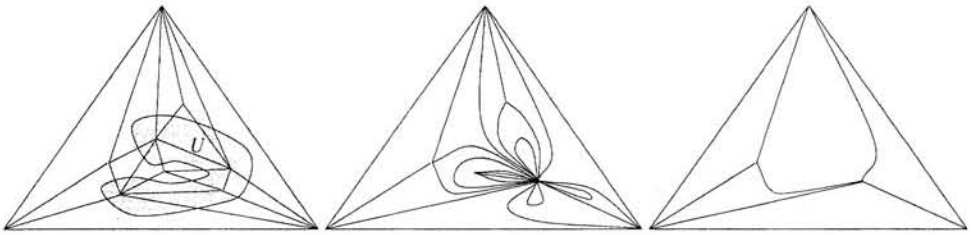


FIGURE 3.2. Coalescing a set of vertices. First some multiple edges are created, but these can be deleted, together with some of the vertices, to obtain a triangulation.

triangulation, and let $Q = (Q_v : v \in V)$ be some packing of disklike sets with nerve T. Suppose that U is some nonempty connected subset of vertices, and suppose that z is some vertex that is not in U. We describe what we mean by *coalescing* Q_U *while keeping* z. The idea is that we want to consider the packing $\hat{Q}$ that is the same as Q except for the fact that all the sets Q_v, $v \in U$ are united into one, Q_U. $\hat{Q}$ might not be a good packing, however, because its nerve might not be a planar triangulation, and the set Q_U might not be disklike.

First, we look at the combinatorial picture. Coalescing the sets Q_v, $v \in U$ is analogous to coalescing the vertices U into one. Consider a realization of T in the plane. Let $\hat{T}$ be a picture obtained from T by collapsing the vertices in U. See Figure 3.2. Generally, in $\hat{T}$ there are some loops and multiple edges, but every 2-cell determined by $\hat{T}$ has at most three edges of $\hat{T}$ on its boundary. We want to have a triangulation without any loops or multiple edges. Therefore we must remove the excess edges from $\hat{T}$, and do this without having 2-cells with too many edges on their boundary.

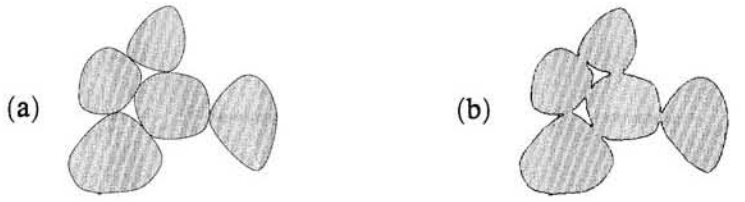


FIGURE 3.3. Initially, after coalescing, one may obtain a nondisk-like set, but this is easily remedied.

Consider a loop l in $\hat{T}$; that is, an edge having the same vertex at both its endpoints. If we just delete l from $\hat{T}$, then we might have some resulting 2-cell, which is a quadrilateral, and we do not want that. The loop l determines two regions in $\hat{T} - l$. If we delete l together with all the edges and vertices of $\hat{T}$ that are in one of these regions, then in the resulting picture all the 2-cells are still 1-, 2-, or 3-gons. Because we want to keep z, we delete with l all the stuff that is in a region that does not contain z.

Similarly, we deal with multiple edges. If two edges of $\hat{T}$ have the same two vertices as endpoints, then we delete one of them, together with all the stuff in one of the two regions determined by these two edges that does not contain z.

It is straightforward to verify that this procedure is consistent, and that one obtains a triangulation T'. (This is so provided that z neighbors with some vertex that is not in U. If all the neighbors of z are in U, then T' consists of two vertices and an edge between them.) In T', z and the vertices neighboring with z are still present. Let V' be the set of vertices of T', and assume that the vertex corresponding to the coalesced set U is u.

Let $Q' = (Q'_v : v \in V')$ be the packing defined by $Q'_v = Q_v$, $v \in V' - \{u\}$, $Q'_u = Q_U$. Then the packing Q' has the nerve T', and the only remaining problem is that the set Q'_u might not be disklike. See Figure 3.3(a). We then modify Q'_u slightly to make it disklike, as in Figure 3.3(b). Note that this can be done with an arbitrarily minute modification of Q'_u, while keeping Q' as a packing with the same combinatorics.

The packing obtained in this manner is called *the packing obtained from* Q *by coalescing* U *while keeping* z. This procedure is done similarly for triangulations with boundary, and for infinite triangulations, provided the boundary of U is finite.

Proof of Incompatibility Theorem. First we make an easy reduction to the case where all the sets Q_v, P_v, $v \in V - \{a, b, c, d\}$ are topological disks. If Q_v $(v \neq a, b, c, d)$, say, is not a topological disk, then we can adjoin the connected components of its complement to Q_v, except for the one that intersects the other sets Q_w, $w \neq v$. Because Q_v is disklike, the resulting set Q'_v would be a topological disk. We apply this procedure to all the Q_v, P_v, $v \in V - \{a, b, c, d\}$ that are not topological disks, and obtain two packings Q', P'

that satisfy the hypotheses of the theorem. If Q'_w and P'_w $(w \neq a, b, c, d)$ are incompatible, then surely, the same holds for Q_w and P_w. We may, and do, therefore assume that the sets Q_v, P_v, $v \in V - \{a, b, c, d\}$ are topological disks.

The proof proceeds by induction on the number of vertices in T. The base of the induction is the case in which there are no vertices in T other than a, b, c, d. Because D_1 and D_3 are disjoint, and $P_a \subset D_1$, $P_c \subset D_3$, there cannot be any edge between a and c in T. If there is an edge between b and d, then Q_d must intersect $Q_b \subset D_2$. This would force $Q_d \cap P_b \neq \varnothing$, because Q_d is connected and intersects D_2 and D_4, and P_b is connected and intersects D_1 and D_3. By our assumptions, however, $Q_d \cap P_b = \varnothing$, and therefore there cannot be an edge between b and d. As a result, since T is a triangulation, there must be some vertices other than a, b, c, d. This takes care of the base of the induction. (As a warm-up, the reader may want to examine the case where T has 5 vertices.)

Set $J = V - \{a, c, d\}$. Let $a = v_0, v_1, \ldots, v_n, v_{n+1} = c$ be the neighbors of d in T in clockwise order. Note that $b \notin \{v_0, v_1, \ldots, v_{n+1}\}$. If Q_{v_i} and P_{v_i} are incompatible for some $i = 1, 2, \ldots, n$, then we are done. So assume that Q_{v_i} and P_{v_i} are compatible for $i = 1, 2, \ldots, n$.

Our method of proof is to find a nonempty set of vertices $H \subset \{v_1, v_2, \ldots, v_n\}$ so that Q_H is disjoint from P_{J-H}. After this is done, we examine the packings obtained by coalescing Q_H into Q_d and P_H into P_d. Then the induction hypothesis establishes the theorem.

The proof is divided into three sections.

a. Constructing H. Let E be the connected component of $D - (P_J \cup Q_d)$ whose boundary intersects Q_d. See Figure 3.4 on page 136. P_J is disjoint from Q_d, and is connected. Therefore E is well defined. We think of E as a quadrilateral whose edges are, in clockwise order, $E_a = \partial E \cap D_1$, $E_P = \partial E \cap P_J$, $E_c = \partial E \cap D_3$, $E_Q = \partial E \cap Q_d$. Note that there is no loss of generality in assuming that E_Q is a simple curve, because we can easily modify Q_d slightly so that it has a nice boundary without disrupting the hypotheses of the theorem. Thus we assume that this is the case. We think of E_Q and E_P as being oriented from D_1 to D_3, with E_Q having E on its left, and E_P having E on its right.

With the intention of introducing some more notations, we walk along E_P. As we walk on E_P, we visit, in consecutive order, the sets $P_{v_1}, P_{v_2}, \ldots, P_{v_n}$. This gives a sequence of arcs $\gamma_1, \gamma_2, \ldots, \gamma_n$. In other words, E_P consists of these arcs in this order, and each arc γ_i lies on the boundary of P_{v_i}. Let p_i be the terminal point of the arc γ_i, $i = 1, 2, \ldots, n$ (this is also the initial point of γ_{i+1} for $i < n$), and let p_0 be the initial point of γ_1.

For $i = 1, 2, \ldots, n$, v_i neighbors with d. Pick some arbitrary point q_i in the intersection $Q_{v_i} \cap Q_d$. It is important to note that the points $q_1, q_2, \ldots, q_n$ appear in that order on E_Q. (E_Q is oriented from D_1 to D_3.) Again, let i be some index in the range $i = 1, 2, \ldots, n$. We say that v_i is an *invader*

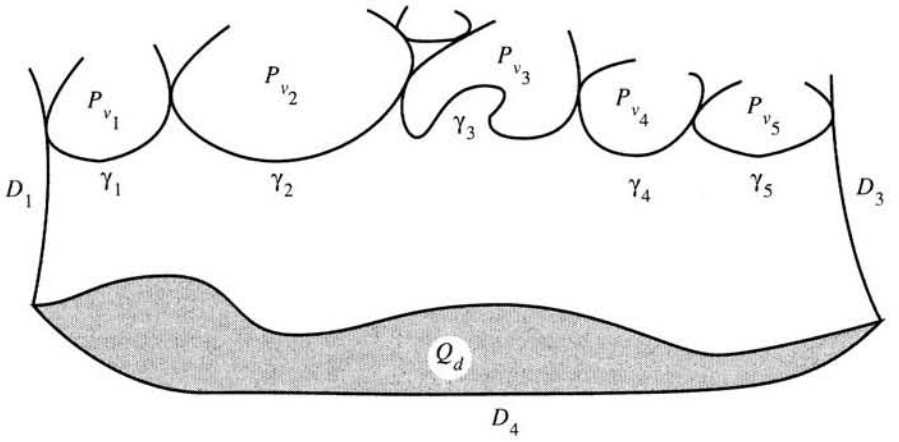


FIGURE 3.4. The quadrilateral E and its accessories.

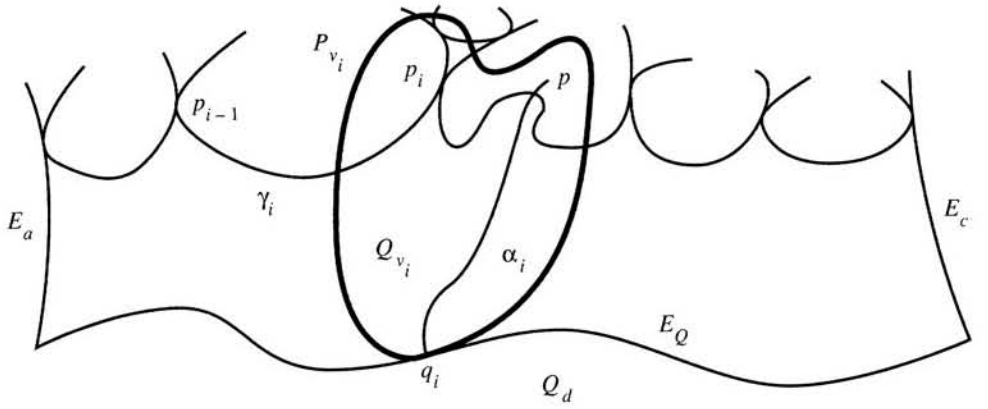


FIGURE 3.5. v_i is an invader.

if Q_{v_i} intersects $P_{J-\{v_i\}} - \{p_{i-1}, p_i\}$. See Figure 3.5. Consider some invader v_i and some point $p \in Q_{v_i} \cap (P_{J-\{v_i\}} - \{p_{i-1}, p_i\})$. As q_i and p are in $Q_{v_i} - \text{interior}(P_{v_i})$, there is some simple curve, say β_i, connecting them in $\text{interior}(Q_{v_i} - P_{v_i})$, because Q_{v_i} and P_{v_i} are compatible. We orient β_i from q_i to p. Let r be the first point of β_i that is in P_J (β_i has some point of P_J, namely p), and let α_i be the part of β_i extending from q_i to r. α_i is a cross-cut of the quadrilateral E. It has one endpoint, q_i, in E_Q and one endpoint, r, in E_P.

Note that r, the endpoint of α_i that is on E_P, is not in γ_i. If $r \neq p$, then this is clear, because the curve β_i avoids P_{v_i}, except perhaps at its endpoints. If $r = p$, then we know that $r \neq p_{i-1}, p_i$, and since $p \in P_{J-\{v_i\}}$, $p \notin \gamma_i$.

For every invader v_i there are two possibilities. Either α_i separates γ_i from E_a in E, or it separates γ_i from E_c in E. (See Figure 3.5.) In the first case we call i a *left invader*, and in the second case we call i a *right invader*. Note that

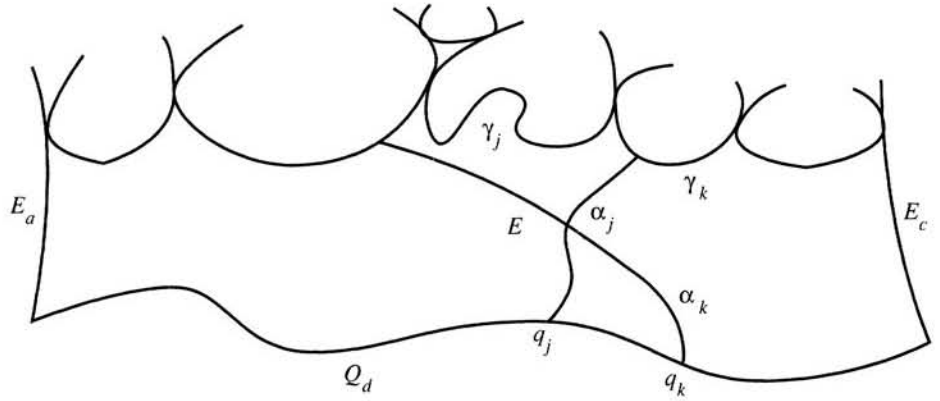


FIGURE 3.6. α_j and α_k must cross, but this is impossible.

if i is a left invader, then α_i separates in E every one of the points q_h, $h < i$ from every one of the arcs γ_m, $m \geqslant i$. Similarly, if i is a right invader, then α_i separates in E every one of the points q_h, $h > i$ from every one of the arcs γ_m, $m \leqslant i$.

For notational convenience, let 0 be considered a right invader, and let $n+1$ be considered a left invader. Since 0 is a right invader, $n+1$ is left invader, and every invader is either right or left, there will be some indices $0 \leqslant j < k \leqslant n+1$ so that j is a right invader, k is a left invader, and there is no invader between j and k. (To find such j, k, start with $j = 0$, $k = n+1$. If there is no invader between j and k, stop. Otherwise pick some invader i in the range $j < i < k$. If i is a right invader, set $j := i$. Otherwise, set $k := i$. Continue in this manner until a situation is reached in which there is no invader i, $j < i < k$.) Suppose that $j < k$ are such indices, and let $H = \{v_{j+1}, v_{j+2}, \ldots, v_{k-1}\}$.

b. Showing that $H \neq \varnothing$, and that Q_H is disjoint from P_{J-H}. To prove $H \neq \varnothing$, it is sufficient to demonstrate that $j + 1 \neq k$. Wishing to arrive at a contradiction, assume that $j + 1 = k$. Since $n > 0$, either $j > 0$ or $k < n + 1$ (or both). Because these cases are symmetric, we assume that $k < n + 1$. As k is a left invader, it is obvious that $k > 1$ and therefore, also $j = k - 1 > 0$. Again, because k is a left invader, α_k separates q_j from γ_k in E. Similarly, α_j separates q_k from γ_j in E. Since γ_j comes before γ_k on E_P and q_j comes before q_k on E_Q, this implies that the curves α_j and α_k must cross. (See Figure 3.6.) This is impossible, because $v_k \neq v_j$, and $\text{relint}(\alpha_j) \subset \text{interior}(Q_{v_j})$, $\text{relint}(\alpha_k) \subset \text{interior}(Q_{v_k})$. This contradiction shows that $k > j + 1$, thus establishing $H \neq \varnothing$.

Now we will see that $Q_H \cap P_{J-H} = \varnothing$; i.e., that $Q_{v_h} \cap P_{J-H} = \varnothing$ for $j < h < k$. Let h be an index in the range $j < h < k$. If $Q_{v_h} \cap P_{J-\{v_h\}} = \varnothing$, then obviously $Q_{v_h} \cap P_{J-H} = \varnothing$. So assume otherwise, and let p be a point in $Q_{v_h} \cap P_{J-\{v_h\}}$. We need to show that $p \notin P_{J-H}$. Because v_h is not an invader,

we know that p must be p_h or p_{h-1}. Since these cases are treated similarly, we only consider the case $p = p_h$.

The point p_h is in the intersection $P_{v_h} \cap P_{v_{h+1}}$. Because the intersection of any three sets in the packing P is empty, no other set P_u $(u \neq v_h, v_{h+1})$ contains p_h. Thus we only need to worry about $P_{v_{h+1}}$. If $h+1 < k$, then $v_{h+1} \in H$, and if $h+1 = k = n+1$, then $v_{h+1} = c \notin J$. Therefore we may, and do consider only the case $h+1 = k < n+1$. Now, $q_h, P_h \in Q_{v_h} - \text{interior}(P_{v_h})$, and Q_{v_h} and P_{v_h} are compatible. Therefore, there is some curve, say α, that connects q_h and p_h in $\text{interior}(Q_{v_h} - P_{v_h})$.

The curve α cannot cross α_k, because $\text{interior}(Q_{v_h})$ and $\text{interior}(Q_{v_k})$ are disjoint. Similarly, $\text{relint}(\alpha)$ is disjoint from E_Q. Also, $\text{relint}(\alpha)$ is disjoint from E_P, because it certainly is disjoint from $\gamma_h \subset P_{v_h}$, and v_h is not an invader. Thus $\text{relint}(\alpha) \subset E$. But the two endpoints of α, q_h, and $p_h = p_{k-1}$ are separated in E by α_k, because v_k is a left invader. This means that α must cross α_k; a contradiction. This contradiction establishes $Q_{v_h} \cap P_{J-H} = \varnothing$, and therefore $Q_H \cap P_{J-H} = \varnothing$.

c. Applying the inductive hypothesis. Let Q' and P' be the packings obtained from Q and P, respectively, by coalescing $H \cup \{d\}$ while keeping b. Let $T' = T'(V')$ be the triangulation that is the common nerve of these two packings, and let d' be the vertex of T' corresponding to the coalesced set $H \cup \{d\}$. Because of part b, we have $Q'_{d'} \cap P'_{V'-\{a,c,d'\}} = \varnothing$, provided that the touch-up modification done in the coalescing procedure is not too big. The inductive hypothesis now applies to the packings Q' and P', and establishes the theorem. □

4. Rigidity of infinite circle packings that almost fill the sphere

In this section we prove Theorem 1.1.

4.1. **Definitions.** Let $P = (P_v : v \in V)$ be a finite or infinite packing in S^2. An *interstice* of P is a connected component of the complement of $P_V = \bigcup_{v \in V} P_v$ whose boundary is formed by finitely many of the sets P_v. The carrier of P, $\text{carrier}(P)$, is the union of all the interstices and all the sets P_v. A connected component of $S^2 - \text{carrier}(P)$ is called a *singularity* of P. A singularity is *parabolic* if it consists of a single point. And $\text{singular}(P)$ denotes the union of the singularities of P: $\text{singular}(P) = S^2 - \text{carrier}(P)$.

Proof of Rigidity Theorem 1.1. Let $P = (P_v : v \in V)$ be another circle packing whose nerve is T, and let $[a, b, c]$ be some triangle in T. By making an appropriate initial normalization, we assume that $P_v = Q_v$ for $v = a, b, c$. (The Möbius transformation that takes the three intersection points of the circles P_a, P_b, P_c to the corresponding intersection points of Q_a, Q_b, Q_c also takes P_a, P_b, P_c to Q_a, Q_b, Q_c, respectively.

If $Q_v = P_v$ for each $v \in V$, then we are done. With the intention of reaching a contradiction, we assume that this is not the case. Starting from $[a, b, c]$, we walk on triangles of T by moving from a triangle to an adjacent triangle. We can do this, and reach some first triangle where the three circles corresponding to the vertices are not the same in both packings. So, without loss of generality, we assume that $Q_d \neq P_d$, where d is the vertex other than b that makes a triangle with a and c. We further assume that the point ∞ is contained in the interstice corresponding to the triangle $[a, b, c]$. Thus our packings lie in the plane.

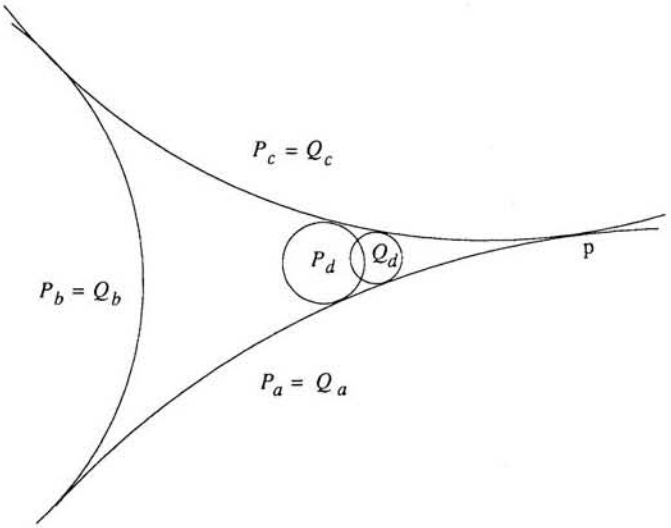
There are two possibilities. Either Q_d is smaller than P_d, and the situation is as in Figure 4.1(a) on the next page, or the other way around. Both cases are treated similarly, and we assume that the situation is as in Figure 4.1(a). Let p be the point of intersection of P_a and P_c, which is also the point of intersection of Q_a and Q_c. Pick some number $\beta > 1$ with $\beta - 1$ small, and expand the packing Q by a homothety with center p and expanding ratio β. Continue to denote the resulting packing by Q. The modified picture is given in Figure 4.1(b).

The boundary of $P_V = \bigcup_{v \in V} P_v$ is also the boundary of $\bigcup_{v \in V} \text{interior}(P_v)$. Therefore it is nowhere dense, and thus is of Baire category I. Because singular(Q) is countable, the set of translations S for which singular$(S(Q))$ intersects ∂P_V is also of Baire category I. By Baire's theorem, this shows that there is an arbitrarily small translation S for which singular$(S(Q))$ is disjoint from ∂P_V. Apply such a translation to Q, and make it small enough so that qualitatively, Figure 4.1(b) is still correct, except in a small neighborhood of p. We continue to denote the resulting packing by Q.

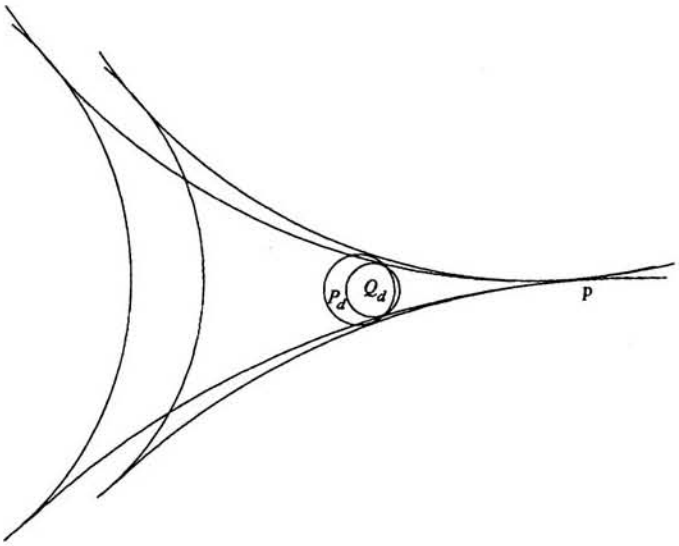
The complement of $Q_a \cup Q_b \cup Q_c \cup Q_d$ consists of a quadrilateral that contains the other sets of the packing Q, and two trilaterals (one of the trilaterals contains infinity). Let D_Q be the (closure of the) quadrilateral region, and let Q'_a, Q'_b, Q'_c, Q'_d be the arcs of D_Q that are on Q_a, Q_b, Q_c, Q_d, respectively. We have $D_Q = (Q'_a, Q'_b, Q'_c, Q'_d)$. Let the quadrilateral $D_P = (P'_a, P'_b, P'_c, P'_d)$ in $S^2 - (P_b \cup P_b \cup P_c \cup P_d)$ be defined in the same manner. These two quadrilaterals are indicated in Figure 4.2. Define new packings $Q' = (Q'_v : v \in V)$ and $P' = (P'_v : v \in V)$, by $Q'_v = Q_v$, $P'_v = P_v$ for $v \neq a, b, c, d$, and P'_v, Q'_v as defined above for $v = a, b, c, d$.

Our two quadrilaterals D_Q, D_P, and the two packings Q', P' are in the right relative position to apply the Incompatibility Theorem. (See Remark 3.2.) If we could apply the theorem to them, then we could conclude that there is some $v \in V - \{a, b, c, d\}$ so that Q'_v and P'_v are not compatible. This is impossible, because these sets are circles, and this contradiction would finish the proof. The problem is that these packings are not finite. We overcome this problem by using the fact that singular(Q') is disjoint from $\partial P'_V$. We use this fact to cook up finite packings from Q' and P'.

Let $\mathfrak{H}$ be the collection of all interstices of the packing P', and let $\mathfrak{L}$ be the collection of all the singularities of the packing P'. The collection



(a) Case Q_d is smaller than P_d.



(b) After expanding Q.

FIGURE 4.1

$\{H,\ \text{interior}(P_v'),\ \text{interior}(L) : H \in \mathfrak{H},\ v \in V,\ L \in \mathfrak{L}\}$ is an open cover for the compact set $\text{singular}(Q')$, because $\text{singular}(Q') \cap \partial P_V' = \varnothing$. Let $\{H,\ \text{interior}(P_v'),\ \text{interior}(L) : H \in \mathfrak{H}',\ v \in V',\ L \in \mathfrak{L}'\}$ be a finite sub-

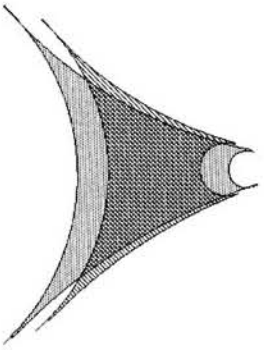


FIGURE 4.2. The quadrilaterals D_Q and D_P.

cover; that is, $\mathfrak{H}' \subset \mathfrak{H}$, $V' \subset V$, $\mathfrak{L}' \subset \mathfrak{L}$ are finite subsets with

$$G \stackrel{\text{def}}{=} \left(\bigcup_{v \in V'} \text{interior}(P'_v) \right) \cup \left(\bigcup_{H \in \mathfrak{H}'} H \right) \cup \left(\bigcup_{L \in \mathfrak{L}'} \text{interior}(L) \right),$$

containing singular(Q'). Let V_1 be the set of vertices $v \in V$ so that Q'_v is not contained in G. V_1 is a finite set, because for any infinite sequence of distinct circles in the packing Q', the radii of the circles must tend to zero, and any accumulation point for the sequence is necessarily in singular(Q'), which is a compact subset of the open set G.

Let V_2 be some finite connected set of vertices that contains $V_1 \cup \{a, b, c, d\}$, and let $\hat{P}$ and $\hat{Q}$ be the packings obtained from P' and Q', respectively, by coalescing each connected component of $V - V_2$ while keeping a. (Recall the technique of coalescing from 3.3.) Because $\hat{P}$ and $\hat{Q}$ are finite, we can now apply the incompatibility theorem. We conclude from it that there are sets $\hat{P}_w$ and $\hat{Q}_w$, both corresponding to the same vertex $w \neq a, b, c, d$ that are not compatible. This vertex w cannot be in V_2, because the sets of $\hat{P}$ and $\hat{Q}$ that correspond to vertices in V_2 are circles, and circles are always compatible.

Let W be the connected component of $V - V_2$ that coalesced to form w. The union $Q'_W = \bigcup_{v \in W} Q'_v$ is a connected set whose closure is contained in G. The union defining G is a finite union of disjoint open sets. Because Q'_W is connected, $\overline{Q'_W}$ is contained in one set of this union, say $\overline{Q'_W} \subset \text{interior}(F)$, with $F \in \mathfrak{H}' \cup \mathfrak{L}'$ or $F = P'_u$ for some $u \in V'$. If F is not disjoint from P_W, then F is contained in P_W. Thus $\overline{Q'_W}$ is either contained in the interior of P_W, or is disjoint from P_W. $\hat{P}_w$ and $\hat{Q}_w$ are compatible, therefore, provided that the modifications done during the coalescing procedure are small enough. This contradicts our previous conclusion, and thereby completes the proof of the theorem. □

It is possible to obtain a generalization of this theorem to packings by convex sets. This is given at the end of the next section.

5. Rigidity of packings that almost fill a disk

5.1. **Rigidity Theorem, hyperbolic case.** *Let $T = T(V)$ be an infinite, planar triangulation. Let $Q = (Q_v : v \in V)$ be a circle packing in the open unit disk U with nerve T, and suppose that $U - \text{carrier}(Q)$ is at most countable. Let P be another circle packing in U with nerve T. Further assume that the singularity of P that corresponds to the singularity $S^2 - U$ of Q is also equal to $S^2 - U$. Then Q and P are Möbius equivalent.*

We need the following lemma, which is topological in nature.

5.2. **Lemma.** *Let $T = T(V)$ be a finite triangulation of the sphere, and let $a \leftrightarrow b$ be an edge in T. Suppose that $P = (P_v : v \in V)$ and $Q = (Q_v : v \in V)$ are two nondegenerate packings of topological disks on S^2 with nerve T. Further suppose that $P_a \subset \text{interior}(Q_a)$; $Q_b \subset \text{interior}(P_b)$; and $Q_z \subset \text{interior}(P_z)$ [or $P_z \subset \text{interior}(Q_z)$] for some $z \neq a, b$ in V. Then Q_v and P_v are incompatible for some vertex $v \in V - \{a, b, z\}$.*

Proof. The idea is to take a double cover, and obtain a situation to which the Incompatibility Theorem can be applied. The statement of the lemma is symmetric in Q and P, and thus we consider only the case $Q_z \subset \text{interior}(P_z)$.

Pick some points $s_a \in \text{interior}(P_a)$ and $s_z \in \text{interior}(Q_z)$. Consider a double cover M of S^2, in the topological sense, branched over s_a and s_z. The Packings P, Q lift to packings $P' = (P'_v : v \in V')$, $Q' = (Q'_v : v \in V')$ in M, where to each vertex $v \in V - \{a, z\}$ correspond two vertices in V', say v_1, v_2, and to each of a and z corresponds one vertex in V', say a' and z', respectively. Figure 5.1(a) illustrates the sets corresponding to the vertices a and b downstairs, and Figure 5.1(b) illustrates those corresponding to a', b_1, b_2 in the double cover M.

For convenience, we assume that $P_a \cap P_b$ contains precisely two distinct points, say p, p', and that similarly $Q_a \cap Q_b = \{q, q'\}$. There clearly is no loss of generality in this assumption, because we can make slight modifications to P_a near P_b and to Q_a near Q_b. Let $p_1, p_2, p'_1, p'_2, q_1, q_2, q'_1, q'_2$ be the corresponding points in the double cover M. Let D_Q be the closure of the connected component of $M - (Q'_{b_1} \cup Q'_{a'} \cup Q'_{b_2})$ that contains the other sets of Q', and let D_P be the closure of the connected component of $M - (P'_{b_1} \cup P'_{a'} \cup P_{b'_2})$ that contains the other sets of P'. We view D_Q and D_P as quadrilaterals with vertices q_1, q_2, q'_1, q'_2 and p_1, p_2, p'_1, p'_2, respectively. Let P''_{a_1} and P''_{a_2} be the two edges of the quadrilateral D_P that are on the boundary of $P'_{a'}$. Let P''_{b_1} and P''_{b_2} be the two edges of D_P that are on P'_{b_1} and P'_{b_2}, respectively. Similarly define the edges of D_Q, $Q''_{a_1}, Q''_{a_2}, Q''_{b_1}, Q''_{b_2}$. See Figure 5.2.

We now define modified packings $P'' = (P''_v : v \in V'')$, $Q'' = (Q''_v : v \in V'')$. Set $V'' = (V' \cup \{a_1, a_2\}) - \{a'\}$. P''_v and Q''_v have been defined above for $v = a_1, a_2, b_1, b_2$. Set $P''_v = P'_v$ and $Q''_v = Q'_v$ for other $v \in V''$. The

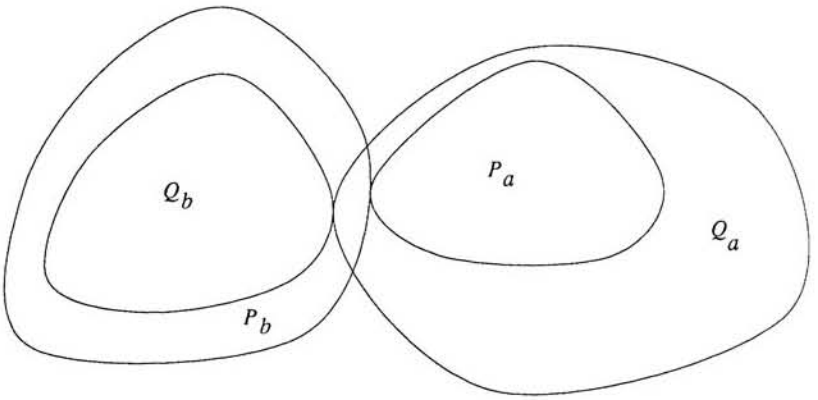


FIGURE 5.1(a). The sets corresponding to a and b.

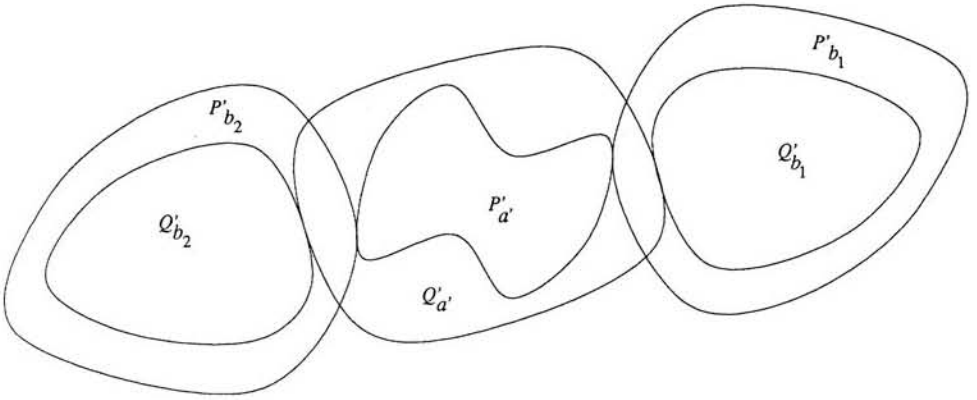


FIGURE 5.1(b). Sets corresponding to a', b_1, b_2 in M.

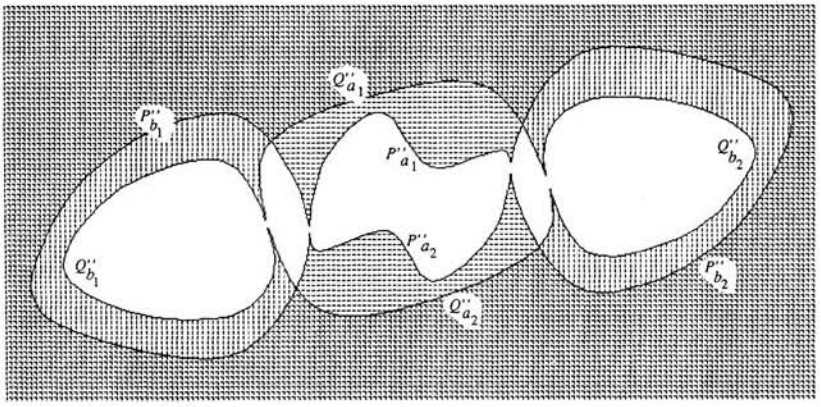
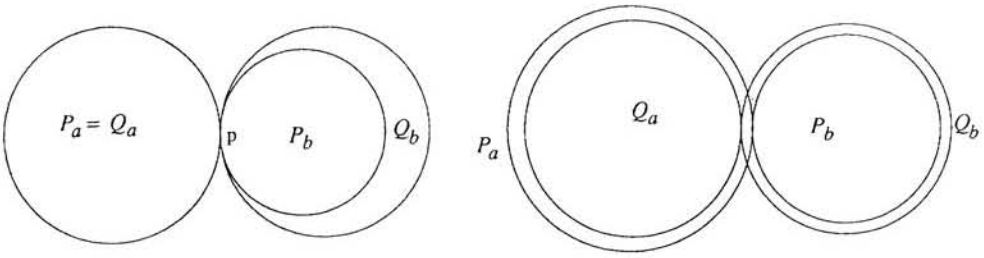


FIGURE 5.2. The quadrilaterals D_P and D_Q.



(a) Case $P_b \subsetneqq Q_b$. (b) After expanding and translating P.

FIGURE 5.3

packings Q'', P'' have as nerve a common triangulation T'' of a quadrilateral with boundary a_1, a_2, b_1, b_2 (provided the labeling in the branching $v \to v_1, v_2$ is done consistently for both packings), which can roughly be described as a double cover of T branched over the vertex z and the edge $a \leftrightarrow b$. It is clear that the incompatibility theorem applies to these packings and gives a vertex $w \in V'' - \{a_1, a_2, b_1, b_2\}$ for which Q''_w and P''_w are incompatible. It is impossible that $w = z'$, because $Q''_{z'} \subset \text{interior}(P''_{z'})$. Thus $w = v_j$, say, with $v \in V - \{a, b, z\}$, and $j = 1$ or $j = 2$. But this implies that Q_v and P_v are incompatible, as needed. □

Proof of 5.1. After a few initial normalizations, we reach a situation where the above lemma can be applied.

Let Q_s be the singularity $S^2 - U$ of the packing Q, and let P_s be the corresponding singularity of P. (We assume that the symbol s is not a vertex of T.)

Pick some edge $a \leftrightarrow b$ in T, and let p be the point of intersection of Q_a and Q_b. After renormalizing by a Möbius transformation taking U onto U (i.e., a hyperbolic isometry), we may, and do, also assume that $P_a \cap P_b = \{p\}$, and that the outward unit normal of Q_a at p is the same as that of P_a.

There is some (unique) $\alpha > 0$ so that expanding P_a by a homothety with center p and expansion α takes P_a to Q_a. Apply this expansion to the packing P, and continue to denote the resulting packing by P. Now P_s is either contained in Q_s, or contains Q_s, depending on whether $\alpha \geqslant 1$ or $\alpha \leqslant 1$, respectively.

Case 1. $P_b \subsetneqq Q_b$. See Figure 5.3(a). Expand the packing P by a homothety with center p. Make the expansion ratio $\beta > 1$, but small enough so that afterwards $P_a \supsetneqq Q_a$, $P_b \subsetneqq Q_b$, and $P_s \subset \text{interior}(Q_s)$ or $Q_s \subset \text{interior}(P_s)$. ($P_s \subset \text{interior}(Q_s)$ would be the case, if $\alpha \geqslant 1$, and $Q_s \subset \text{interior}(P_s)$, otherwise.) Now translate the packing P slightly in the direction of the outward normal to P_a and Q_a at p. After the translation, we have $Q_a \subset \text{interior}(P_a)$, $P_b \subset \text{interior}(Q_b)$, and $P_s \subset \text{interior}(Q_s)$ or $Q_s \subset \text{interior}(P_s)$. See Figure 5.3(b). Finally, make a small perturbing translation that does not destroy the

above relations, but makes all the singularities of Q, except possibly Q_s, be disjoint from the boundary of $P_V = \bigcup_{v \in V} P_v$. By Baire's category theorem, this can be done, as in the proof of 1.1.

We assume that $P_s \subset \text{interior}(Q_s)$. The other possibility is dealt with similarly. Let Z be a set of vertices whose boundary is finite, so that all the sets P_v, $v \in Z$ are contained in $\text{interior}(Q_s)$, and P_{V-Z} is bounded away from P_s. (To see that one can find such a set of vertices, let V_n be the set of vertices whose distance from a is at most n. Let C_n be the simple closed path in the boundary of V_n so that P_{C_n} separates $P_{V_n - C_n}$ from P_s. As $n \to \infty$, $P_{C_n} \to \partial P_s$. Thus, for sufficiently large n, one can take Z to be the connected component of $V - C_n$ that does not contain a.)

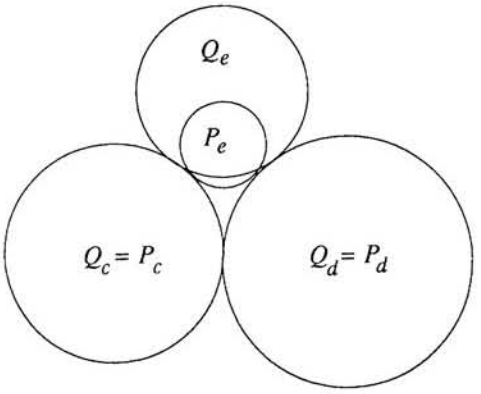
Let P' and Q' be the packings obtained by coalescing Z to some vertex, say z, while keeping a. We make further coalescings, as in the proof of 1.1, but keep the vertices a, b, z. Then to each set in these packings we adjoin the connected components of its complement which do not intersect Q_a (to make them into topological disks). Eventually, we obtain finite packings P'', Q'' with a common triangulation as nerve; corresponding sets being compatible; and $P''_z \subset \text{interior}(Q''_z)$, $Q''_a \subset \text{interior}(P''_a)$, and $P''_b \subset \text{interior}(Q''_b)$. This contradicts the lemma, and thus Case 1 is ruled out.

Case 2. $Q_b \subsetneqq P_b$. We make the same argument in this case, only now take $0 < \beta < 1$ and interchange the roles of a and b.

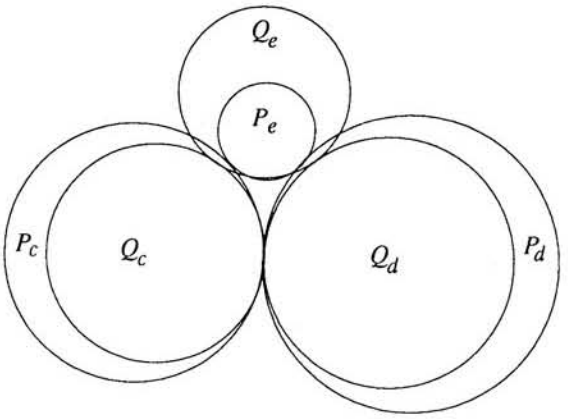
Case 3. $Q_b \not\subset P_b$ and $P_b \not\subset Q_b$. This case is impossible, because the boundaries of P_b and Q_b both pass through p and have the same unit outward normal there.

Case 4. $Q_b = P_b$. If $Q_v = P_v$ for all $v \in V$, then we are done, so assume otherwise. We also have $Q_a = P_a$, and therefore there is some triangle $[c, d, e]$ in T with $Q_c = P_c$, $Q_d = P_d$, and $Q_e \neq P_e$. We then have either Q_e bigger than P_e and $Q_e \cup Q_c \cup Q_d$ separates P_e from every P_v, $v \neq c, d, e$, as in Figure 5.4(a) (see p. 146), or the other way around. Both situations are dealt with similarly, so assume that Q_e is bigger than P_e. Let p' be the point of intersection of P_c and P_d (also $Q_c \cap Q_d = \{p'\}$). Expand the packing P by a homothety with center p' and expansion ratio $\beta > 1$ with $\beta - 1$ small. Then we have the sets Q_v, P_v, $v = c, d, e$ as in Figure 5.4(b), and $P_s \subset \text{interior}(Q_s)$, if $\alpha \geqslant 1$, or $Q_s \subset \text{interior}(P_s)$, otherwise. Now, if necessary, modify Q_e so that its interior would contain P_e, as in Figure 5.4(c), while keeping Q as a packing with the same combinatorics. (That is, Q_e must still touch the sets it used to touch. It may perhaps no longer be a circle, though.) A very slight translation would now give $Q_d \subset \text{interior}(P_d)$, and a contradiction can be reached as in case 1, with d replacing a and e replacing b. □

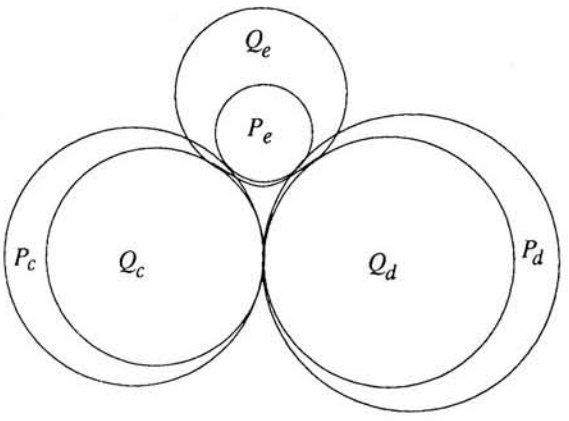
In the proof above, except for the first normalization, all the transformations applied were translations and positive homotheties. Going through the proof, one sees that (except for this initial normalization), we did not use the fact that



(a) Case 4, with Q_e bigger than P_e.

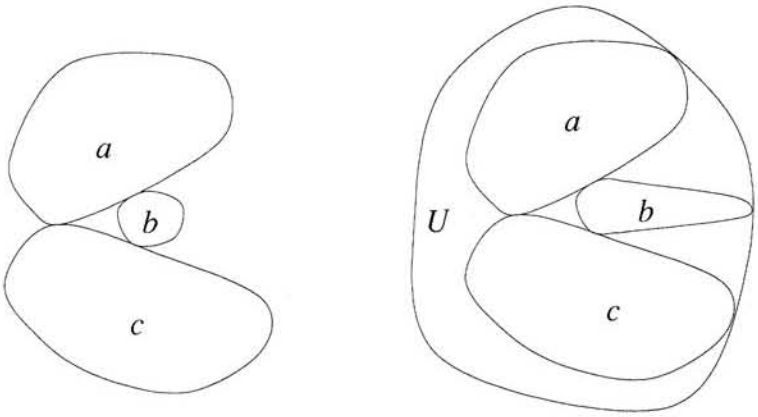


(b) After expanding P.



(c) After modifying Q_e.

FIGURE 5.4



(a) Cannot find a circumscribing convex U.

(b) After a harmless modification, such a U exists.

FIGURE 5.5

we are working with circles and, in fact, strict convexity and smoothness are enough. Thus we have the following generalization of 5.1.

5.3. **Theorem.** *Let $T = T(V)$ be an infinite, planar triangulation. Let $Q = (Q_v : v \in V)$ be a packing with nerve T of smooth strictly convex sets in an open convex set U. Suppose that $U - \text{carrier}(Q)$ is at most countable. Then Q is rigid up to three degrees of freedom, in the following sense. Let $a \leftrightarrow b$ be some edge in T, and let P be another packing in U with nerve T. Further assume that P_v is positively homothetic to Q_v for $v \in V$; the singularity of P that corresponds to the singularity $S^2 - U$ of Q is also equal to $S^2 - U$; $P_a \cap P_b = Q_a \cap Q_b$; and the outward unit normal of P_a at this intersection point is the same as that of Q_a. Then $P = Q$.*

The only need for smoothness is to insure that the packings be nondegenerate. One could dispense with the smoothness hypothesis, if one restricts the word 'packing' to mean 'nondegenerate packing'.

We also prove the following theorem, which can be seen as a generalization of 1.1.

5.4. **Theorem.** *Let $T = T(V)$ be a (possibly infinite) planar triangulation. Let $Q = (Q_v : v \in V)$ be a packing with nerve T of smooth strictly convex bodies in the plane. Suppose that ∞ is contained in one of the interstices of Q, say in the interstice corresponding to the triangle $[a, b, c]$ of T. Further suppose that* singular(Q) *is at most countable. Then Q is rigid, in the following sense. Let P be another packing in the plane with nerve T; P_v positively homothetic to Q_v for each v; and $P_v = Q_v$ for $v = a, b, c$. Then $Q = P$.*

Proof. First assume that there is some smooth convex body U that contains $Q_a \cup Q_b \cup Q_c$, and each of Q_v, $v = a, b, c$, intersects the boundary of U. Let T' be the triangulation obtained from T by adding one additional vertex, say s,

in the triangle $[a, b, c]$, adjoining edges from s to a, b, c, and splitting the triangle $[a, b, c]$ into three smaller triangles: $[a, b, s], [b, c, s], [c, a, s]$. Let P' and Q' be the packings obtained from P and Q by adding another set $Q'_s = P'_s = S^2 - \text{interior}(U)$. These packings obviously have nerve T'.

Now, the proof of 5.1 Case 4 can be applied here, because the fact that $Q'_s = P'_s$ is not a singularity, but a packed set, only makes things easier. We get $P' = Q'$, which gives $P = Q$.

There may be a situation where no such set U exists, as in Figure 5.5(a). However, we are free to manipulate the sets P_v, Q_v, $v = a, b, c$, provided we do not modify the parts of their boundaries that bound the component of $S^2 - (Q_a \cup Q_b \cup Q_c)$ that contains the other sets of the packings. Thus we easily reduce the situation to the case where such a U exists. See Figure 5.5(b). □

6. Some problems

One is naturally led to the following conjecture, which probably also occurred to other circle packers.

6.1. **Conjecture.** *Let T be an infinite triangulation with at most countably many ends. Then there exists a circle packing P in S^2 whose nerve is T and so that all the singularities of P are either circles or single points. This packing is unique, up to Möbius transformations.*

With the additional assumption that T has bounded valence, the uniqueness part of this conjecture can be proved. The proof uses the techniques presented here, and the fact that in the bounded valence case, two packings with the same triangulation as nerve induce homeomorphisms between the boundaries of the corresponding singularities, provided these boundaries are simple closed curves. This fact follows from the analogous property for quasiconformal maps, and perhaps might also be true without the bounded valence restriction.

As mentioned in the introduction, the conjecture holds when T has one end.

The observant reader may have noticed that our rigidity results for packings of convex sets other than circles do not deal with the case where infinity is (on the boundary of) a singularity of the packing. The reason for this is that we need homotheties to perturb the singularities, and the point at infinity is a fixed point for the homotheties. However, this difficulty does not rule out some kind of rigidity for packings having $\{\infty\}$ as a singularity. For example, one may ask:

6.2. *Problem.* Let P and Q be packings of smooth strictly convex bodies in the plane, both having a triangulation $T = T(V)$ as their nerve. Suppose that $\text{carrier}(P) = \text{carrier}(Q) = \mathbb{R}^2$; P_v is positively homothetic to Q_v for each v; and $P_v = Q_v$ for $v = a, b, c$, where $[a, b, c]$ is some triangle of T. Does it follow that $Q = P$?

Added in proof. Conjecture 6.1 is true; the proof will appear in a joint work with Zheng-Xu He.

ACKNOWLEDGMENTS

I am deeply thankful to my teachers Bill Thurston and Peter Doyle, and to Richard Schwartz, Burt Rodin, and Zheng-Xu He for stimulating discussions relating to packings.

REFERENCES

[An1] E. M. Andreev, *On convex polyhedra in Lobačevskiĭ spaces*, Mat. Sb. (N.S.) **81** (1970), 445–478; English transl. in Math. USSR Sb. **10** (1970), 413–440.

[An2] ———, *On convex polyhedra of finite volume in Lobačevskiĭ space*, Mat. Sb. (N.S.) **83** (1970), 256–260; English transl. in Math. USSR Sb. **12** (1970), 255–259.

[BFP] I. Bárány, Z. Füredi, and J. Pach, *Discrete convex functions and proof of the six circle conjecture of Fejes Tóth*, Canad. J. Math. **36-3** (1984), 569–576.

[CR] I. Carter and B. Rodin, *An inverse problem for circle packing and conformal mapping*, preprint.

[He1] Zheng-Xu He, *An estimate for hexagonal circle packings*, J. Differential Geom. (to appear).

[He2] ———, *Solving Beltrami equations by circle packing*, Trans. Amer. Math. Soc. (to appear).

[Ro1] B. Rodin, *Schwartz's lemma for circle packings*, Invent. Math. **89** (1987), 271–289.

[Ro2] ———, *Schwartz's lemma for circle packings* II, J. Differential Geom. **30** (1989), 539–554.

[RS] B. Rodin and D. Sullivan, *The convergence of circle packings to the Riemann mapping*, J. Differential Geom. **26** (1987), 349–360.

[Sch1] O. Schramm, *Packing two-dimensional bodies with prescribed combinatorics and applications to the construction of conformal and quasiconformal mappings*, Ph.D. thesis, Princeton, 1990.

[Sch2] ———, *Uniqueness and existence of packings with specified combinatorics*, Israel J. Math. (to appear).

[Ste] K. Stephenson, *Circle packings in the approximation of conformal mappings*, Bull. Amer. Math. Soc. **23** (1990), 407–415.

[Th1] W. P. Thurston, *The geometry and topology of 3-manifolds*, Princeton Univ. Lecture Notes, Princeton, NJ.

[Th2] ———, *The finite Riemann mapping theorem*, invited talk at the International Symposium in Celebration of the Proof of the Bieberbach Conjecture, Purdue University, March 1985.

DEPARTMENT OF MATHEMATICS, PRINCETON UNIVERSITY, PRINCETON, NEW JERSEY, 08544

Current address: Mathematics Department, University of California at San Diego, La Jolla, California 92093

E-mail address: oschramm@math.ucsd.edu

Invent. math. 107, 543–560 (1992)

Inventiones mathematicae
© Springer-Verlag 1992

How to cage an egg

Oded Schramm

University of California, San Diego, Department of Mathematics, La Jolla, CA 92093, USA

Oblatum 18-V-1991

Summary. This paper proves that given a convex polyhedron $P \subset \mathbb{R}^3$ and a smooth strictly convex body $K \subset \mathbb{R}^3$, there is some convex polyhedron Q combinatorically equivalent to P which *midscribes* K; that is, all the edges of Q are tangent to K. Furthermore, with some stronger smoothness conditions on ∂K, the space of all such Q is a six dimensional differentiable manifold.

1 Introduction

For any $n=3, 4, 5, \ldots$ one can find a convex n-gon in the plane so that all its vertices lie on the unit circle. In fact, the convex hull of any n distinct points on the unit circle gives such an n-gon. The situation in 3-space is rather different. Steinitz [St] has shown that there are convex polyhedra $P \subset \mathbb{R}^3$ so that there is no combinatorically equivalent polyhedron having all the vertices on the unit sphere. By duality, it follows that there are combinatorial types of convex polyhedra that are not realizable with all the (2-dimensional) faces tangent to the unit sphere.

Schulte [Schu] has generalized the above result of Steinitz, by showing that for pairs of integers (m, d) satisfying $0 \leqq m < d$, $d > 2$, $(m, d) \neq (1, 3)$ there are combinatorial types of d-dimensional polytopes which are not (m, d)-scribable; that is, they cannot be realized in d-space in such a way that all the m-dimensional faces are tangent to the unit $(d-1)$-sphere. The case $(m, d) = (1, 3)$, is an exception. Koebe [Koe] has claimed that every 3 dimensional convex polyhedron is combinatorically equivalent to a polyhedron which *midscribes* the unit sphere; that is, all its edges are tangent to the unit sphere. However, the proof in [Koe] is only for polyhedra which are simple or simplicial. The result for general polyhedra follows from Andreev's Theorem [An1, An2], as Thurston [Th, Chap. 13] observed. The midscribing polyhedron is unique, up to projective transformations which preserve the sphere.

Schulte also introduced the question of replacing the Euclidean sphere in the above by other convex bodies. It is the purpose of this paper to generalize the Koebe-Andreev-Thurston result in this direction by proving:

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_5, © Springer Science+Business Media, LLC 2011

1.1 Midscribability Theorem. *Let P be a (3-dimensional) convex polyhedron, and let $K \subset \mathbb{R}^3$ be a smooth strictly convex body. Then there exists a convex polyhedron $Q \subset \mathbb{R}^3$ combinatorially equivalent to P which midscribes K.*

With stronger smoothness conditions on K, one can say more:

1.2 Theorem. *In Theorem 1.1, if for some integer $k \geqq 1$ ∂K is C^{k+1}-smooth and has positive Gaussian curvature everywhere, then the space of such midscribing Q is a six dimension C^k-smooth manifold.*

This also implies that the space of midscribing Q is C^∞, if ∂K is C^∞ and has positive Gaussian curvature.

A few words about the proof of Theorem 1.1. The main part of the proof is establishing the theorem for the case where ∂K is C^2-smooth and has positive Gaussian curvature everywhere, and then the general case follows by approximations and convergence. We consider the space of all configurations; where a configuration is, by definition, an indexed collection of planes and points in $\mathbb{R}^3$; the points being indexed by the vertices of P, and the planes are indexed by the faces of P. A configuration corrresponds to a P-type K-midscribing polyhedron if the obvious conditions hold: the point corresponding to any vertex v of P belongs to every plane corresponding to a face of P containing the vertex v, the line segment joining the points corresponding to two vertices which share an edge in P is tangent to K, etc. With the presence of the stronger smoothness assumptions on K, these different conditions determine differentiable submanifolds of configuration space, and so, each point in the intersection of these submanifolds corresponds to a P-type K-midscribing polyhedron. We use geometric observations and a combinatorial argument to show that any intersection of these submanifolds is transverse. Transversality then implies that if we have a P-type K-midscribing polyhedron Q, then for a sufficiently small perturbation K' of K a K' midscribing P-type polyhedron Q' can be found. This enables us to transport midscribing polyhedra from one K to another, and the proof is then easily completed.

Certainly, the central part of this argument is the proof of transversality. The method in which we establish transversality is in many respects reminiscent of Cauchy's approach to rigidity of polyhedra ([Ca, Ro]). In a forthcoming paper the author intends to elaborate on this matter, and to use some of the techniques presented here to obtain new proofs of infinitesimal Cauchy rigidity and related results.

As noted above, Thurston's proof that the sphere can be midscribed by a polytope of given combinatorics is an application of Andreev's Theorem. As Thurston observed, Andreev's Theorem can be reinterpreted to obtain the Circle Packing Theorem, which says that any planar graph can be realized as the tangency graph of a circle packing on the sphere. In [Schr] the Circle Packing Theorem is generalized to packings of shapes other than circles. Using this generalization, Theorem 1.1 is proved there in the case that P is a simple or a simplicial polyhedron. Although this may not be apparent to the reader, the techniques of this paper are also closely related to the theory of packings in two dimensions. In fact, much of the arguments here were conceived as part of a differentiable proof giving uniqueness and existence of packings, in the spirit of [Schr]. Later, the method of [Schr] superseded the differentiable proof. Fortunately, this differentiable approach is most useful for the purposes of this work.

It is worthwhile to note that Thurston's route from circle packings to sphere-midscribing polyhedra can be reversed, and so we obtain a new proof of the Circle Packing Theorem as an additional reward for our efforts. (See 6.1.)

Acknowledgements. I would like to express thanks to Egon Shulte for suggesting to try and use my packing theorems to obtain polyhedra that midscribe given convex bodies. I am also very grateful to Bill Thurston, who advised me to try the differentiable approach to obtain a uniqueness result for packings.

2 Preliminaries

The main purpose of this section is to define some terms we later use, and to introduce notations. We freely use some of the most elementary notions in convexity (good sources for these are, for example, [Eg, Gr]), and these are defined here only if some confusion is anticipated. Similarly, we assume some very rudimentary knowledge of differential topology, especially transversality of submanifolds.

A *convex body* in $\mathbb{R}^3$ is a compact convex subset of $\mathbb{R}^3$ which has interior points. A *smooth* convex body is a convex body which has a unique supporting plane at every boundary point (this is equivalent to the boundary being C^1-smooth), and a *strictly convex* body, is a convex body whose boundary does not contain any nontrivial line segment.

Given points $x, y \in \mathbb{R}^n$, we use the notation $[x, y]$ to denote the line segment joining x and y.

A *polyhedron* in $\mathbb{R}^3$ is the convex hull of finitely many points in $\mathbb{R}^3$ which are not all coplanar. Let P be some arbitrary polyhedron in $\mathbb{R}^3$, which will be fixed henceforth. We denote the set of vertices of P by V, the set of edges by E, and the set of (2-dimensional) faces by F. A *P-type* polyhedron will mean a polyhedron which is combinatorically equivalent to P.

The *extended graph* of P, $G^+ = G^+(P)$, is the graph whose vertex set is $V \cup F$, and whose edges are $E \cup VF$, where VF denotes the collection of all pairs $\langle v, f \rangle$ so that $v \in V$, $f \in F$ and the vertex v belongs to the face f. A vertex i in the graph G^+ will be called a *V-vertex* if $i \in V$, or an *F-vertex* if $i \in F$. Hopefully, the fact that a face of P is a vertex of G^+ will cause no confusion. An edge $\langle i, j \rangle$ of G^+ is a V-V edge if $\langle i, j \rangle \in E$, and is termed a V-F edge if $\langle i, j \rangle \in VF$.

G^+ has an embedding in S^2, canonical up to homeomorphism, which can be visualized on the boundary of P by choosing an interior point in every face and connecting it to the vertices of that face. We will always consider G^+ with this embedding, which determines a triangulation of S^2. As above, if i, j are vertices of G^+ which are joined by an edge, then $\langle i, j \rangle$ will denote that edge. Similarly, the notation $\langle i, j, k \rangle$ will be used for a triangle of G^+; that is, a triangle in the triangulation induced by G^+.

A polyhedron $Q \subset \mathbb{R}^3$ is said to *midscribe* a smooth strictly convex body $K \subset \mathbb{R}^3$ if every edge of Q is tangent to K (that is, to ∂K).

2.1 Observations. *Let $K \subset \mathbb{R}^3$ be a smooth and strictly convex body, and let Q be a K-midscribing polyhedron.*

(1) *If m is a plane containing a face of Q, then the polygon $m \cap Q$ circumscribes the smooth convex set $m \cap K$ in the plane m.*

(2) *If v, u are vertices of Q, then the line segment $[v, u]$ intersects K.*

Proof. (1) is obvious. By definition, (2) holds if $\langle v, u\rangle$ is an edge of Q. Otherwise, let E_v be the set of edges of Q incident with v. Each edge $\langle v, w\rangle \in E_v$ contains some point $p_{\langle v, w\rangle} \in K$. But $[v, u]$ intersects the convex hull of the points p_e, $e \in E_v$, and K is convex. Therefore (2) holds. □

3 Configuration space and its submanifolds

In this section K will denote an arbitrary smooth strictly convex body in $\mathbb{R}^3$.

Let Z be the space $(\mathbb{R}^3)^V \times (Gr_2)^F$, where Gr_2 is the space of all oriented affine planes in $\mathbb{R}^3$. In other words, a point $z \in Z$ gives a choice of a point in $\mathbb{R}^3$ for each vertex $v \in V$ of P, and a choice of an oriented plane for each face $f \in F$. Z will be called the *configuration space* and a point $z \in Z$ will be called a *configuration*. For a configuration $z \in Z$ we denote by z_v the point in $\mathbb{R}^3$ corresponding to a vertex $v \in V$, and similarly, z_f will denote the plane corresponding to a face $f \in F$.

For each V-F edge $\langle v, f\rangle \in VF$, let $S_{\langle v, f\rangle}$ be the submanifold of Z consisting of the configurations $z \in Z$ for which z_v is contained in the plane z_f. $S_{\langle v, f\rangle}$ is a C^∞-smooth submanifold of Z having codimension 1.

Let $Z_+ = Z_+(K)$ be the set of configurations $z \in Z$ which avoid K in the sense that $z_v \notin K$ for each $v \in V$. Z_+ is an open set in Z.

Let Z_c be the set of configurations z so that for each face $f \in F$ all the points z_v corresponding to vertices v which do not belong to the face f are below the plane z_f. Because the plane $z_f \in Gr_2$ is oriented, it is meaningful to refer to the open half space *above* z_f, and to the one *below* z_f. Z_c is also open in Z.

For each V-V edge $\langle v, u\rangle$ of P, let $S_{\langle v, u\rangle} = S_{\langle v, u\rangle}(K)$ be the set of all configurations $z \in Z_+$ so that (the relative interior of) the line segment joining z_v and z_u is tangent to ∂K.

3.1 Lemma. *If ∂K is C^{k+1}-smooth (for some integer $k \geq 1$), and has positive Gaussian curvature everywhere, then $S_{\langle v, u\rangle}$ is a codimension 1 C^k-smooth submanifold of Z_+.*

The proof will be sketchy, because it has little relevance to our interests.

Proof. Let $\psi\colon \mathbb{R}^2 \to \partial K$ be a C^{k+1}-smooth coordinate chart for ∂K. Define a mapping

$$\tilde{\psi}\colon \mathbb{R}^2 \times (\mathbb{R}^2 - \{0\}) \times \mathbb{R}_+ \to (\mathbb{R}^3)^{\{v, u\}}$$

by

$$\tilde{\psi}(a, b, \lambda)_v = \psi(a) + d\psi_a(b), \qquad \tilde{\psi}(a, b, \lambda)_u = \psi(a) - \lambda d\psi_a(b).$$

(We identify $T_a\mathbb{R}^2$ with $\mathbb{R}^2$, and consider $T_{\psi(a)}\partial K$ as a linear plane in $\mathbb{R}^3$.) Then the segment $[\tilde{\psi}(a, b, \lambda)_u, \tilde{\psi}(a, b, \lambda)_v]$ is tangent to ∂K at $\psi(a)$. Because ψ is C^{k+1}-smooth, it follows that $\tilde{\psi}$ is C^k-smooth.

From the assumption that ∂K is positively curved it is not hard to (geometrically) deduce that $d\tilde{\psi}$ is everywhere nonsingular. It then follows that $\tilde{\psi} \times \mathrm{id}$ can be used for a coordinate chart for $S_{\langle v, u\rangle}$, where id is the identity map on $(\mathbb{R}^3)^{V - \{v, u\}} \times (Gr_2)^F$. The lemma follows. □

Suppose that Q is a P-type polyhedron in $\mathbb{R}^3$, and consider a particular combinatorial equivalence between Q and P. Then there is a configuration z corresponding to Q: for each $v \in V$, $z_v \in \mathbb{R}^3$ is the position of the vertex corresponding to v in Q, and for each $f \in F$, z_f is the plane containing the face corresponding to f in Q, oriented so that all the points z_v with v a vertex not in f are below z_f. Clearly, this configuration z will be in

$$Z_P \overset{\text{def}}{=} Z_c \cap (\cap_{\langle v, f\rangle \in VF} S_{\langle v, f\rangle}).$$

Conversely, a configuration contained in Z_P corresponds to a polyhedron Q of combinatorial type P. We will say that such a configuration *realizes* P, or that z is a *realization* of P, and $P(z)$ will denote the actual polyhedron in $\mathbb{R}^3$ which corresponds to z.

If a P-type polyhedron Q actually midscribes K, then the corresponding configuration z will be in

$$Z_{P,K} \overset{\text{def}}{=} Z_P \cap Z_+ \cap (\cap_{\langle v, u\rangle \in E} S_{\langle v, u\rangle}).$$

Conversely, every $z \in Z_{P,K}$ is a realization of a K-midscribing P-type polyhedron. So $Z_{P,K}$ is the set we are really interested in. It is left to the reader to convince himself that the above statements are obvious.

4 Transversality

The section is mostly devoted to the following theorem.

4.1 Transversality Theorem. *Let $K \subset \mathbb{R}^3$ be a strictly convex body with C^2-smooth positively curved boundary. Fix some edge $\langle v_0, v_1\rangle \in E$ of P, and let $f_2 \in F$ be one of the two faces containing this edge. Let x_0 and x_1 be two points in $\mathbb{R}^3 - K$ so that the line segment joining them is tangent to K, and let y_2 be an oriented plane containing x_0 and x_1. Denote by Z^i, $i=0, 1, 2$, the set of configurations $z \in Z$ so that $z_{v_0} = x_0$, $z_{v_1} = x_1$, and $z_{f_2} = y_2$, respectively. Let $E' \overset{\text{def}}{=} E - \{\langle v_0, v_1\rangle\}$, and let $VF' \overset{\text{def}}{=} VF - \{\langle v_0, f_2\rangle, \langle v_1, f_2\rangle\}$. Then the intersection*

$$Z'_{P,K} \overset{\text{def}}{=} Z_+ \cap Z_c \cap Z^0 \cap Z^1 \cap Z^2 \cap (\cap_{\langle i, j\rangle \in E' \cup VF'} S_{\langle i, j\rangle}) \tag{4.1}$$

is transverse.[1]

Note that $Z^0 \cap Z^1 \subset S_{\langle v_0, v_1\rangle}$, that $Z^0 \cap Z^2 \subset S_{\langle v_0, f_2\rangle}$, and that $Z^1 \cap Z^2 \subset S_{\langle v_1, f_2\rangle}$. Therefore $Z'_{P,K}$ is contained in $Z_{P,K}$. In fact, the configurations $z \in Z'_{P,K}$ are precisely the realizations of P-type polytopes which midscribe K and satisfy

[1] Recall that the intersection $\cap_n M_n$ of differentiable submanifolds of a differentiable manifold M is *transverse* if at every point z in the intersection the codimension of the intersection of the tangent spaces $\cap_n T_z M_n$ is equal to the sum of the codimensions $\sum_n \operatorname{codim}(T_z M_n)$, where the tangent spaces $T_z M_n$ are naturally considered as subspaces of $T_z M$. In particular, if the intersection is empty, then it is transverse.

the additional conditions $z_{v_0}=x_0$, $z_{v_1}=x_1$ and $z_{f_2}=y_2$. The proof of the theorem will be divided into several subsections.

4.2 Codimension count. Z_+ and Z_c are open in Z, and thus have codimension 0. Each of Z^0, Z^1, Z^2 has codimension 3 in Z. ($\dim(Gr_2)=3$). Each of the $S_{\langle v,u\rangle}$ and the $S_{\langle v,f\rangle}$ have codimension 1. So the sum of the codimensions for the intersection (4.1) is $3+3+3+|E'|+|VF'|$, which is equal to $6+|E|+|VF|$.

We will now show that $|VF|=2|E|$. Consider any V-F edge $\langle v,f\rangle\in VF$. There are precisely two edges of E which are contained in the face f and have v as one of their vertices. Conversely, for each edge $\langle w,u\rangle\in E$ there are four V-F edges $\langle v,f\rangle\in VF$ so that v is w or u and f is a face containing $\langle w,u\rangle$. Thus we have exhibited a 4 to 2 correspondence between VF and E. This establishes $|VF|=2|E|$.

Now we see that the sum of the codimensions is $6+3|E|$. By Euler's formula this is equal to $3|V|+3|F|=\dim(Z)$. Thus, in order to prove transversality, we must show that at each configuration z contained in the intersection (4.1), the intersection of the tangent spaces

$$T_z Z^0 \cap T_z Z^1 \cap T_z Z^2 \cap (\cap_{\langle i,j\rangle\in E'\cup VF'} T_z S_{\langle i,j\rangle}) \tag{4.2}$$

has dimension 0; i.e., contains only the zero vector. Here we think of $T_z Z^0$, the tangent space at z of the submanifold Z^0 of Z, as being contained in $T_z Z$, the tangent space of Z at z, and similarly for the other submanifolds of Z.

4.3 The labeling associated with a tangent vector. Assume that $z\in Z'_{P,K}$, and let $t\in T_z Z$ be a tangent vector at z. In the following, we will associate with t a labeling of the edges of G^+. Every edge of G^+ will be labeled with two labels, one label near each of its two endpoints. The possible labels are $+, 0, -$. We will use the notation $l_{i,j}$ to denote the label on the edge $\langle i,j\rangle$ near the vertex i. For $v\in V$ and for $f\in F$, $t_v\in T_{z_v}\mathbb{R}^3$ and $t_f\in T_{z_f} Gr_2$ will denote the components of t corresponding to v and f, respectively. (If $\pi_v\colon Z\to\mathbb{R}^3$ is the projection $\pi_v(z)=z_v$, then $t_v \overset{\text{def}}{=} d\pi_v(t)$, and similarly for t_f.)

Consider first an edge of type V-V, $\langle v,u\rangle\in E$. Because $z\in S_{\langle v,u\rangle}$, the line segment joining z_v to z_u is tangent to K. Therefore, since K is smooth, it follows that there is a unique plane, say $m_{\langle v,u\rangle}$, which is tangent to K and contains z_v and z_u. We will refer to the open half space determined by $m_{\langle v,u\rangle}$ which is disjoint from K as being *above* $m_{\langle v,u\rangle}$, and the other open half space determined by $m_{\langle v,u\rangle}$ as being *below* $m_{\langle v,u\rangle}$. With respect to t_v we discriminate between three possibilities: t_v either points above $m_{\langle v,u\rangle}$, points below $m_{\langle v,u\rangle}$, or is parallel to $m_{\langle v,u\rangle}$. It t_v points above $m_{\langle v,u\rangle}$, we label the part of the edge $\langle v,u\rangle$ which is near v with a $+$ (i.e., we set $l_{v,u}=+$). If it points below, we place there a $-$, and if parallel, a 0. Similarly, the vector t_u determines the labeling of $\langle v,u\rangle$ near u, and so this procedure defines the labeling for all the V-V edges.

Now consider any V-F edge $\langle v,f\rangle$. In the case where t_v points above z_f, we label $l_{v,f}=+$, if it points below z_f, we set $l_{v,f}=-$, and if parallel, $l_{v,f}=0$.

We define similarly the labeling of $\langle v,f\rangle$ near f. Informally, if every infinitesimal variation of z_f in the direction t_f gives a plane $y\in Gr_2$ having v below it, then we label with a $-$. If every such variation gives a plane $y\in Gr_2$ having v above it, then we label $l_{f,v}=+$. Otherwise, a 0.

In order to make this more precise, we need a convenient interpretation for t_f. Associate with each plane $y\in Gr_2$ the unique affine function on $\mathbb{R}^3$ which

is positive above y, negative below y, and whose gradient at every point has norm 1. Denote this function by $\phi(y)$. This defines an embedding ϕ of Gr_2 into the vector space $A(\mathbb{R}^3, \mathbb{R})$ of all affine functions on $\mathbb{R}^3$. In this way, to t_f corresponds the tangent vector $d\phi(t_f) \in T_{\phi(z_f)} A(\mathbb{R}^3, \mathbb{R})$. As $A(\mathbb{R}^3, \mathbb{R})$ is a vector space, the tangent space $T_{\phi(z_f)} A(\mathbb{R}^3, \mathbb{R})$ is canonically isomorphic to $A(\mathbb{R}^3, \mathbb{R})$, and thus we will regard $d\phi(t_f)$ as an affine function on $\mathbb{R}^3$. We set $l_{f,v} = +, 0, -$, depending on whether $d\phi(t_f)(v) > 0$, $d\phi(t_f)(v) = 0$, or $d\phi(t_f)(v) < 0$, respectively.

Having completed the definition of the labeling associated with t, we now discuss its properties. We continue our assumptions that $z \in Z'_{P,K}$, and that $t \in T_z Z$ is a tangent vector at z.

4.4 Lemma. *Let $\langle v, u\rangle$ be any* V-V *type edge, and let $e, f \in F$ be the two faces containing this edge. Then $l_{v,u}$, the label of $\langle v, u\rangle$ near v, is 'between' $l_{v,f}$ and $l_{v,e}$. Saying that a label l is between two labels means that if one of the two labels is '0', then l is equal to the other label, and if both labels are equal, then l is also equal to them. (l can be any of $+$, 0, $-$, if the two labels are $+$ and $-$.)*

Proof. There is clearly no loss of generality in assuming that $z_v = 0$, and, for simplicity of notation, we will adopt this assumption. Let w_1, $w_2 \in V$ be the neighbors of v other than u which belong to the faces e, f, respectively. We think of t_v as a vector in $\mathbb{R}^3$. Because z_{w_1}, z_{w_2} and z_u are linearly independent, we can write

$$t_v = -\lambda_1 z_{w_1} - \lambda_2 z_{w_2} + \lambda_3 z_u. \tag{4.3}$$

Now, on $[z_v, z_{w_1}]$ there is some point of K, and this point must be below $m_{\langle v,u\rangle}$, because K lies below $m_{\langle v,u\rangle}$, except for the unique point in $m_{\langle v,u\rangle} \cap K$. Since $z_v \in m_{\langle v,u\rangle}$, this implies that z_{w_1} is below $m_{\langle v,u\rangle}$. On the other hand z_{w_1} is on the plane z_e and is below z_f. A similar argument shows that z_{w_2} is on z_f and below z_e and $m_{\langle v,u\rangle}$. Noting that z_u and $0 = z_v$ are in $m_{\langle v,u\rangle} \cap z_e \cap z_f$, it follows from the above and from (4.3) that $l_{v,e} = \operatorname{sign}(\lambda_2)$, that $l_{v,f} = \operatorname{sign}(\lambda_1)$, and that $l_{v,u} = \operatorname{sign}(\lambda_1)$ if $\lambda_2 = 0$ or $\operatorname{sign}(\lambda_1) = \operatorname{sign}(\lambda_2)$, and that $l_{v,u} = \operatorname{sign}(\lambda_2)$ if $\lambda_1 = 0$. This shows that $l_{v,u}$ is between $l_{v,e}$ and $l_{v,f}$, and proves the lemma. □

4.5 Definitions. Consider our graph G^+ and the labeling of it induced by t, and let i be some vertex of G^+ (either a V-vertex, or an F-vertex). If there is some edge e of G^+ which is incident with i and has a nonzero label near i, then i will be called *live*. Otherwise, i is termed *dead*. Now suppose that $\langle i, j, k\rangle$ is a triangle of G^+. If the two labels near i on $\langle i, j\rangle$ and $\langle i, k\rangle$ are $+$ and $-$ (i.e., $\{l_{i,j}, l_{i,k}\} = \{+, -\}$), then we consider this as a sign change in the triangle $\langle i, j, k\rangle$ at the vertex i. If one of these labels is 0, and the other is $+$ or $-$, then we consider it as half a sign change. If both labels are 0, then it is not considered a sign change if i is dead, it is considered half a sign change if i is a live V-vertex, and it is considered a full sign change if i is a live F-vertex. See Fig. 4.1.

The reason for the difference in the way V- and F-vertices are treated is that we have included V-V edges in G^+, but did not include F-F edges. The following lemma provides some motivation for these definitions.

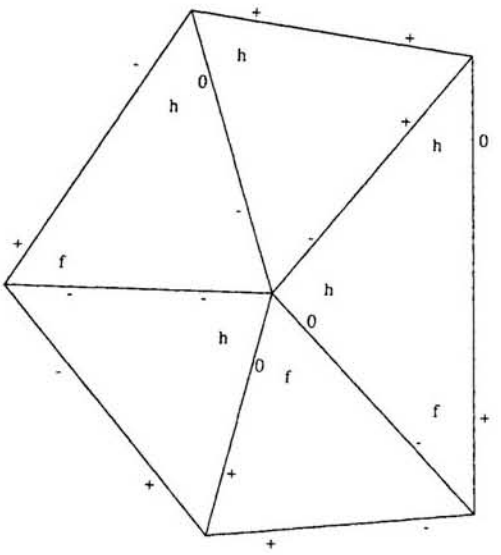


Fig. 4.1. An example of a portion of the labeled graph. The center vertex is an F-vertex, the others are V-vertices. A full sign change is marked by *f*, and half sign changes are marked *h*. (Non sign changes are unmarked)

4.6 Lemma (Sign changes around a vertex) *Let i be any* V-*vertex or* F-*vertex.* (a) *The total number of sign changes at i, counted according to the definition above, is at most* 2. (b) *i is a live vertex if, and only if, $t_i \neq 0$.*

Proof. Consider the case where i is an F-vertex. Recall the definition of the affine function $d\phi(t_i)$. Let g be the restriction of $d\phi(t_i)$ to the plane z_i, and let $V_i \subset V$ be the set of V-vertices which belong to the face i of P.

Note that the correspondence $v \to z_v$, restricted to V_i, gives a realization of the abstract polygon which is the i face of P as a polygon that lies in the plane z_i and whose vertices are the z_v, $v \in V_i$. Denote this realization by $P_i(z)$.

Consider now the case where i is a dead vertex; i.e., all the labels near i are 0. Then g is zero on all the vertices of $P_i(z)$, and, being affine, it follows that g is zero on the plane z_i. So, in that case, $d\phi(t_i)$ and $\phi(z_i)$ are both zero on z_i, and, consequently, $d\phi(t_i)$ is a scalar multiple of $\phi(z_i)$. But the gradient of all the affine functions in $\phi(Gr_2)$ is 1 and $d\phi(t_i) \in T_{\phi(z_i)}\ \phi(Gr_2)$. Therefore $d\phi(t_i)=0$, and $t_i=0$. Conversely, it $t_i=0$, then obviously $d\phi(t_i)=0$ and i is dead. Thus part (b) is proved in the case that i is an F-vertex.

If g is positive on all the vertices of $P_i(z)$, then all the labels near i will be $+$, and there are no sign changes at i. Similarly, if it is negative on all these points, then all the labels near i will be $-$. It remains to consider the case where g is not all positive, all negative, or all zero on z_v, $v \in V_i$. Then there is a line, say L, contained in z_i where g is zero, and g is positive on one half plane determined by this line and negative in the other. If the intersection of L with $P_i(z)$ is an edge of $P_i(z)$, then two adjacent labels around i are zero and the other labels are either all $+$ or all $-$. (The labels which are 0 are those corresponding to the vertices of $P_i(z)$ which are in L.) Thus the total number of sign changes around i is 2 in this case. Otherwise, L intersects the

boundary of $P_i(z)$ in at most two points. Every such intersection point corresponds to precisely one sign change in the total. (An intersection point in the relative interior of an edge of $P_i(z)$ corresponds to a sign change between a $+$ and a $-$. An intersection point which occurs at a vertex z_w of $P_i(z)$ corresponds to two half sign changes: each between the $l_{i,w}=0$ label corresponding to w and the labels corresponding to its neighbors in the face i.) This completes the proof in the case $i\in F$.

Suppose now that $i\in V$. The proof for this case will be, in some sense, dual to the above. Assume, for convenience of notation, that $z_i=0$. We think of t_i as a vector in $\mathbb{R}^3$. If $t_i=0$, then clearly all the labels near i are 0, and i is dead. Conversely, if $l_{i,f}=0$ for all faces $f\in F$ incident with i, then t_i belongs to all the corresponding planes z_f, and the intersection of these planes is $\{z_i\}=\{0\}$, so $t_i=0$. Thus (b) is verified.

To prove (a), we may, and will, assume $t_i\neq 0$. Denote by V_i the set of vertices $v\in V$ which neighbor with i, and denote by C the convex hull of $P(z)\cup\{t_i\}$. Because $t_i\neq 0=z_i$, there are at most two faces of C which contain $[t_i, z_i]$. Each face of C which contains $[t_i, z_i]$ contains at most two points of the type z_v, $v\in V_i$. Let $v\in V_i$, and let $f, g\in F$ be the two faces which contain the edge $\langle i, v\rangle$. By the above, it is sufficient that we prove that the sum of the sign changes at i in the triangles $\langle f, i, v\rangle$ and $\langle v, i, g\rangle$ is

(1) 0, if z_v does not belong to a face of C which contains $[t_i, z_i]$,
(2) $\frac{1}{2}$, if z_v belongs to only one face of C containing $[t_i, z_i]$, and this face contains some other point of the form z_u, $u\in V_i$,
(3) at most 1, in any case.

For the proof of (1), suppose that z_v does not belong to a face of C which contains $[t_i, z_i]$. Then the affine hull of $\{t_i, z_i, z_v\}$ must contain interior points of $P(z)$. Let $p=\lambda_1 z_v+\lambda_2 t_i$ be such an interior point. (Recall $z_i=0$.) p is below z_f and below z_g, but 0, $z_v\in z_f\cap z_g$, and therefore t_i is either below z_f and below z_g (if $\lambda_2>0$), or above both of them (if $\lambda_2<0$). This implies that $l_{i,f}$ and $l_{i,g}$ are either both $+$ or both $-$. Using Lemma 4.4, it follows that $l_{i,v}=l_{i,f}=l_{i,g}$, and we see that (1) holds.

To verify (2), assume now that z_v and z_u belong to a face, say C_0, of C, containing $[t_i, z_i]$, $v\neq u\in V_i$, and that C_0 is the only face of C containing t_i, z_v, z_i. The first assumption, implies that u must belong to the face f or to g, say to f. Therefore, because C_0 and z_f both contain the points z_i, z_v, z_u, C_0 is contained in z_f. Thus $t_i\in z_f$, and $l_{i,f}=0$. On the other hand, the second assumption shows that z_g does not contain a face of C containing $[t_i, z_i]$. This gives $t_i\notin z_g$, and therefore $l_{i,g}\neq 0$. Using this and $l_{i,f}=0$, Lemma 4.4 gives $l_{i,g}=l_{i,v}\neq l_{i,f}=0$, and (2) follows.

(3) follows immediately from Lemma 4.4, and this completes the proof of the lemma. □

4.7 Orientation Lemma. *Let $\langle i,j\rangle$ be an edge in G^+, and suppose that $t\in T_z S_{\langle i,j\rangle}$. Then the labels $l_{i,j}$ and $l_{j,i}$ are complementary. That is, $l_{i,j}=+$ if $l_{j,i}=-$, $l_{i,j}=0$ if $l_{j,i}=0$, and $l_{i,j}=-$ if $l_{j,i}=+$.*

If we suppose that $t\in\bigcap_{\langle i,j\rangle\in E\cup VF} T_z S_{\langle i,j\rangle}$, then the lemma implies that on every edge of G^+ the labels are complementary, and so we can associate with t a partial orientation of G^+: an edge is oriented from $+$ to $-$, and the edges

labeled with two 0's remain unoriented. This is the reason for calling this lemma the 'Orientation Lemma'.

Proof of Lemma 4.7 Consider the case where $\langle i,j\rangle$ is an edge of type V-V. Let $z(s)$ be a smooth curve in $S_{\langle i,j\rangle}$ satisfying $z(0)=z$ and $\left.\frac{dz(s)}{ds}\right|_{s=0}=t$. If $l_{i,j}=+$, the t_i points above the plane $m_{\langle i,j\rangle}$, and for $s>0$ sufficiently small $z(s)_i$ will be above this plane. If also $l_{j,i}=+$, then $z(s)_j$ will be above that plane for $s>0$ sufficiently small. But this would mean that the line segment $[z(s)_j, z(s)_i]$ joining $z(s)_j$ and $z(s)_i$ lies above this plane, which contradicts $z(s)\in S_{\langle i,j\rangle}$, because K does not intersect the half space above $m_{\langle i,j\rangle}$. Thus we see that two + labels cannot occur on the edge $\langle i,j\rangle$.

If $l_{i,j}=+$, $l_{j,i}=0$, then, for sufficiently small $s>0$, the segment $[z(s)_j, z(s)_i]$ is above $m_{\langle i,j\rangle}$, except possibly for points arbitrarily close to z_j. Again this gives a contradiction, because K is disjoint from a neighborhood of z_j and does not intersect the half space above $m_{\langle i,j\rangle}$. Thus we see that a + and a 0 are impossible.

Because $-t$ is also in $T_z S_{\langle i,j\rangle}$, we see that two − labels, or a − and a 0 are also impossible. The only remaining possibilities are + and −, or two 0's, as required.

The proof for the case that $\langle i,j\rangle$ is a V-F type edge is similar (and simpler). □

4.8 Lemma (Sign changes in a quadrilateral) *Let $\langle v, u\rangle$ be a* V-V *type edge in G^+, and let the two faces of P which contains this edge be $e, f\in F$. Suppose that $t\in T_z S_{\langle v,u\rangle}\cap T_z S_{\langle v,e\rangle}\cap T_z S_{\langle v,f\rangle}\cap T_z S_{\langle u,e\rangle}\cap T_z S_{\langle u,f\rangle}$. If among v, u, e, f there is one live vertex, then the sum of the total number of sign changes in both triangles $\langle v, u, e\rangle$ and $\langle v, u, f\rangle$ of G^+ is at least one.* (*See Definition* 4.5.) *If among these four vertices there is more than one live vertex, then this number of sign changes is at least two.*

Proof. No new geometric arguments will appear in the proof – the lemma is a combinatorial consequence of the Orientation Lemma and Lemma 4.4. Consider some triangle $\langle i, j, k\rangle$ of G^+, and assume that not all the labels in it are 0 and that $t\in T_z S_{\langle i,j\rangle}\cap T_z S_{\langle j,k\rangle}\cap T_z S_{\langle k,i\rangle}$. We will show, by a case by case analysis, that the total number of sign changes in this triangle is at least one. Suppose, without loss of generality, that $l_{i,j}\neq 0$, and suppose, by symmetry, that $l_{i,j}=+$. Then $l_{j,i}=-$, by the Orientation Lemma. We check the three possibilities for the label $l_{i,k}$. If $l_{i,k}=-$, then we already have one full sign change in $\langle i, j, k\rangle$ at i. Consider the possibility $l_{i,k}=0$. Then there is half a sign change at i, and $l_{k,i}=0$, again by the Orientation Lemma. If $l_{k,j}=+, -$, then we also have half a sign change at k, and so only $l_{k,j}=0$ needs to be discussed. This however leads to $l_{j,k}=0$, and the half sign change appears at j. It remains to check $l_{i,k}=+$, and, by symmetry, only the possibility $l_{j,k}=-$ needs consideration. But then $l_{k,i}=-$ and $l_{k,j}=+$, giving a sign change at k.

We return to the situation of the lemma. Assume that the total number of sign changes in $\langle v, u, e\rangle$ and $\langle v, u, f\rangle$ is less than 2. By the above, it follows that in at least one of these triangles, say in $\langle v, u, e\rangle$, all the labels are 0. In particular, we have $l_{v,u}=l_{v,e}=0$. From Lemma 4.4 it then follows that $l_{v,f}=0$ also, and by symmetry $l_{u,f}=0$. From the Orientation Lemma we

get $l_{f,u}=l_{f,v}=0$, and so all the labels on both triangles are 0. If, for instance, v is live, than there are two half sign changes at v, one in each triangle. Thus there is a total of one sign change in our quadrilateral (the two neighboring triangles) at each live V-vertex. Also, there clearly is one sign change for each live F-vertex. This completes the proof of the lemma. □

Proof of Theorem 4.1 Suppose now that t is in the intersection (4.2). As we pointed out in the subsection devoted to the codimension count, it is enough to prove $t=0$. Consider the graph G^+ with the labeling induced by t. Note that $t\in S_{\langle i,j\rangle}$ for every $\langle i,j\rangle\in E\cup VF$, and so Lemma 4.8 can be freely applied.

The basic idea is the oldest trick in the book of combinatorics – a double count. We will count the number of sign changes in two different ways. On the one hand, Lemma 4.6 tells us that there are 'few' sign changes, and on the other hand Lemma 4.8 shows that there are 'many' sign changes.

Each triangle of G^+ has one V-V edge, and two V-F edges. We pair the triangles of G^+ across V-V edges, to form quadrilaterals. Let Q_1 be the set of such quadrilaterals which have exactly one live vertex, and let Q_2 be the set of such quadrilaterals which have at least two live vertices. By Lemma 4.8, the total number of sign changes in each quadrilateral in Q_n is at least n ($n=1, 2$). So the total number of sign changes in G^+ is at least $|Q_1|+2|Q_2|$. But Lemma 4.6 implies that the total number of sign changes in G^+ is at most $2l$, where l is the number of live vertices in G^+. Thus we arrive at the inequality

$$|Q_1|+2|Q_2|\leqslant 2l. \tag{4.4}$$

We think of this inequality as saying that there is a set of vertices L, the set of live vertices, which are incident to relatively few quadrilaterals; relative in relation to $l=|L|$. This is like having a set which is large when compared to the size of its boundary.

Our intention is to prove $l=0$. Then Lemma 4.6(b) gives $t_i=0$ for each vertex i, and thus $t=0$, as required. There are fairly direct (traditional) ways to prove $l=0$ from inequality (4.4), and from the fact that there is a triangle of dead vertices $-\langle v_0, v_1, f_2\rangle$. We prefer to take a shortcut.

Let G' be the planarly embedded graph obtained from G^+ in the following manner. In each quadrilateral of Q_1 that has a V-type live vertex delete the V-V edge, and insert into the quadrilateral an F-F edge joining the two F-type vertices of that quadrilateral. Because the graph G' is planar, G' can be embedded in the plane so that all its edges are straight line segments (see [Tu], or prove by induction on the number of vertices). Furthermore, there is such a straight embedding so that the unbounded region in the complement of the embedded G' is the region corresponding to the triangle $\langle v_0, v_1, f_2\rangle$, whose vertices are all dead. We think of G' as being thusly embedded.

We estimate in two different ways the sum of the angles of G' at the live vertices. On the one hand, this is clearly $2\pi l$. To make another estimate of this quantity, let q be some quadrilateral in Q_1, and let i be the live vertex of q. By our construction of G', the angle of q at i is an interior angle of a triangle, and is thus smaller than π. Since the sum of the angles of any quadrilateral is 2π (except for the quadrilateral containing $\langle v_0, v_1, f_2\rangle$), the sum of angles at live vertices for any quadrilateral in Q_2 is at most 2π, and is in fact smaller

than 2π unless all its vertices are live. Summing over all quadrilaterals the total of the angles at the live vertices we see that

$$\pi|Q_1|+2\pi|Q_2|\geqq 2\pi l.$$

Comparison with (4.4) shows that this inequality must be an equality. Thus our estimates for each quadrilateral are, in fact, realized. This shows that the vertices in every quadrilateral are either all dead, or all alive. By connectedness, since there are some dead vertices, all the vertices are dead. This completes the proof of Theorem 4.1. □

We now obtain corollaries from the Theorem.

4.9 Corollary (Rigidity) *$Z'_{P,K}$ is a discrete set.*

Proof. Because (4.1) is a transverse intersection of C^1-smooth submanifolds, the intersection $Z'_{P,K}$ is a C^1-smooth submanifold whose codimension is equal to the sum of the codimensions. By our codimension count, this means that $Z'_{P,K}$ is a 0-dimensional submanifold; that is, it is discrete. □

The following elementary lemma will be needed for the proof of Theorem 1.2.

4.10 Lemma. *Let $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m, M$ be C^1 submanifolds of a C^1 manifold Z. Suppose that*

$$z\in(\cap_{i=1}^n X_i)\cap(\cap_{i=1}^m Y_i) \tag{4.5}$$

and that intersection is transverse at z. Further suppose that

$$M\supset\cap_{i=1}^n X_i.$$

Then the intersection $M\cap(\cap_{i=1}^m Y_i)$ is transverse at z.

Proof. We will use the notation $\dot{X}_i$ for the tangent space of X_i at z, $T_z X_i$. Similarly $\dot{Y}_i=T_z Y_i$, $\dot{M}=T_z M$. When $U\subset W$ are vector subspaces of $T_z Z$ we will denote the codimension of U in W by $\operatorname{codim}(U, W)$. The codimension of U in $T_z Z$ is denoted by $\operatorname{codim}(U)$. In the following computation we use the fact that $\operatorname{codim}(\cap_i U_i, W)\leqq\sum_i \operatorname{codim}(U_i, W)$, and that $\operatorname{codim}(U)=\operatorname{codim}(U, W)+\operatorname{codim}(W)$, if $U, U_i\subset W$. Using the transversality of the intersection (4.5) and these properties, we have:

$$\begin{aligned}
&\sum_{i=1}^n \operatorname{codim}(\dot{X}_i)+\sum_{i=1}^m \operatorname{codim}(\dot{Y}_i)+\operatorname{codim}(\dot{M})\\
&\quad=\operatorname{codim}((\cap_{i=1}^n \dot{X}_i)\cap(\cap_{i=1}^m \dot{Y}_i))+\operatorname{codim}(\dot{M})\\
&\quad=\operatorname{codim}((\cap_{i=1}^n \dot{X}_i)\cap(\cap_{i=1}^m \dot{Y}_i), \dot{M})+2\operatorname{codim}(\dot{M})\\
&\quad\leqq\operatorname{codim}(\cap_{i=1}^n \dot{X}_i, \dot{M})+\operatorname{codim}(\cap_{i=1}^m \dot{Y}_i\cap\dot{M}, \dot{M})+2\operatorname{codim}(\dot{M})\\
&\quad=\operatorname{codim}(\cap_{i=1}^n \dot{X}_i)+\operatorname{codim}(\cap_{i=1}^m \dot{Y}_i\cap\dot{M})\\
&\quad\leqq\sum_{i=1}^n \operatorname{codim}(\dot{X}_i)+\sum_{i=1}^m \operatorname{codim}(\dot{Y}_i)+\operatorname{codim}(\dot{M}).
\end{aligned}$$

This show that equality holds throughout, and therefore $\operatorname{codim}(\cap_{i=1}^{m} \dot{Y}_i \cap \dot{M}) = \operatorname{codim}(\dot{M}) + \sum_{i=1}^{m} \operatorname{codim}(\dot{Y}_i)$. That is the required transversality at z, and the proof is complete. □

Proof of Theorem 1.2 We do not yet prove the implicit claim that $Z_{P,K}$ is nonempty, but prove the rest. Consider some point $z \in Z_{P,K}$. For showing that $Z_{P,K}$ is a six dimensional submanifold near z there is clearly no loss of generality in assuming that $z \in Z'_{P,K}$, because of the freedom in the choice of x_0, x_1, y_2. Thus we will make this assumption.

Consider the set $M = Z_+ \cap S_{\langle v_0, v_1 \rangle} \cap S_{\langle v_0, f_2 \rangle} \cap S_{\langle v_1, f_2 \rangle}$. It is easy to see that M is a C^k-smooth submanifold having codimension 3, by showing that the intersection defining M is transverse. (Recall that the $S_{\langle v,u \rangle}$ are C^k-smooth, by Lemma 3.1). In fact, the transversality of that intersection follows from Theorem 4.1, because we could have chosen other vertices and another face in place of v_0, v_1, f_2.

We have

$$Z_{P,K} = M \cap Z_c \cap (\cap_{\langle i,j \rangle \in E' \cup VF'} S_{\langle i,j \rangle}). \tag{4.6}$$

Since $M \supset Z^0 \cap Z^1 \cap Z^2$ and the intersection (4.1) is transverse, it follows from Lemma 4.10 that the intersection (4.6) is transverse at z. Therefore, in a neighborhood of z, $Z_{P,K}$ is a C^k-smooth submanifold whose codimension is equal to the sum of the codimensions in the intersection (4.6). So $\operatorname{codim}(Z_{P,K}) = 3 + |E'| + |VF'| = \operatorname{codim}(Z'_{P,K}) - 6$, using our results from the codimension count subsection. Thus $\dim(Z_{P,K}) = 6$, and this completes the proof. □

Remark. It is probably true, and perhaps not too hard to prove, that the assumptions in Theorem 1.1 are sufficient to guarantee that $Z_{P,K}$ is a 6-manifold, perhaps not smooth.

5 Existence of midscribing configurations

In this section we will prove the Midscribability Theorem 1.1. The basic method is the following. We start with some realization of P, $z^0 \in Z_P$, and show that there is a convex body, K^0, which has positively curved C^2-smooth boundary, and which $P(z^0)$ midscribes. Then we consider a curve of C^2-smooth convex bodies with positively curved boundary, K^s, $0 < s < 1$, which joins K^0 to $K^1 = K$. Take some $s \in [0, 1)$, and assume that K^s can be P-midscribed. It easily follows from Theorem 4.1 applied to K^s that $K^{s'}$ can be P-midscribed for all $s' \in [0, 1]$ sufficiently close to s. This shows that the set of $s \in [0, 1)$ for which K^s is P-midscribable is a relatively open set. If we could show that the set of $s \in [0, 1]$ for which K^s is P-midscribable is a closed set, then it would follow that this set is $[0, 1]$, and that $K = K^1$ is P-midscribable. To show closure, the natural approach is to take a limit of midscribing configurations, and to claim that this gives a midscribing configuration. The difficulty in doing this is that the points z_v may escape to infinity, or plunge into K^s. That is where the proof needs a little care, and where the details come in.

The following lemma will enable us to make sure that the vertices stay within a compact subset of $\mathbb{R}^3$.

5.1 Lemma. *For any smooth strictly convex body K and for any polyhedron Q midscribing K, there are at most two vertices of Q whose distances from K are greater than* 2 diameter(K).

Proof. Recall that if $p, q \in \mathbb{R}^3$ are two distinct vertices of Q, then the line segment joining p and q must intersect K. (Observation 2.1(2).) Now consider three distinct vertices of Q, p, q, $r \in \mathbb{R}^3$, and assume, without loss of generality, that $0 \in K$. If the angles between the vectors p and q, and between the vectors p and r are $\geqq 2\pi/3$, then the angles between $-p$ and q, and between $-p$ and r are $\leqq \pi/3$. This would imply that the angle between q and r is $\leqq 2\pi/3$. Therefore, at least one of the three angles determined by p, q, r at 0 is $\leqq 2\pi/3$. Assume that this is the angle between p and q. Let x be the point closest to 0 on the line segment joining p and q. Because the angle between p and q is at most $2\pi/3$, it follows that $2\|x\| \geqq \text{minimum}(\|q\|, \|p\|)$. But $\|x\| \leqq \text{diameter}(K)$, because $0 \in K$, and x is the closest point to 0 on the segment joining p and q which contains some point of K. Thus $\text{minimum}(\|q\|, \|p\|) \leqq 2\,\text{diameter}(K)$, and the lemma follows. □

5.2 Lemma. *Let v_0, v_1 be two neighboring vertices of P. For every constant $C>0$, there exists a realization of P, $z \in Z_P$, with the following property. The edge joining z_{v_0} and z_{v_1} has length C, but the distance from the midpoint of this edge to any vertex z_v, $v \neq v_0, v_1$ is less than $\frac{1}{2}$.*

Proof. We start with an arbitrary realization $Q = P(z)$ of P, and modify Q by a projective transformation. By applying an expansion, if necessary, assume that the edge $e = [z_{v_0}, z_{v_1}]$ has length C. Pick a plane L_0 in $\mathbb{R}^3$ which contains the edge e, but is otherwise disjoint from Q. Now think of $\mathbb{R}^3$ as being embedded in $\mathbb{R}^4$, and let $L \subset \mathbb{R}^4$ be an affine three space whose intersection with $\mathbb{R}^3$ is L_0. *Let* x be the midpoint of the edge e, and let o be a point in $\mathbb{R}^4 - \{L \cup \mathbb{R}^3\}$ whose distance to x is some small $\varepsilon > 0$ and which is separated from $Q - e$ by L.

For any point $y \in \mathbb{R}^3$, if the line through y and o intersects L, then let $f(y)$ be the point in the intersection of this line with L. The mapping f defined thusly is a projective map from $\mathbb{R}^3$ to L. (To be more precise, f extends to a projective map from real projective 3-space to the projective closure of L.) The image of Q, $Q' = f(Q)$, is a realization of P in L. The points on the edge e remain fixed under f.

On the other hand, let h be the distance from L to the vertices of Q other than z_{v_0}, z_{v_1}, let y be some vertex of Q other than z_{v_0}, z_{v_1}, and let l be the length of the part of $[y, o]$ which lies in the same side of L as y. As Fig. 5.1 illustrates,

$$d(o, f(y)) = \frac{d(o, L)\, l}{d(y, L)}.$$

Using $d(o, L) \leqq d(o, x) = \varepsilon$, $l \leqq d(y, o) \leqq d(y, x) + d(x, o)$, and $d(y, L) \geqq h$, we see that

$$d(o, f(y)) \leqq \frac{\varepsilon(\text{diameter}(Q) + \varepsilon)}{h}.$$

Thus, because the points x, z_{v_0}, z_{v_1} remain fixed under f, the polyhedron $Q' = f(Q)$ satisfys the required conditions, provided ε is sufficiently small. □

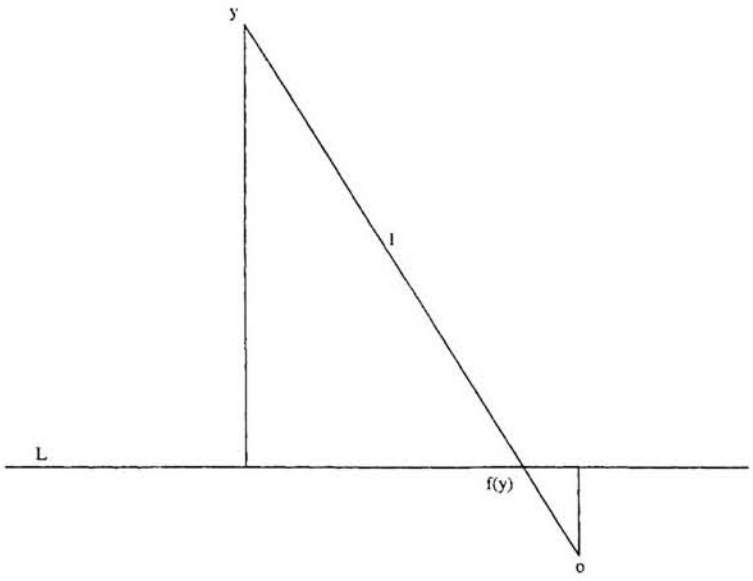


Fig. 5.1

Proof of Theorem 1.1 Let v_0, v_1, f_2 be as in Theorem 4.1. We start with a realization of P, $P(z^0)$, $z^0 \in Z_P$, which satisfies the conclusion of Lemma 5.2 with $C=6$. For each edge e of $P(z^0)$, let a_e be the midpoint of the edge. From the properties of $P(z^0)$ it follows that the diameter of the set of all such a_e is <1. For each edge e of $P(z^0)$ choose some plane m_e which supports $P(z^0)$ at that edge; i.e., $P(z^0) \cap m_e = e$. Now let K^0 be a C^2-smooth strictly convex body of diameter <1 with positively curved boundary and whose boundary contains each a_e and is tangent to m_e at a_e (for each edge e of $P(z^0)$). Clearly, such a K^0 exists. (To get an explicit construction, one can start with the convex hull of small balls B_e, with each B_e tangent to m_e at a_e from the side of m_e which contains interior$(P(z^0))$, and then minutely modify this convex body away from the a_e to make its boundary positively curved and C^2). By construction, the polyhedron $P(z^0)$ midscribes K^0.

Denote by x_0 and x_1 the two vertices $z^0_{v_0}$ and $z^0_{v_1}$ of $P(z^0)$, let e_0 be the edge of $P(z^0)$ joining them, and let $y_2 = z^0_{f_2}$. Let K^1 be a copy of K, rescaled and translated so that it has diameter <1 and is tangent to m_{e_0} at a_{e_0}, from the same side that K^0 is. It is clearly enough if we show that K^1 can be P midscribed.

Now choose a curve K^s, $s \in [0, 1]$, of strictly convex bodies joining K_0 and K_1. We require that each of the sets K^s, $s \in (0, 1)$, is tangent to m_{e_0} at a_{e_0}, that its boundary be C^2-smooth and positively curved, and that its diameter be less than 1 (and the curve must be continuous, in the Hausdorff metric). Clearly, there exists such a curve.

Let A be the set of $s \in [0, 1]$ such that there exists a realization $z^s \in Z_P$ which midscribes K^s and satisfys $z^s_{v_0} = x_0$, $z^s_{v_1} = x_1$, $z^s_{f_2} = y_2$. Clearly, $0 \in A$, because $P(z^0)$ midscribes K^0. Our goal is to show that $1 \in A$, by demonstrating that $A \cap [0, 1)$ is relatively open and A is closed in $[0, 1]$.

Showing that $A \cap [0, 1)$ is open in $[0, 1]$ is easy after the preparations in the previous section: Consider an $s \in A - \{1\}$. The K^s midscribing configuration

z^s is in Z'_{P,K^s}. By Theorem 4.1, applied to K^s, the intersection (4.1) is transverse at z^s. This implies that if the intersecting submanifolds in (4.1) are perturbed slightly, an intersection close to z^s remains. For K' a smooth strictly convex body sufficiently close to K^s, the corresponding submanifolds are close to those of K^s. (For each $\langle u, w\rangle \in E'$ there is a homeomorphism of $S_{\langle u,w\rangle}(K^s)$ to $S_{\langle u,w\rangle}(K')$ which is arbitrarily close to the identity on compact subsets of Z_+, if K' is sufficiently close to K^s. For example, one can take as this homeomorphism the mapping $z \to z'$ so that $z'_i = z_i$ for $i \in F \cup V - \{u, w\}$, and $z'_u = z_u + b' - b$ and $z'_w = z_w + b' - b$, where b is the point of tangency of $[z_u, z_w]$ with K^s and b' is the point of $\partial K'$ which has the same outer unit normal in K' as p has in K^s. The submanifolds other than Z_+ in the intersection (4.1) which are not of the form S_e, $e \in E'$, are the same for K and K'.) Thus $Z'_{P,K^{s'}}$ is nonempty for $s' \in [0, 1]$ sufficiently close to s, proving that $A - \{1\}$ is open in $[0, 1)$.

To prove that A is closed in $[0, 1]$, let $r_1, r_2, r_3, \ldots$ be a sequence in A converging to some $r \in [0, 1]$. Since $r_n \in A$, $n = 1, 2, \ldots$, there exists a configuration, z^{r_n} say, in $Z'_{P,K^{r_n}}$. The length of the edge $[x_0, x_1]$ is 6, and the midpoint of this edge is in every K^s, and therefore, because diameter$(K^s) < 1$, the distance from x_0 or x_1 to K^s is greater than 2 diameter(K^s). Together with Lemma 5.1, this implies that for every vertex $v \in V$ the sequence $\{z^{r_n}_v : n = 1, 2, \ldots\}$ is bounded. From this we also conclude that for every $f \in F$ the sequence $\{z^{r_n}_f : n = 1, 2, \ldots\}$ is contained in a compact subset of Gr_2, because each plane $z^{r_n}_f$ contains some vertex $z^{r_n}_v$. Therefore there is some configuration $z \in Z$ which is an accumulation point for the sequence $\{z^{r_n}\}$. Since we can replace the sequence $\{r_n\}$ by some subsequence, we may, and will, assume that actually $z^{r_n} \to z$ as $n \to \infty$.

We want to show that z is a midscribing configuration for K^r. It follows from the convergence of z^{r_n} to z and the corresponding property for z^{r_n} that z_v lies in the plane z_f, whenever the vertex $v \in V$ belongs to the face $f \in F$. Another consequence of the convergence is that for $v, u \in V$ $[z_v, z_u]$ intersects K^r (recall Observation 2.1(2)), and is tangent to K^r, if $\langle v, u\rangle \in E$. 'Tangent' here is taken in a broad sense – allowing the tangency to be at an endpoint of the segment $[z_v, z_u]$, or even allowing $z_v = z_u \in \partial K^r$. It still requires proof to show that $z_v \notin K^r$, $v \in V$.

To prove that $z_v \notin K^r$, $v \in V$, let V_0 be a connected component (in the graph of P) of the set of vertices $v \in V$ for which $z_v \in K^r$. Because $[z_v, z_u]$ is tangent to K^r for $\langle v, u\rangle \in E$, it follows that all the points z_v, $v \in V_0$ are the same point, say p_0, and that $p_0 \in \partial K^r$.

Let $f \in F$ be any face of P, and denote by V_f the collection of vertices of P which belong to f. Because $P(z^{r_n})$ midscribes K^{r_n} for each n, we know that the convex hull of $\{z^{r_n}_v : v \in V_f\}$ contains $K^{r_n} \cap z^{r_n}_f$ (Observation 2.1(1)). In the limit, since K^r is strictly convex, it follows that the convex hull of $\{z_v : v \in V_f\}$ contains $K^r \cap z_f$.

Now, the face z_{f_2} intersects the interior of K^r, since $z_{f_2} = z^0_{f_2} = y_2$, which is not equal to m_{e_0}, but passes through the point a_{e_0} where m_{e_0} and K^r are tangent (because $z_{f_2} \supset \{x_0, x_1\}$). Using the above fact that the convex hull of $\{z_v : v \in V_{f_2}\}$ contains the two dimensional smooth set $K^r \cap z_{f_2}$, we see that at least three points in $\{z_v : z \in V_{f_2}\}$ are not in K^r. Therefore $V - V_0$ contains at least three vertices.

Now let U_0 be the set of vertices in $V - V_0$ which neighbor with some vertex in V_0. For each $v \in U_0$, the segment $[z_v, p_0]$ is tangent to K^r at p_0. Therefore, all the points z_v, $v \in U_0$, lie in the plane, m_{p_0} say, which is tangent to K^r at

p_0. Because K^r is strictly convex, p_0 is the only point of K^r in m_{p_0}. Therefore, if u, v are distinct vertices in U_0, then $[z_u, z_v]$ must pass through p_0, since $[z_u, z_v]$ has to intersect K^r. This implies that there are at most 2 distinct vertices in U_0. But, as noted above, there are more than 2 vertices in $V-V_0$, and we see that U_0 is a set of at most two vertices which separates the graph of P. This contradicts the well known elementary fact that graphs of polyhedra are 3-connected (see [Gr, Theorem 11.3.1]), and this contradiction shows that $V_0=\emptyset$. Therefore $z_v \notin K^r$ for $v \in V$.

To show that $z \in Z'_{P,K^r}$, all that remains to be seen is that $z \in Z_c$. For this, let $f \in F$ be a face of P, and let $w \in V$ be some vertex which does not belong to the face f. Because for each $n=1, 2, 3, \ldots$, z^r_w is below the plane z^r_f, in the limit, z_w is either on z_f, or below z_f. We have noted above that the convex hull of $\{z_v: v \in V_f\}$ must contain $z_f \cap K^r$. From this and $z_w \notin K^r$, we see that if $z_w \in z_f$, then the segment $[z_w, z_v]$ will be disjoint from K^r for some $v \in V_f$, but we know that that is impossible. Therefore z_w is below the plane z_f, and this establishes $z \in Z_c$.

Having shown that $z \in Z'_{P,K^r}$, we conclude that $r \in A$. Thus A is a closed set. Because $A-\{1\}$ is open and closed in $[0, 1)$, and is nonempty ($0 \in A$), it follows that $A-\{1\}=[0, 1)$. Now $1 \in A$, because A is closed in $[0, 1]$. Therefore K^1, and also K, can be P-midscribed. This completes the proof. □

6 Concluding remarks

As pointed out in the introduction, Thurston's proof of the midscribability of the sphere can be reversed, to give another proof of:

6.1 Corollary (Circle Packing Theorem) *Let G be a planar graph. Then there exists a circle packing on the sphere having G as its tangency graph; that is, the circles of the packing are in one to one correspondence with the vertices of G and two circles touch if and only if the corresponding vertices of G share an edge.*

Proof. It is easy to see that one can extend G to a triangulation of S^2 by adding vertices and edges to G, but without adding an edge between two vertices which are already in G. Since it is sufficient to prove the claim for the extended graph, we may, and will, assume that G is a triangulation. Then there is some polyhedron, say Q, whose graph is precisely G. Theorem 1.1 implies that there is such a polyhedron Q which midscribes the unit sphere $S^2 \subset \mathbb{R}^3$. For each vertex v of Q, let C_v be the set of points on S^2 which are visible from v; that is, the set of $p \in S^2$ so that the segment $[p, v]$ intersects S^2 only in p. Then the collection of these sets C_v forms a circle packing on S^2 and two circles are tangent if and only if the corresponding vertices share an edge of Q. □

Of course, our result here can be used to generalize the circle packing theorem to packings on the boundary of convex bodies. However, this already appears in [Schr].

Remark. Perhaps the best setting for the midscription problem is not in $\mathbb{R}^3$, but in S^3 (S^3 considered as the double cover of projective 3-space, rather than as the one point compactification of $\mathbb{R}^3$). This is like allowing polyhedra obtained

from convex polyhedra by projective transformations. One advantage to working in S^3 is that duality between vertices and faces, which did not appear in our discussion but lurked in the background, is more apparent. Another advantage is that there is no need to worry about vertices escaping to infinity, and so the Lemmas 5.1 and 5.2 can be deleted from the discussion. With the appropriate adjustment of the definitions and the situation in Theorem 4.1 to S^3, one gets that if the points x_0, x_1 are not antipodal, and the plane (that is, codimension 1 oriened great sphere) y_2 contains interior points of K, then $Z'_{P,K} \neq \emptyset$. In fact, $Z'_{P,K}$ consists of one single point. In this setting, it follows that the manifold of solutions is homeomorphic to the Möbius group.

References

[An1] Andreev, E.M.: On convex polyhedra in Lobačevskiĭ spaces. Mat. Sb., Nov. Ser. **81** (123), 445–478 (1970); English translation in Math. USSR, Sb. **10**, 413–440 (1970)

[An2] Andreev, E.M.: On convex polyhedra of finite volume in Lobačevskiĭ space. Mat. Sb., Nov. Ser. **83** (125), 256–260 (1970); English translation in Math. USSR, Sb. **12**, 255–259 (1970)

[Ca] Cauchy, A.L.: Sur les polygones et polyèdres, second Mémoire. J. Éc. Polytechnique **19**, 87–98 (1813)

[Eg] Eggleston, H.G.: Convexity. Cambridge: Cambridge University Press 1958

[Gr] Grünbaum, B.: Convex Polytopes. New York: Wiley 1967

[G-S] Grünbaum, B., Shephard, G.: Some problems on polyhedra. J. Geom. **29**, 182–190 (1987)

[Koe] Koebe, P.: Kontaktprobleme der konformen Abbildung. Ber. Verh. Saechs. Akad. Wiss. Leipzig, Math.-Phys. Kl. **88**, 141–164 (1936)

[Ro] Roth, B.: Rigid and flexible frameworks. Am. Math. Mon. **88**, 6–21 (1981)

[Schr] Schramm, O.: Existence and uniqueness of packings with specified combinatorics. Isr. J. Math. (to appear)

[Schu] Schulte, E.: Analogues of Steinitz's theorem about non-inscribable polytopes. In: Böröczky, K., Tóth, G.F. (eds.) Intuitive geometry. Siófok, 1985 (Colloq. Math. Soc. János Bolyai, vol. 48, pp. 503–516). Amsterdam: North-Holland 1987

[St] Steinitz, E.: Über isoperimetrische Probleme bei konvexen Polyedern. J. Reine Angew. Math. **159**, 133–143 (1928)

[Th] Thurston, W.P.: The geometry and topology of 3-manifolds. Princeton University Notes 1982

[Tu] Tutte, W.T.: How to draw a graph. Proc. Lond. Math. Soc. **52**, 743–767 (1963)

Annals of Mathematics, **137** (1993), 369–406

Fixed points, Koebe uniformization and circle packings

By ZHENG-XU HE and ODED SCHRAMM*

Contents

Introduction
1. The space of boundary components
2. The fixed-point index
3. The Uniqueness Theorem
4. The Schwarz–Pick lemma
5. Extension to the boundary
6. Maximum modulus, normality and angles
7. Uniformization
8. Domains in Riemann surfaces
9. Uniformizations of circle packings

Addendum

Introduction

A domain in the Riemann sphere $\hat{\mathbb{C}}$ is called a *circle domain* if every connected component of its boundary is either a circle or a point. In 1908, P. Koebe [Ko1] posed the following conjecture, known as Koebe's Kreisnormierungsproblem: *Any plane domain is conformally homeomorphic to a circle domain in* $\hat{\mathbb{C}}$. When the domain is simply connected, this is the content of the Riemann mapping theorem. The conjecture was proved for finitely connected domains and certain symmetric domains by Koebe himself ([Ko2], [Ko3]); for domains with various conditions on the "limit boundary components" by R. Denneberg [De], H. Grötzsch [Gr], L. Sario [Sa], H. Meschowski

*The authors were supported by N.S.F. Grants DMS-9006954 and DMS-9112150, respectively. The authors express their thanks to Mike Freedman, Dennis Hejhal, Al Marden, Curt McMullen, Burt Rodin, Steffen Rohde and Bill Thurston for conversations relating to this work. Also thanks are due to the referee, and to Steffen Rohde, for their careful reading and subsequent corrections.

The paper of Sibner [Si3] served as a very useful introduction to the subject.

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_6, © Springer Science+Business Media, LLC 2011

([Me1], [Me2]), K.L. Strebel ([Str1], [Str2]), L. Bers [Be], A. Haas [Haa] and others; and for domains quasiconformally homeomorphic to a circle domain by R.J. Sibner [Si1], [Si2].

In this article we prove the following theorem:

THEOREM 0.1. *Any domain Ω in $\hat{\mathbb{C}}$, whose boundary $\partial\Omega$ has at most countably many components, is conformally homeomorphic to a circle domain Ω^* in $\hat{\mathbb{C}}$. Moreover Ω^* is unique up to Möbius transformations, and every conformal automorphism of Ω^* is the restriction of a Möbius transformation.*

The uniqueness of Ω^* in the above theorem can fail if Ω has uncountably many boundary components. The theory of quasiconformal maps and the Beltrami equation (cf. [LV]) can be used to show that the complement of a Cantor set in $\hat{\mathbb{C}}$ of nonzero area provides such an example. This is done by placing a nonzero Beltrami differential supported on the Cantor set and solving the Beltrami equation to obtain a quasiconformal map which is conformal outside the Cantor set.

A *circle domain* in a Riemann surface is a domain, whose complement's connected components are all closed geometric disks and points. Here a geometric disk (or, in short, a disk) means a topological disk, whose lifts in the universal cover of the Riemann surface (which is the hyperbolic plane, the euclidean plane or the sphere) are round. As a consequence of Theorem 0.1 we have the following theorem:

THEOREM 0.2. *Let Ω be an open Riemann surface with finite genus and at most countably many ends. Then there is a closed Riemann surface R^* such that Ω is conformally homeomorphic to a circle domain Ω^* in R^*. Moreover the pair (R^*, Ω^*) is unique up to conformal homeomorphisms.*

Circle domains are closely related to circle packings, and results similar to the above hold for circle packings. Recall that a *circle packing* P is a collection of closed geometric disks with disjoint interiors. The (*tangency*) *graph*, or *nerve*, of a circle packing P is a graph, whose vertices are in one-to-one correspondence with the packed sets, and an edge appears in the graph if and only if the corresponding disks are tangent. The *carrier* of a circle packing P is the union of the packed disks and the finite interstices (the connected components of the complement, whose boundaries lie on finitely many of the packed disks). The circle packing theorem says that, for any triangulation $\mathcal{T}$ of the 2-sphere, there is a circle packing in $\hat{\mathbb{C}}$, unique up to Möbius transformations, whose graph is combinatorially equivalent to the 1-skeleton of $\mathcal{T}$. This theorem was first discovered by Koebe [Ko4], who obtained it as a limiting case for his uniformization theorem of finitely connected domains as circle

domains.[1] But the circle packing theorem was unnoticed, or forgotten, until W.P. Thurston [Th1] rediscovered it as a corollary of E.M. Andreev's theorem ([An1], [An2]). Thurston [Th2] then conjectured that finite circle packings can approximate the Riemann map from a simply connected domain to the unit disk. This conjecture was proved by B. Rodin and D. Sullivan [RoSu], and much research on circle packings followed.

A slightly modified proof of Theorem 0.1 yields the following generalization of the circle packing theorem to infinite triangulations:

THEOREM 0.3. *Let $\mathcal{T}$ be a triangulation of a domain in $\hat{\mathbb{C}}$ with at most countably many boundary components. Then there is a circle packing P in $\hat{\mathbb{C}}$ whose graph is combinatorially equivalent to the 1-skeleton of $\mathcal{T}$ and whose carrier is a circle domain. Moreover P is unique up to Möbius transformations.*

This theorem was conjectured in [Sch2]. More generally we have the following theorem:

THEOREM 0.4. *Let $\mathcal{T}$ be a triangulation of a finite-genus open surface with at most countably many ends. Then there are a closed Riemann surface R and a circle packing $P \subset R$ whose graph is combinatorially equivalent to the 1-skeleton of $\mathcal{T}$ and whose carrier is a circle domain in R. Moreover R and P are unique up to conformal or anticonformal homeomorphisms.*

As a special case of Theorem 0.3 we have the following corollary:

COROLLARY 0.5. *Let $\mathcal{T}$ be a triangulation of a simply connected plane domain. Then there is a circle packing P in $\mathbb{C}$ whose graph is combinatorially equivalent to the 1-skeleton of $\mathcal{T}$ and whose carrier is either the (euclidean) plane $\mathbb{C}$ or the unit disk U. Moreover P is unique up to Möbius transformations.*

The 1-skeleton of $\mathcal{T}$ is called a *parabolic* graph if the carrier of the circle packing P of the above corollary is equal to $\mathbb{C}$; otherwise it is called a *hyperbolic* graph. In [HeSch] and [BSte2] some combinatorial criteria are given for determining if a graph is parabolic or hyperbolic.

The uniqueness (or rigidity) statement of Corollary 0.5 was previously proved by the second author in [Sch2]. In the restricted case, when there is a uniform upper bound on the valences of all the vertices in the triangulation $\mathcal{T}$, the existence statement of Corollary 0.5 follows from [BSte1], and the uniqueness can be obtained by the methods of [RoSu] or [He2] (see also [Ro1]).

[1]The authors thank Horst Sachs for this reference.

This work originated with the proof of Theorem 0.3, and then we noticed that the techniques apply to circle domains as well. However the proof of Theorem 0.1 does not mention circle packings.

Theorem 0.1 has two parts: uniqueness and existence. The uniqueness is derived from an analysis of fixed-point indices of mappings. Given a domain $\Omega \subset \mathbb{C}$ and a mapping $f : \partial\Omega \to \mathbb{C}$, the fixed-point index of f can be defined as the total number of fixed points of a continuous map $F : \Omega \to \mathbb{C}$ whose restriction to $\partial\Omega$ is f, counting multiplicities. (One should restrict f to having no fixed points, and F to having only isolated fixed points.) This number does not depend on the choice of F. Conformal maps have only positive-multiplicity fixed points, and so, if the index of f is negative, then one can conclude that there is no continuous F whose restriction to $\partial\Omega$ is f and whose restriction to Ω is conformal. The first fundamental observation in the proof of uniqueness is that the index of any orientation-preserving homeomorphism that takes a circle to a circle is nonnegative; thus much information is available for domains that have circles as boundary components. As was known already to Strebel [Str1], this observation implies that a conformal homeomorphism between circle domains with countably many boundary components that extends continuously to the boundary is a Möbius transformation. However Strebel was not able to prove the continuous extension to the boundary, as is done below.[2]

One of the main tools we use both in the uniqueness and existence parts of Theorem 0.1 is the following Schwarz–Pick lemma for multiply connected domains:

THEOREM 0.6 (Schwarz–Pick lemma for multiply connected domains). *Let $U \subset \mathbb{C}$ denote the open unit disk and let A and A^* be Jordan domains in $\mathbb{C}$ with $A \supset U \supset A^*$. Let Ω be a domain, which is obtained from A by the deletion of a closed disjoint union of at most countably many closed (geometric) disks and points in A. Similarly let Ω^* be a domain obtained from A^* by the deletion of a closed disjoint union of at most countably many closed disks and points. Suppose that $f : \Omega \to \Omega^*$ is a conformal homeomorphism between Ω and Ω^* and that $f^B(\partial A) = \partial A^*$. Then f is a contraction in the hyperbolic metric in the following sense: If $p, q \in \Omega \cap U$ are distinct, then*

$$d_{\mathrm{hyp}}\big(f(p), f(q)\big) \leqslant d_{\mathrm{hyp}}(p, q),$$

where $d_{\mathrm{hyp}}(\cdot,\cdot)$ denotes the distance in the hyperbolic metric of U. Furthermore, if equality holds for one pair $p \neq q$, then $\Omega \subset U$ and f is a restriction to Ω of a hyperbolic isometry.

[2]The authors thank Dennis Hejhal for pointing out this work of Strebel.

This is already an interesting result in the finite-connectivity case; that is, when the collection of deleted disks and points is finite. Our proof of this theorem is also based on fixed-point arguments. A version of the Schwarz–Pick lemma for circle packings was previously obtained by Rodin ([Ro1], [Ro2]) and by A. Beardon and K. Stephenson ([BSte1], [BSte3]).

Both the proof of the continuous extension to the boundary and the proof of existence use transfinite induction, where the induction is with respect to the "complexity" of the boundary, as in Sibner's paper [Si3]. We prove existence by taking limits of maps from simpler subdomains of Ω. A maximum modulus principle and normality results, as well as the Schwarz–Pick lemma, are needed to conclude that the limit has the required properties.

1. The space of boundary components

In this section we recall the definition and properties of the space of boundary components of an open planar set.

We will use the term *domain* for a connected open set in the Riemann sphere $\hat{\mathbb{C}}$. Let Ω be a domain in $\hat{\mathbb{C}}$. The collection $B(\Omega)$ of boundary components of Ω has the structure of a compact Hausdorff space. One way to describe the topology on $B(\Omega)$ is the following: On the boundary of Ω, $\partial\Omega$, consider the equivalence relation $\sim$ in which $z \sim w$ if and only if z and w belong to the same connected component of $\partial\Omega$. Then $B(\Omega)$ is the set of equivalence classes of $\sim$, $B(\Omega) = \partial\Omega/\sim$, with the quotient topology.

An intrinsic way to describe $B(\Omega)$ is as the space of ends of Ω, denoted by $\mathcal{E}(\Omega)$. An end $e \in \mathcal{E}(\Omega)$ is a function that assigns to each compact subset $F \subset \Omega$ a connected component $e(F)$ of $\Omega - F$ in such a way that $e(F) \supset e(F^*)$ whenever $F \subset F^*$. The topology on the collection of ends is then defined as the minimal topology containing all sets of the form $\{e \in \mathcal{E}(\Omega) : e(F) = C\}$, where F is some compact subset of Ω and C is some connected component of $\Omega - F$. It is then not hard to check that $B(\Omega)$ is naturally homeomorphic to the space of ends. (The end e_K corresponding to a boundary component $K \in B(\Omega)$ is the end for which $e_K(F)$ is the connected component of $\Omega - F$ whose closure intersects K, whenever $F \subset \Omega$ is compact.) Since the space of ends is clearly independent of the embedding of Ω in $\hat{\mathbb{C}}$, we have the following well-known fact:

FACT 1.1. *Let $f : \Omega \to \Omega^*$ be a homeomorphism of connected open sets in $\hat{\mathbb{C}}$. Then f induces a canonical homeomorphism $f^B : B(\Omega) \to B(\Omega^*)$.*

In the following, we shall use the notation f^B for that homeomorphism induced by f.

This article deals with domains Ω, where $B(\Omega)$ is at most countable (and nonempty), and henceforth we make this assumption on $B(\Omega)$. From Baire category considerations it follows that any countable compact Hausdorff space has isolated points. Given a topological space $\mathcal{X}$, let $\mathcal{X}'$ be $\mathcal{X} - \{\text{its isolated points}\}$. Then $\mathcal{X}'$ is closed in $\mathcal{X}$. Now, for each ordinal α, we define $\mathcal{X}^\alpha$ by transfinite induction (see [Hau] for background on ordinals and transfinite induction): let $\mathcal{X}^0 = \mathcal{X}$; for successor ordinals $\alpha = \beta + 1$ let $\mathcal{X}^\alpha = (\mathcal{X}^\beta)'$; and for limit ordinals α define $\mathcal{X}^\alpha = \bigcap_{\beta<\alpha} \mathcal{X}^\beta$. Then, for each ordinal α, $\mathcal{X}^\alpha$ is closed in $\mathcal{X}$.

Since $\mathcal{X}' \subset \mathcal{X}$, we have $\mathcal{X}' \neq \mathcal{X}$ for every closed nonempty subset $\mathcal{X}$ of $B(\Omega)$. There is some ordinal β so that $B(\Omega)^\beta = \emptyset$. Let γ be the minimal ordinal with that property. By compactness, γ must be a successor ordinal and γ is, of course, countable. Let α be its predecessor. We will refer to α as the *rank* of Ω. The rank is also defined as the only ordinal α such that $B(\Omega)^\alpha$ is finite and nonempty. If $B(\Omega)^\alpha$ contains n points, $0 < n < \infty$, then the pair (α, n) will be called the *type* of Ω. For every ordinal β, an isolated point of $B(\Omega)^\beta$ will be called a boundary component of *rank* β.

2. The fixed-point index

Definition. Let γ be an oriented Jordan curve in the plane $\mathbb{C}$. Let $f : \gamma \to \mathbb{C}$ be a continuous map without fixed points. The (fixed-point) *index* of f, denoted by $\text{index}(f)$, is defined to be the winding number with respect to the origin 0 of the closed curve $f(z) - z$ as z varies in γ. In other words, $\text{index}(f)$ is the winding number of $g \circ \gamma$ around 0, where $g(z) = f(z) - z$ and γ is parametrized in accordance with its orientation.

If the domain of definition of f is a finite, disjoint union of oriented Jordan curves in $\mathbb{C}$ (and f has no fixed points), then the index of f is defined as the sum of the indices of the restrictions of f to the individual curves.

Definition. Let $f : A \to \mathbb{C}$ be continuous, where $A \subset \mathbb{C}$, and suppose that $z \in \text{int}(A)$ (the interior of A) is an isolated fixed point of f. The *index* of f at z, denoted by $\text{index}(f, z)$, is defined as the index of the restriction of f to ∂D, where $D \subset A$ is a closed disk that contains z in its interior, but does not contain any other fixed point of f, and where ∂D is positively oriented with respect to D. Considering homotopies makes it clear that $\text{index}(f, z)$ does not depend on the particular choice of D. The index of f at z will also be called the *multiplicity* of the fixed point z.

THEOREM 2.1 (Poincaré–Hopf). *Let $A \subset \mathbb{C}$ be a compact set that is the closure of its interior and whose boundary consists of finitely many disjoint*

Jordan curves. Let the boundary components of A be positively oriented with respect to A. Suppose that $f : A \to \mathbb{C}$ is continuous, has only isolated fixed points and has no fixed points on the boundary of A. Then the index of the restriction of f to ∂A is equal to the sum of the indices of f at all its fixed points.

Remarks. A simple and elementary proof of Theorem 2.1 is given below. However, for the sake of the reader whose topology is a little rusty, we will now try to put Theorem 2.1 in perspective. The common formulation of the Poincaré–Hopf theorem (cf. [Mi], [GPo]) is that on a compact smooth manifold without boundary the number of zeroes, counting multiplicities, of any continuous vector field with only isolated zeroes is equal to the Euler characteristic of the manifold. This relates to our setting in the following way: A function $f : A \to \mathbb{C}$ gives a vector field $v(z) = f(z) - z$, and the fixed points of the function f correspond to the zeroes of the vector field v.

The boundary enters the picture in the usual manner: given a compact smooth manifold M with boundary ∂M and a zeroless vector field v on ∂M (with values taken in the tangent space of M, denoted by TM), the number of zeroes of a continuous extension of v to all of M with only isolated zeroes is independent of the extension and is therefore an invariant of the pair (M, v). That is an easy consequence of the Poincaré–Hopf theorem, as follows: Let v_1, v_2 be two such extensions and let M' be a manifold identical to M, but with the reversed orientation. Glue M and M' along the boundary (via the identification of ∂M and $\partial M'$) to form a manifold M^* without boundary. Now v is a vector field on ∂M with values in TM^*. Extend it to some continuous vector field v' on $M' \subset M^*$ with only isolated zeroes. Now consider two vector fields on M^*: w_1, which is v_1 on M and v' on M', and w_2, which is v_2 on M and v' on M'. The Poincaré–Hopf theorem implies that the number of zeroes of w_1 and w_2 is the same, and that says that the number of zeroes of v_1 and v_2 is the same.

The Lefschetz fixed-point theorem is also very much related. Its relevant consequence here is that an orientation-preserving homeomorphism of the sphere $\hat{\mathbb{C}}$ with only isolated fixed points has, counting multiplicities, 2 fixed points. (The number 2 appears as the Euler characteristic of the sphere.) This fact is in the background of this article. For example, consider a conformal homeomorphism of $\hat{\mathbb{C}}$. By composing with a Möbius transformation, one can arrange to get at least 3 fixed points and, as we shall see below, it is impossible in this case to get fixed points with negative index. Therefore, by the Lefschetz theorem, there must be nonisolated fixed points. This says that the composition must be the identity map, and the original conformal homeomorphism must be a Möbius transformation. This simple proof that

any conformal homeomorphism of $\hat{\mathbb{C}}$ is a Möbius transformation can be seen as the prototype for the proofs of our rigidity results below.

Proof of Theorem 2.1. Suppose first that f has no fixed points. Define $F : A \to \mathbb{C} - \{0\}$ by $F(z) = f(z) - z$. The boundary of A, with the orientation induced by A, is obviously trivial in the first homology of A. Therefore its image under F in the first homology of $\mathbb{C} - \{0\}$ is also 0. This means that the index of the restriction of f to the boundary of A is 0. And so the theorem holds in this case.

Note that the assumptions imply that f will have at most finitely many fixed points. If f has fixed points $z_1, \ldots, z_k$, then let $D_1, \ldots, D_k$ be small disjoint closed disks in $\operatorname{int}(A)$ containing $z_1, \ldots, z_k$ in their respective interiors. Let A^* be the closure of $A - \bigcup_{j=1}^{k} D_j$. In A^* the map f has no fixed points, and so the index of the restriction of f to ∂A^* is 0. Each ∂D_j has the opposite orientations as a boundary component of A^*. Therefore the index of the restriction of f to ∂D_j is negated if one takes the orientation induced by A^*. We conclude that the index of the restriction of f to ∂A is equal to the index of the restriction of f to $\bigcup_{j=1}^{k} \partial D_j$. The theorem follows. □

A variation of the following lemma appears in [Str1].

CIRCLE INDEX LEMMA 2.2. *Let J, K be Jordan curves in $\mathbb{C}$, positively oriented with respect to the Jordan domains that they bound (in $\mathbb{C}$); and let $f : J \to K$ be an orientation-preserving homeomorphism with no fixed points. Then*

(1) $\operatorname{index}(f) = \operatorname{index}(f^{-1})$.

(2) *If J is contained in the closure of the Jordan domain determined by K, or K is contained in the closure of the Jordan domain determined by J, then* $\operatorname{index}(f) = 1$.

(3) *If the intersection of K and J contains at most 2 points, then* $\operatorname{index}(f) \geqslant 0$.

(4) *If J and K are circles, then* $\operatorname{index}(f) \geqslant 0$.

Proof. Let $g : K \to \mathbb{C}$ be defined by $g(z) = f^{-1}(z) - z$. Then the winding number around the origin of $g \circ f : J \to \mathbb{C}$ is clearly the same as that of g. But $g \circ f(z) = z - f(z)$, which has the same winding number around 0 as $z \to f(z) - z$. Therefore part (1) holds.

Let $\hat{J}$ and $\hat{K}$ denote the closures of the Jordan domains determined by J and K, respectively. To prove part (2) suppose that $J \subset \hat{K}$. Let $h : J \times [0,1] \to \hat{K}$ be a homotopy from the identity map of J to some constant $\mathrm{c} \in \operatorname{int}(\hat{K})$ with the property that $h(z,t) \notin K$ for $t > 0$. Define $H(z,t) = f(z) - h(z,t)$ for $z \in J, t \in [0,1]$. Then 0 is not in the image of H, and H is a homotopy from $z \to f(z) - z$ to $z \to f(z) - \mathrm{c}$. Since $\mathrm{c} \in \operatorname{int}(\hat{K})$, the map $z \to f(z) - \mathrm{c}$

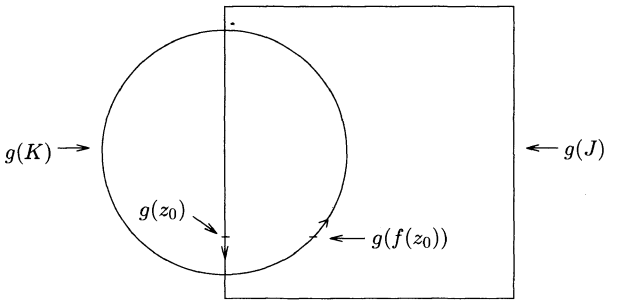


FIGURE 2.1. After the isotopy.

has the winding number 1 around 0. Therefore the map $z \to f(z) - z$ has the winding number 1 around 0, and index$(f) = 1$. Using (1), we see that (2) holds.

Consider now the case where the interiors of $\hat{K}$ and $\hat{J}$ are disjoint. Let $h : J \times [0,1] \to \mathbb{C} - \text{int}(\hat{K})$ be a homotopy from the identity map on J to some constant c $\notin \hat{K}$ with the property that $h(z,t) \notin \hat{K}$ for $t > 0$. Define $H(z,t) = f(z) - h(z,t)$ for $z \in J, t \in [0,1]$, as in the previous paragraph. Then H is a homotopy from $z \to f(z) - z$ to $z \to f(z) - \text{c}$ in $\mathbb{C} - \{0\}$. The winding number of $z \to f(z) - \text{c}$ around 0 is zero, because c $\notin \hat{K}$. This shows that the winding number of $z \to f(z) - z$ around 0 is also zero. Therefore index$(f) = 0$, in the case where the interiors of $\hat{K}$ and $\hat{J}$ are disjoint.

To prove part (3), it remains to consider the case where J and K intersect in exactly 2 points and neither $J \subset \hat{K}$ nor $K \subset \hat{J}$. In that case, let $h : (J \cup K) \times [0,1] \to \mathbb{C}$ be an isotopy from the identity map on $J \cup K$ to a map $g : J \cup K \to \mathbb{C}$ with the property that $g(J)$ is the square and $g(K)$ is the circle indicated in Figure 2.1. Let $H(z,t) = h(f(z),t) - h(z,t)$ for $z \in J, t \in [0,1]$. Then H is a homotopy in $\mathbb{C} - \{0\}$ from $z \to f(z) - z$ to $z \to g(f(z)) - g(z)$; therefore index(f) is the winding number of $z \to g(f(z)) - g(z)$ around 0. If z_0 is some point where $g(f(z)) - g(z)$ is real and positive, then $g(z_0)$ must be in the line segment joining the 2 intersection points of $g(J)$ and $g(K)$; obviously $g(f(z_0))$ is to the right of this line segment. Therefore, as z moves near z_0 in the positive direction along J, the imaginary part of $g(z)$ decreases and the imaginary part of $g(f(z))$ increases. This implies that the imaginary part of $g(f(z)) - g(z)$ is increasing near every point $z_0 \in J$ where the curve $z \to g(f(z)) - g(z)$ crosses the positive real ray. Thus the winding number around 0 of this curve is nonnegative, and the proof of (3) is complete.

Now part (4) follows from (2) and (3). □

We now recall another well-known fact.

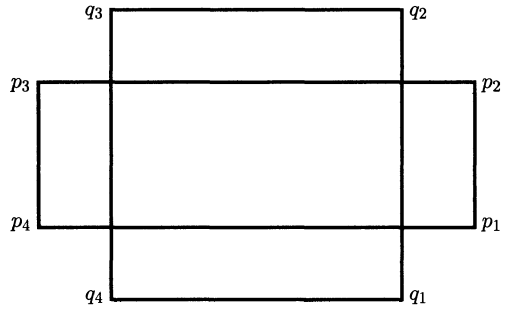


FIGURE 2.2. These rectangles are conformally unequivalent.

FACT 2.3. *Let $\Omega \subset \mathbb{C}$ be open and connected and let $f : \Omega \to \mathbb{C}$ be an analytic map that is not the identity. Then f has only isolated fixed points in Ω; and if z is a fixed point of f, then* $\operatorname{index}(f, z) \geqslant 1$.

Proof. Since f is not the identity, $f(z) - z$ is not identically 0 and therefore has only isolated zeroes in Ω. So the fixed points of f are isolated. The multiplicity of a 0 of $f(z) - z$ is the fixed-point index of f at that point. Thus $\operatorname{index}(f, z) \geqslant 1$ at each fixed point z. □

Example. Consider the two rectangles in Figure 2.2. Let f be a homeomorphism between them that maps p_1, p_2, p_3, p_4 to q_1, q_2, q_3, q_4, respectively. Then it is easy to see that the fixed-point index of f is -1. One concludes that there is no conformal homeomorphism between these rectangles whose continuous extension to the boundary takes each p_i to q_i; that is, the quadrilaterals have distinct conformal moduli. This result is hardly surprising, but it is meant to exhibit a simple application of the fixed-point index, which is very much in the spirit of our arguments below.

The next corollary, which is more or less immediate from the above observations, will be most useful.

COROLLARY 2.4. *Let $f : \Omega \to \Omega^*$ be a conformal homeomorphism between bounded plane domains that extends continuously to a homeomorphism $F : \overline{\Omega} \to \overline{\Omega}^*$. Let $B_0 \subset B(\Omega)$ be a finite collection of boundary components of Ω. Suppose that the following conditions hold:*

(1) *Ω has at most countably many boundary components;*

(2) *all of the boundary components in $B(\Omega) - B_0$ and $B(\Omega^*) - f^B(B_0)$ are circles and points, and all of the boundary components in B_0 are Jordan curves;*

(3) *B_0 contains the boundary component of Ω that is contained in the unbounded component of $\mathbb{C} - \Omega$; and similarly, $f^B(B_0)$ contains the boundary component of Ω^* that is contained in the unbounded component of $\mathbb{C} - \Omega^*$;*

(4) *F has no fixed points in any of the boundary components in B_0.*
Let n be the index of the restriction of F to the boundary components in B_0. Then f has at most n fixed points in Ω. Furthermore, if S is a set of fixed points of f, then the total number of fixed points for f in S, counting multiplicity, is at most n.

Note that conditions (4) and (3) imply that f is not the identity and therefore, being conformal, has only isolated fixed points in Ω. One conclusion of the corollary is that $n \geqslant 0$.

Proof. Let S be a finite set of fixed points of f and let m be the total number of fixed points in S, counting multiplicity. Because f is conformal, all of the fixed points of f have positive multiplicity. Therefore it is sufficient to show that $m \leqslant n$.

Our first goal is to make a perturbation to the case that F has no fixed points in $\partial\Omega$. For any constant $c \in \mathbb{C}$ with $|c|$ sufficiently small, the map $z \to F(z) + c$ will have at least m fixed points, counting multiplicities, in some neighborhood of S. (This follows from Rouché's theorem, and it is also clear from the topological definition.)

Let $J \in B(\Omega) - B_0$ be a boundary component of Ω. Since J and $f^B(J)$ are circles or points, the set of complex numbers c such that the mapping $z \to F(z) + c$ has a fixed point in J is a closed set with empty interior in $\mathbb{C}$. Therefore, since the collection of boundary components is countable, by Baire category considerations there are complex numbers c arbitrarily close to 0 so that $z \to F(z) + c$ has no fixed points on the boundary components in $B(\Omega) - B_0$. Since F has no fixed points on the boundary components in B_0, we can find a $c \in \mathbb{C}$ arbitrarily close to 0 such that $z \to F(z) + c$ will have no fixed points in $\partial\Omega$. Pick such a c with $|c|$ sufficiently small that

(a) $z \to f(z) + c$ has at least m fixed points in Ω, counting multiplicity,

(b) the index of the restriction of the map $z \to F(z) + c$ to the boundary components in B_0 is still n. Define $F_c(z) = F(z) + c$.

If $\{z_0\}$ is a boundary component of Ω consisting of a single point, then $F_c(z_0) \neq z_0$, since F_c has no fixed points in $\partial\Omega$.

When J is a circle in $\mathbb{C}$, we will denote by $D(J)$ the closed disk in $\mathbb{C}$ determined by J. Let B_+ be the set of boundary components J in $B(\Omega) - B_0$, which are circles and for which $D(J) \cap D(F_c(J)) \neq \emptyset$. Let B_- consist of all the other circle boundary components in $B(\Omega) - B_0$.

The set B_+ is necessarily finite. To see this, consider an infinite sequence of distinct circles in B_+. The radii of these circles must tend to 0. By taking a subsequence, if necessary, we may assume that this sequence of circles converges to some point. This point would necessarily be a fixed point for F_c on $\partial\Omega$, which gives a contradiction and shows that B_+ is finite.

Let $\Sigma = \overline{\Omega} \cup (\bigcup_{J \in B_-} D(J))$. We now continuously extend the map F_c to a map $G : \Sigma \to \mathbb{C}$ by letting the restriction of G to every $D(J)$, $J \in B_-$, be an arbitrary homeomorphism onto $D(F_c(J))$, which agrees with F_c on J. By the definition of B_-, the fixed points of G will be exactly those of F_c.

Since B_+ is finite, Σ is a closed set, whose boundary consists of finitely many Jordan curves $(B_0 \cup B_+)$. What is the index of the restriction of G to $\partial\Sigma$? The index of the restriction of G to the boundary components in B_0 is n. On the other hand, the Circle Index Lemma 2.2 tells us that the index of the restriction of G to every $J \in B_+$ is nonnegative, since J and $G(J)$ are circles, provided that we consider J with the orientation induced by $D(J)$ and that G is orientation preserving. But for $J \in B_+$ the disks $D(J)$ and $D(G(J))$ are disjoint from Σ and $G(\Sigma)$, respectively; hence G is orientation preserving, and the orientation we must consider for J is the opposite orientation from the one induced by $D(J)$. Therefore the index of the restriction of G to every $J \in B_+$ is nonpositive. Thus the index of the restriction of G to $\partial\Sigma$ is at most n.

We now appeal to Theorem 2.1, and conclude that the total number of fixed points of G, counting multiplicities, is at most n. Certainly the same would be true for F_c. This gives $n \geqslant m$, since every fixed point of G has positive multiplicity, and so the proof of the corollary is complete. □

3. The Uniqueness Theorem

Recall that a circle domain in $\hat{\mathbb{C}}$ is a connected open subset of $\hat{\mathbb{C}}$, all of whose boundary components are circles and points. We now restate the uniqueness part of Theorem 0.1.

THE UNIQUENESS THEOREM 3.1. *Let Ω, Ω^* be circle domains in the Riemann sphere having at most countably many boundary components. Suppose that $f : \Omega \to \Omega^*$ is a conformal homeomorphism of Ω onto Ω^*. Then f is a restriction to Ω of a Möbius transformation and, in particular, Ω and Ω^* are Möbius equivalent.*

BOUNDARY EXTENSION THEOREM 3.2. *Let Ω, Ω^* be open connected sets in the Riemann sphere and let $f : \Omega \to \Omega^*$ be a conformal homeomorphism between them. Let W be an open subset of $B(\Omega)$, which is at most countable. Suppose that the boundary components of Ω corresponding to elements of W are all circles and points and that the corresponding (under f) boundary components of Ω^* are also all circles and points. Then f extends continuously to the boundary components in W and extends to a homeomorphism between $\bigcup\{K : K \in W\} \cup \Omega$ and $\bigcup\{K^* : K^* \in f^B(W)\} \cup \Omega^*$.*

In the hypotheses of the theorem, we do not assume that $f^B(K)$ is a circle when $K \in W$ is a circle or that $f^B(K)$ is a point when $K \in W$ is a point.

The proof of Theorem 3.2 will appear in the following sections. Now we will see how the Uniqueness Theorem 3.1 follows from our Boundary Extension Theorem 3.2. This was also pointed out by Strebel [Str1] (who was unable to prove Theorem 3.2).

Proof of the Uniqueness Theorem 3.1. Assuming Theorem 3.2, we apply it with $W = B(\Omega)$ to conclude that f extends continuously to a homeomorphism F from $\overline{\Omega}$ to $\overline{\Omega}^*$ (closures in $\hat{\mathbb{C}}$).

Since we are free to normalize in the domain and range by Möbius transformations, we assume, without loss of generality, that $\infty \in \Omega$ is a fixed point for f and that f has the form

$$f(z) = z + \frac{a_1}{z} + \frac{a_2}{z^2} + \cdots$$

near ∞. If all of the coefficients a_j are 0, then $f(z) = z$, and we are done. Suppose that j is the least positive integer with $a_j \neq 0$. Then there is a real number R large enough that $\partial D(0, R) \subset \Omega$ and $|f(z) - z - a_j z^{-j}| < |a_j z^{-j}|$ for $z \in \partial D(0, R)$, where $D(0, R)$ is the disk centered at 0 with radius R. The index of the restriction of f to $\partial D(0, R)$ is the same as the winding number around 0 of the restriction of $z \to a_j z^{-j}$ to $\partial D(0, R)$, because the homotopy $H : \partial D(0, R) \times [0, 1] \to \mathbb{C}$, $H(z, t) = (1 - t)(f(z) - z) + t a_j z^{-j}$, has no zeroes. But that winding number is $-j$, which is negative. If we now look at the restriction of f to $\Omega \cap D(0, R)$, this gives a contradiction to Corollary 2.4 with $B_0 = \{\partial D(0, R)\}$. This contradiction shows that all of the coefficients a_j are 0, completing the proof. □

4. The Schwarz–Pick lemma

The generalization Theorem 0.6 of the Schwarz–Pick lemma is of central importance in this work;[3] we now start its proof.

Note that the assumptions in Theorem 0.6 imply that Ω is a domain that has at most countably many boundary components, and all but one of them are circles and points. The same holds for Ω^*.

The following lemma, which will also be used in the proof of Theorem 3.2, shows that Theorem 0.6 reduces to Theorem 3.2.

LEMMA 4.1. *Theorem* 0.6 *is true under the additional hypothesis that* f *extends to a homeomorphism of* $\overline{\Omega} - \partial A$ *onto* $\overline{\Omega}^* - \partial A^*$.

Proof. Let $p \neq q$ be two points in Ω and assume first that $f(p) = p$ and $f(q) = q$. Suppose that f is not the identity. Let $h(z) = az$, where

[3] A generalization of this Schwarz–Pick lemma for noninjective mappings will appear in a subsequent paper.

$|a| < 1$, but a is sufficiently close to 1 that the composition $h \circ f$ still has 2 distinct fixed points, p' near p and q' near q. Now $h(\overline{A^*}) \subset U$. Therefore there is a Jordan curve $\gamma \subset \Omega$, which separates $h(\Omega^*)$ from ∂A. Since $h \circ f$ maps γ into $h(\Omega^*)$, the index of the restriction of $h \circ f$ to γ is 1, by part (2) of Lemma 2.2. Let Σ be the intersection of Ω with the Jordan domain bounded by γ. The points p' and q' must certainly be in Σ, because they are in $\Omega \cap h(\Omega^*)$. We therefore get a contradiction to Corollary 2.4 by considering the restriction of $h \circ f$ to Σ and taking $B_0 = \{\gamma\}$. This contradiction shows that f must be the identity if it fixes 2 distinct points.

We now deal with the general case. Suppose that $p, q \in \Omega \cap U$, $p \neq q$, and $d_{\text{hyp}}(f(p), f(q)) \geqslant d_{\text{hyp}}(p, q)$. Then there is a Möbius transformation g, which takes $f(p)$ to p, takes $f(q)$ to q and takes U into U. (If g_1, and g_2 are hyperbolic isometries of U taking p and $f(p)$, respectively, to 0, then one can take $g(z) = g_1^{-1}(g_2(z)g_1(q)/g_2(f(q)))$.) Then, by the above, the composition $g \circ f$ is the identity on Ω^*. Now g, being the inverse of f, maps $A^* \subset U$ onto $A \supset U$. Hence g is a Möbius transformation, which maps U onto U, and it is therefore a hyperbolic isometry. This implies that f is the restriction of a hyperbolic isometry and completes the proof of the lemma. □

5. Extension to the boundary

In this section we will prove Theorem 3.2 using transfinite induction. In the inductive step we will encounter the situation where a conformal map f extends continuously to all but perhaps one of the boundary components of its domain, and we will have to prove that it also continuously extends to that one boundary component. The following lemmas, like Lemma 4.1, deal with the situation where there is only one boundary component to which we do not know if f extends.

LEMMA 5.1. *Let $f : \Omega \to \Omega^*$ be a conformal homeomorphism between connected open subsets of the unit disk U, which extends continuously to a homeomorphism from $\overline{\Omega} - \partial U$ onto $\overline{\Omega^*} - \partial U$. Suppose that ∂U is a boundary component of both Ω and Ω^* and that $f^B(\partial U) = \partial U$. Further suppose that J and $J^* = f^B(J)$ are corresponding boundary components of Ω and Ω^*, all other boundary components of Ω and Ω^* are circles and points and there are at most countably many such boundary components. Consider U with the hyperbolic metric. Then, given any $\epsilon > 0$, f is bi-Lipschitz (in the hyperbolic metric) on the set L_ϵ of points in Ω having hyperbolic distance $> \epsilon$ from J. Furthermore the restriction of f to L_ϵ extends to a bi-Lipschitz homeomorphism from U onto U.*

Proof. We will first verify a Lipschitz condition near 0. Assume that J does not separate 0 from Ω. Let $a > 0$ be some number less than the *euclidean* distance from 0 to J and set $g(z) = z/a$. Then $g(J)$ does not intersect $\overline{U}$. Let $\gamma \subset g(\Omega)$ be some Jordan curve separating $g(J)$ from U and let Σ be the connected component of $g(\Omega) - \gamma$ containing 0. The restriction of the map $f \circ g^{-1}$ to Σ then satisfies the conditions of Lemma 4.1. We conclude from that lemma that $f \circ g^{-1}$ is hyperbolic-length-decreasing on $\Sigma \cap U$. But the restriction of g to the disk $D(0, a/2)$ of euclidean radius $a/2$ around 0 is Lipschitz in the hyperbolic metric, with the Lipschitz constant $l = l(a)$ depending only on a. Since $f \circ g^{-1}$ is contracting, it follows that f is Lipschitz with constant $l(a)$ on $D(0, a/2) \cap \Omega$.

For every number $t \in (0, 1)$ let $h(t)$ denote the hyperbolic radius of a circle of euclidean radius t around 0. Since we may precompose f by any hyperbolic isometry, our above Lipschitz condition near 0 translates to every point in $U - J$ as follows: Let p be a point in U whose hyperbolic distance from J is greater than $h(a)$; then f has Lipschitz constant $l(a)$ on $\Omega \cap D_{\text{hyp}}(p, h(a/2))$, where $D_{\text{hyp}}(p, r)$ denotes the disk of hyperbolic radius r centered at p. By taking limits, we find that the continuous extension of f to $\overline{\Omega} - \partial U$, which we continue to denote by f, is Lipschitz with constant $l(a)$ on $\overline{\Omega} \cap D_{\text{hyp}}(p, h(a/2))$.

Let $M \notin \{J, \partial U\}$ be some circle boundary component of Ω, which bounds a disk $D(M)$, and let M^* be the corresponding boundary component of Ω^*. As we have seen, for every point p in M, f satisfies a local Lipschitz condition at p. (The local Lipschitz constant of f at p is, by definition,

$$\limsup d_{\text{hyp}}(f(p_1), f(p_2)) / d_{\text{hyp}}(p_1, p_2)$$

as p_1 and p_2 tend to p, while $p_1 \neq p_2$.) But one can further extend f to $D(M)$ by mapping the hyperbolic center of $D(M)$ to the center of the corresponding disk $D(M^*)$ and extending radially. Clearly the local Lipschitz constant of this extension at any point in $D(M)$ is at most the supremum of the local Lipschitz constants of f in M. We extend f in this way over the interior of every such boundary circle M to obtain a map F.

Now take any $\epsilon' > 0$ and let $R_{\epsilon'}$ be the set of points in U having distance of at least ϵ' from J and which are not separated from ∂U by J. From the above, it follows that F satisfies a uniform local Lipschitz condition on $R_{\epsilon'}$; that is, there is some constant l such that all of the local Lipschitz constants of F in $R_{\epsilon'}$ are bounded by l. Since the same arguments can be applied to the inverses of f and F, and since $F(R_{\epsilon'})$ is bounded away from J^*, we conclude that F^{-1} also satisfies a uniform local Lipschitz condition on $F(R_{\epsilon'})$. It is easy to see that one can modify F in the complement of R_ϵ and then extend F to the whole of U so that the resulting homeomorphism, which we will continue to denote by F, as well as its inverse will satisfy a uniform local

Lipschitz condition on U. (To get an explicit construction take an analytic Jordan curve $\beta \subset \Omega$, which separates J from R_ϵ. Let D_1 be the topological disk bounded by β, let D_2 be the topological disk bounded by $F(\beta)$ and let g_1 and g_2 be Riemann maps from a euclidean geometric disk D to D_1 and D_2, respectively. Note that g_1 and g_2 extend analytically to ∂D. Consider the map $g : \partial D \to \partial D$ defined by $g(z) = g_2^{-1}(F(g_1(z)))$. This map g is analytic and therefore bi-Lipschitz in ∂D. Let G be the radial extension of g to D. Then G is bi-Lipschitz. Let F stay as it is outside D_1 and, in D_1, redefine F by $F(z) = g_2(G(g_1^{-1}(z)))$. In other words, with g_1 and g_2 parametrize $\overline{D}_1$ and $\overline{D}_2$ as euclidean disks and then use the usual euclidean radial extension. Because g_1, g_2 and G are bi-Lipschitz, so is the restriction of F to $\overline{D}_1$.)

Since the domain of F is now U, which is hyperbolically convex, clearly F is bi-Lipschitz (globally). (If $p, q \in U$, take the hyperbolic line segment joining p and q; the image of that segment will be a path of length at most $d_{\text{hyp}}(p, q)l$ joining $F(p)$ and $F(q)$, where l is the uniform local Lipschitz constant of F. This gives $d_{\text{hyp}}(F(p), F(q)) \leqslant l\ d_{\text{hyp}}(p, q)$. A similar argument with F^{-1} gives $l'\ d_{\text{hyp}}(F(p), F(q)) \geqslant d_{\text{hyp}}(p, q)$.) Since $L_\epsilon \subset R_\epsilon$, this completes the proof of Lemma 5.1. □

Remark. Though we will not use this, the above argument can be used to show that the Lipschitz constant of f on the set L_d tends to 1 as $d \to \infty$.

LEMMA 5.2. *Let $f : \Omega \to \Omega^*$ be a conformal homeomorphism between open connected subsets of $\hat{\mathbb{C}}$. Let $J \neq K$ be boundary components of Ω and let J^* and K^* be the corresponding boundary components of Ω^*. Suppose that Ω has at most countably many boundary components and all of them, with the possible exception of J, are circles and points. Similarly assume that all of the boundary components of Ω^*, with the possible exception of J^*, are circles and points. Also assume that f extends continuously to a homeomorphism of $\overline{\Omega} - K$ onto $\overline{\Omega}^* - K^*$. Then K is a circle if and only if K^* is a circle.*

Proof. We assume that K is a circle and K^* is a point. Since the situation is symmetric, it is enough to reach a contradiction in this case. Since K^* is a point, f extends continuously to K and maps K to K^*. By replacing J with some Jordan curve in Ω separating J from K, we assume, without loss of generality, that J and J^* are Jordan curves.

Normalizing with Möbius transformations allows us to assume that the situation is as in Figure 5.1; that is, $K^* = \{0\}$; K is a circle with center 0 which separates 0 from J; J separates K from J^*; and J^* separates J from ∞. By further renormalization we want to replace f with a map which fixes some point in Ω, without losing the above properties. To achieve that let $a = \max\{|z| : z \in J\}$, let r be the (euclidean) radius of K and let d

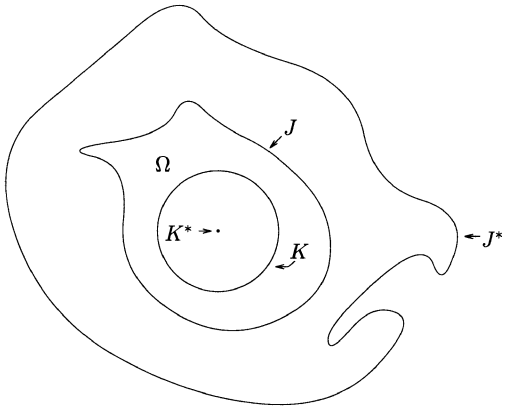


FIGURE 5.1

be the (euclidean) distance from 0 to J^*. Pick some point $p \in \Omega$, so that $|f(p)| < rd/a$, and consider the map $g(z) = f(z)p/f(p)$. We have $g(p) = p$. Since $|p| > r$, it follows that $|p/f(p)| > a/d$ and, therefore, $g^B(J)$ is disjoint from J and separates J from ∞. Let $F : \overline{\Omega} \to \overline{g(\Omega)}$ be the continuous extension of g. By the Circle Index Lemma 2.2, the index of the restriction of F to J is 1 and the index of the restriction of F to K is -1, when K has the orientation induced by Ω; the sum of these is 0. But F has a fixed point at p, which contradicts Corollary 2.4 with $B_0 = \{K, J\}$. This completes the proof of Lemma 5.2. □

Proof of Theorem 3.2. We will prove by transfinite induction on α that, for each countable ordinal α, the map f extends continuously to a homeomorphism from $\Omega \cup (\bigcup_{K \in W_\alpha} K)$ to $\Omega^* \cup (\bigcup_{K^* \in W_\alpha^*} K^*)$, where W_α is the collection of boundary components in W having rank at most α and W_α^* is the collection of corresponding boundary components of Ω^*.

Suppose that this holds for all ordinals $\beta < \alpha$. Let $K \in W$ be a boundary component of rank α and let K^* be the corresponding boundary component of Ω^*. Let $J \subset \Omega$ be a Jordan curve separating K from every other boundary component of Ω that has rank $\geqslant \alpha$, and from every boundary component outside W. And let Σ be the connected component of $\Omega - J$, which has K as a boundary component. Denote by Σ^* the image of Σ under f, $\Sigma^* = f(\Sigma)$, and let F denote the restriction of f to Σ. We shall show that F extends to a homeomorphism from $\overline{\Sigma}$ onto $\overline{\Sigma}^*$. This will clearly suffice to complete the inductive step.

By the inductive hypothesis we already know that F extends to a homeomorphism from $\overline{\Sigma} - K$ onto $\overline{\Sigma}^* - K^*$. Lemma 5.2 shows that K and K^* are either both points or both circles. If K and K^* are points, then it is obvious

that F extends as needed, and we only consider the case where K and K^* are circles. Since we are free to renormalize by Möbius transformations, we assume, without loss of generality, that $K = K^* = \partial U$ and that $\Sigma, \Sigma^* \subset U$. Now Lemma 5.1 applies and shows that the restriction of F to the set of points having hyperbolic distance at least 1, say, from J extends to a bi-Lipschitz homeomorphism $g : U \to U$. We briefly reproduce here, for the convenience of the reader, a standard geometric argument (maybe due to Mostow), which shows that g extends to a self-homeomorphism of $\overline{U}$. (Alternatively one can conclude that g extends to a homeomorphism of $\overline{U}$ from the fact that g is quasiconformal.)

Let l be the bi-Lipschitz constant of g. Consider some straight ray A of infinite hyperbolic length starting at 0. We now show that $g(A)$ has 1 limit point in ∂U. Take some $r > 0$. Since the diameter of a disk of radius $r+2$ is $2(r+2)$, the preimage under g of the disk with center the origin and radius $r+2$ has diameter at most $2(r+2)l$. Therefore the total length of the part of $g(A)$, whose distance from 0 is between r and $r+2$, is at most $2(r+2)l^2$. This implies that

$$\sum_{E \in H(r)} \theta(E) \leqslant 2\pi \frac{2(r+2)l^2}{p(r)},$$

where $H(r)$ denotes the collection of connected components of the intersection of $g(A)$ with the open annulus between the circles of radii r and $r+2$ around 0, and where each $\theta(E)$ denotes the angular diameter of E with respect to 0, $\theta(E) = \sup_{x,y \in E} \angle(x, 0, y)$, and $p(r)$ is the length of the perimeter of a hyperbolic circle with radius r. Since $p(r)$ increases exponentially as $r \to \infty$, it follows that $\lim_{n \to \infty} \sum_{r=n}^{\infty} \sum_{E \in H(r)} \theta(E) = 0$. This shows that $g(A)$, which clearly has some limit points in ∂U, has in fact a unique limit point there. Now we extend g to ∂U by letting $g(p)$ be the unique limit point of the ray $g(A)$, where A is the ray $[0, p)$ and p is any point in ∂U. One easily uses the above inequalities to verify that g extended thusly is continuous. Checking that it is a homeomorphism is also straightforward.

This clearly implies that f extends as needed, completing the inductive step. An appeal to the principle of transfinite induction now establishes the theorem. (Note that it is not necessary to verify the base of the induction, since the inductive hypothesis is empty when $\alpha = 0$. Of course the base of the induction is also standard.) □

6. Maximum modulus, normality and angles

The results of this section, besides their independent interest, will prepare us for the proof of the existence part of Theorem 0.1.

MAXIMUM MODULUS THEOREM 6.1. *Let A and A^* be Jordan domains in $\mathbb{C}$; let Ω be a domain which is obtained from A by the deletion of a closed disjoint union of at most countably many closed (geometric) disks and points in A; and similarly let Ω^* be a domain obtained from A^* by the deletion of a closed disjoint union of at most countably many closed disks and points. Suppose that $f : \Omega \to \Omega^*$ is a conformal homeomorphism between Ω and Ω^* that extends continuously to ∂A, and that $f^B(\partial A) = \partial A^*$. Then*

$$f(z) - z \in \text{convex hull } \{f(w) - w : w \in \partial A\}$$

for every point $z \in \Omega$. In particular,

$$\sup_{z\in\Omega} |f(z) - z| = \max_{w\in\partial A} |f(w) - w|.$$

Proof. Note first that Theorem 3.2 implies that f extends to a homeomorphism between the closures of Ω and Ω^*.

Let $z_0 \in \Omega$ and define $g(z) = f(z) - f(z_0) + z_0$. Then z_0 is a fixed point for g. Assume that $f(z_0) - z_0$ is not in the convex hull of $\{f(w) - w : w \in \partial A\}$. It then follows that 0 is not in the convex hull of $\{g(w) - w : w \in \partial A\}$. Therefore the winding number around 0 of the restriction of $w \to g(w) - w$ to ∂A is 0. This means that the restriction of g to ∂A has index 0. However g has a fixed point at z_0, in contradiction to Corollary 2.4. This proves the first assertion, and the second assertion clearly follows. □

By the same method as above, it is possible to get estimates analogous to Cauchy's estimates for the first and second derivative. This will be done in a subsequent paper.

COROLLARY 6.2 (Normality). *Let Ω be as above and let $f_k : \Omega \to \Omega_k^*$ be a sequence of conformal homeomorphisms such that each f_k and Ω_k^* satisfy the conditions placed on f and Ω^* above. Suppose that the f_k converge uniformly on compact subsets of Ω to a function g. Then g is either a constant map or a conformal homeomorphism, and $g^B(K)$ is a circle or a point for every boundary component $K \in B(\Omega) - \{\partial A\}$. Moreover the convergence to g is uniform on any subset of Ω whose closure does not intersect ∂A.*

Proof. We start with the last assertion. Let E be some subset of Ω whose closure does not intersect ∂A, and let $\epsilon > 0$. There is some Jordan curve $\gamma \subset \Omega$ separating E from ∂A. The convergence on γ is uniform. Therefore there is some integer N so that $|f_m(z) - f_k(z)| < \epsilon$ for all $k, m > N$ and all $z \in \gamma$; equivalently, $|f_m \circ f_k^{-1}(w) - w| < \epsilon$ for $k, m > N$, $w \in f_k(\gamma)$. Now apply the Maximum Modulus Theorem 6.1 to the maps $f_m \circ f_k^{-1}$ to conclude that $|f_m \circ f_k^{-1}(w) - w| < \epsilon$ for all $k, m > N$ and all $w \in f_k(E)$. This gives

$|f_m(z) - f_k(z)| < \epsilon$, for all $k, m > N$ and all $z \in E$, and implies the uniform convergence.

That g is either a constant or a conformal homeomorphism is a well-known consequence of Rouché's theorem. Let $K \in B(\Omega) - \{\partial A\}$, let W be an open set that contains K and whose closure is disjoint from ∂A, and set $W_0 = W \cap \Omega$. By the above, the convergence $f_k \to g$ is uniform on W_0. Let $\epsilon > 0$ and let k be large enough that $|f_k(z) - g(z)| < \epsilon$ for $z \in W_0$. Then clearly the distance from any point in $g^B(K)$ to $f_k^B(K)$ is at most ϵ, and the distance from any point in $f_k^B(K)$ to $g^B(K)$ is at most ϵ. In other words, the Hausdorff distance from $g^B(K)$ to $f_k^B(K)$ is at most ϵ. Since $f_k^B(K)$ is a circle or a point, and since ϵ was arbitrary, we conclude that $g^B(K)$ is a circle or a point, because the collection of circles and points is closed in the Hausdorff metric. This completes the proof of the normality corollary. □

Definition 6.3. Let $\eta \subset \hat{\mathbb{C}}$ be a circular arc with endpoints p, q and let z be some point not on the circle containing η. We define the angle of η from z, denoted by $\operatorname{ang}(z, \eta)$, to be the length of $m(\eta)$, where m is any Möbius transformation taking z to 0 and η into ∂U, the unit circle. (This is the same as the angle at z between the two circular arcs that join z to p and q, respectively, and which are orthogonal to η.)

The definition is clearly Möbius invariant.

ANGLE LEMMA 6.4. *Let Ω be a domain obtained from a Jordan domain $A \subset \hat{\mathbb{C}}$ by the removal of a closed, disjoint, countable union of disks and points in A. Suppose that f is a conformal homeomorphism from Ω onto a circle domain that extends continuously to a homeomorphism from ∂A to a circle. Further suppose that $z_0 \in \Omega$ and D is an open geometric disk containing z_0 such that the boundary of the connected component of $D \cap A$ that contains z_0, is the union of the arcs $\alpha \subset \partial A \cap D$ and $\beta \subset \partial D \cap A$, as in Figure* 6.1. *Then*

$$\operatorname{ang}\big(f(z_0), f(\alpha)\big) \geqslant \operatorname{ang}(z_0, \eta),$$

where η is the arc of ∂D complementary to β.

In a slight abuse of notation we are using f to denote also the continuous extension of f to ∂A.

Proof. By normalizing with Möbius transformations, we assume without loss of generality that f maps ∂A onto ∂D, respecting orientation, and that $D = U$, $z_0 = f(z_0)$ and $\infty \in \Omega$. Striving for a contradiction, we assume that $\operatorname{ang}(f(z_0), f(\alpha)) < \operatorname{ang}(z_0, \eta)$, which is equivalent to $\operatorname{ang}(f(z_0), f(\alpha)) + \operatorname{ang}(z_0, \beta) < 2\pi$. Then, by further normalizing with a hyperbolic isometry of U that fixes z_0, we assume that β and $f(\alpha)$ are disjoint.

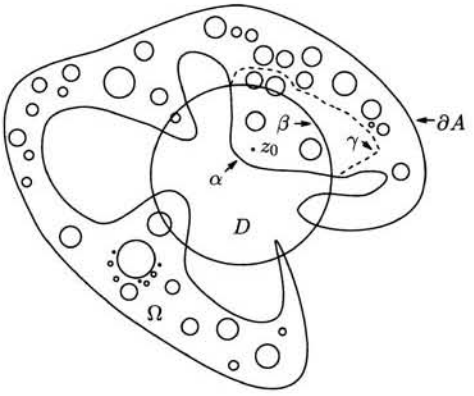


FIGURE 6.1

Now let $\gamma \subset \Omega - \overline{U}$ be some simple curve with endpoints in $\partial A - U$, such that $\gamma \cup \partial A$ separates z_0 from any other connected component of $U \cap A$ and from ∞. (See Figure 6.1.) It is easy to check that such a γ exists. Let Ω^- be the connected component of $\Omega - \gamma$ containing z_0. We now examine the index of the restriction of f to the boundary component K_1 of Ω^- that contains γ and α. Let f_1 denote that restriction and let $H : \alpha \times [0,1] \to U \cup \beta$ be an endpoints-fixing homotopy in $U \cup \beta$ from the identity map on α to some homeomorphism h_1 from α to β. Since $f_1(\alpha) = f(\alpha) \subset \partial U$ is disjoint from β, it follows that $H(z,t) \neq f_1(z)$ for each $z \in \alpha, t \in [0,1]$. Therefore the index of f_1 is equal to the index of the map $f_2 : \beta \cup K_1 - \alpha \to \partial U$, defined by $f_2(z) = f_1(z)$ for $z \in K_1 - \alpha$ and $f_2(z) = f_1 \circ h_1^{-1}(z)$ for $z \in \beta$. But the index of f_2 is clearly 0, because the Jordan domain in $\hat{\mathbb{C}}$ determined by $\beta \cup K_1 - \alpha$ is disjoint from U. So the index of f_1 is 0.

However, every boundary component $K \in B(\Omega^-) - \{K_1\}$ is a circle or a point, and f maps these to circles and points (f extends continuously to the boundary, by Theorem 3.2). Since $f(z_0) = z_0$ and the index of the restriction f_1 of f to K_1 is 0, this gives a contradiction to Corollary 2.4, as usual. This completes the proof of the lemma. □

We will also need the following elementary geometric lemma:

LEMMA 6.5. *Let η be a circular arc with endpoints p, q, say. Suppose that z is a point not on the circle containing η and let δ_z be the circular arc with endpoints p, q that passes through z. Then* $\operatorname{ang}(z,\eta) + 2\theta = 2\pi$, *where θ is the angle between η and δ_z at p (or at q).*

Proof. By normalizing with a Möbius transformation, we assume that $z = \infty$. Let o be the center of the circle containing η. Then $\operatorname{ang}(o,\eta) = \operatorname{ang}(z,\eta)$.

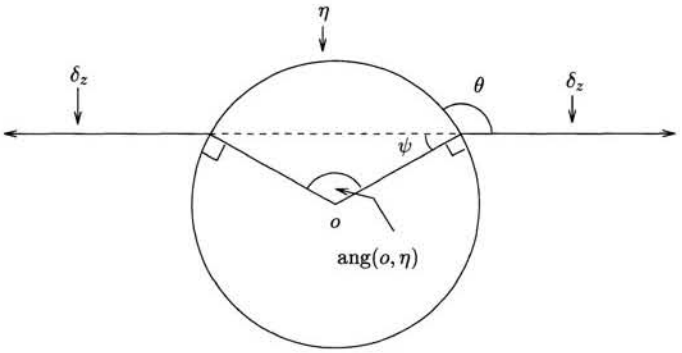


FIGURE 6.2

Now consider the angle ψ indicated in Figure 6.2. Clearly $\psi + \pi/2 = \theta$ and $\mathrm{ang}(o,\eta) + 2\psi = \pi$. From these the lemma follows. □

7. Uniformization

We will now restate and prove the existence part of Theorem 0.1.

UNIFORMIZATION THEOREM 7.1. *Every connected planar domain Ω with at most countably many boundary components is conformally equivalent to a circle domain.*

Proof. The proof will proceed by transfinite induction on the type of Ω. Recall that the type of Ω, denoted by $\mathrm{tp}(\Omega)$, is defined as the pair (λ, n) such that λ is a countable ordinal, n is a positive integer and $B(\Omega)^\lambda$ has n elements. The collection of such pairs is ordered lexicographically; that is, $(\lambda_1, n_1) < (\lambda_2, n_2)$ if $\lambda_1 < \lambda_2$, or $\lambda_1 = \lambda_2$ and $n_1 < n_2$. Since this is a well ordering, one can transfinitely induct with respect to it.

Let (λ, n) be the type of Ω. If $\lambda = 0$, that is, if Ω has finitely many boundary components, then the existence was proved by Koebe. Therefore we will assume that Ω has infinitely many boundary components and the theorem holds for all domains of lesser type. Let K_0 be some boundary component of Ω of rank λ ($K_0 \in B(\Omega)^\lambda$). Let J_k, $k = 1, 2, \ldots$, be a sequence of Jordan domains satisfying $\partial J_k \subset \Omega$, $J_k \subset J_{k+1}$ and $\bigcup_{k=1}^{\infty} J_k \supset \overline{\Omega} - K_0$. Define $\Omega_k = J_k \cap \Omega$.

By the inductive hypothesis, since $\mathrm{tp}(\Omega_k) < \mathrm{tp}(\Omega)$, each Ω_k is conformally homeomorphic to some circle domain. For each k let $f_k : \Omega_k \to \Omega_k^*$ be such a homeomorphism. By normalizing with a Möbius transformation, we assume without loss of generality that $f_k^B(\partial J_k) = \partial U$ and $f_k(z_0) = 0$, where z_0 is some arbitrary point in Ω_1. Since the sequence $\{f_k\}$ is a normal family, by choosing a subsequence if necessary, we also assume that the maps f_k converge

uniformly on compact subsets of Ω. Let f be the limit of the sequence $\{f_k\}$. Then f is either the constant 0 or a conformal homeomorphism $f : \Omega \to \Omega^*$, where $\Omega^* \subset U$. We will consider these two cases separately.

Hyperbolic Case ($f \neq$ constant). Since the maps f_k converge to f uniformly on compacts of Ω, we conclude from Corollary 6.2 that $f^B(K)$ is a circle or a point whenever $K \in B(\Omega) - \{K_0\}$. The application of that corollary is feasible here, because every such K can be separated from K_0 by a Jordan curve in Ω. It only remains to show that $f^B(K_0) = \partial U$.

LEMMA 7.2. *Let W be the connected component of $U - f^B(K_0)$ which contains $\Omega^* = f(\Omega)$. Then W is convex in the hyperbolic metric d_{hyp} of U.*

Proof. Let x, y be 2 distinct points in Ω and let their images under f be $x^*, y^* \in \Omega^*$, $x^* = f(x)$, $y^* = f(y)$. Let $\epsilon > 0$ and let m be such that $x, y \in \Omega_m$, and $d_{\text{hyp}}(x^*, f_m(x)) < \epsilon$ and $d_{\text{hyp}}(y^*, f_m(y)) < \epsilon$. Let e be the hyperbolic line segment joining $f_m(x)$ and $f_m(y)$. For $k > m$ now consider the map $h_k = f_k \circ f_m^{-1}$, whose domain is $f_m(\Omega_m)$. By the Schwarz–Pick lemma, Theorem 0.6, it follows that h_k is a contraction in the hyperbolic metric and extends to a contraction $\hat{h}_k$ of U. Define $e_k = \hat{h}_k(e)$. Then the length of each path e_k is at most the length of e, which is less than $d_{\text{hyp}}(x^*, y^*) + 2\epsilon$. Taking a limit of the e_k, we get some curve $\hat{e}$, which joins x^* and y^*, has length at most $d_{\text{hyp}}(x^*, y^*) + 2\epsilon$ and obviously lies in W.

This shows that any 2 points in Ω^* can be joined by a path in W, whose length is arbitrarily close to the distance between the points. Since Ω^* is open and W is simply connected, this implies that W contains the convex hull of Ω^*; but because $\partial W \subset \partial\Omega^*$, we see that W is the convex hull of Ω^*. This completes the proof of the lemma. □

Knowing now that W is convex, in order to reach a contradiction assume that $W \neq U$. Then there is some hyperbolic line L, which contains a point in ∂W, say p, and has W entirely on one side of it. (Through every point $p \in \partial W \cap U$ passes such a line L.) Let L' be the hyperbolic line that is an arc of a euclidean circle with euclidean radius 1 and whose euclidean midpoint p' lies on the negative real ray. Let g be the hyperbolic orientation-preserving isometry that takes L to L' and p to p' and takes W into the region to the left of L'. (That is, $\text{Re}(w) < \text{Re}(p')$ for $w \in g(W)$.) Now let q be the *euclidean* translation $q(z) = z + 1 - p'$. Then q takes $g(W)$ into U. Furthermore q is clearly (strongly) expanding in the hyperbolic metric at points in $g(W)$ near p'. Therefore the map $q \circ g$ takes Ω^* into U and is expanding in the hyperbolic metric at some point of Ω^* near p, at $x^* = f(x)$, say. Let $\alpha > 1$ be the expansion factor of $q \circ g$ at x^*.

Since $f_k \to f$ and $f_k' \to f'$ as $k \to \infty$, uniformly on compact subsets of Ω, the expansion factor of the map $f \circ f_k^{-1}$ at $f_k(x)$ tends to 1. Let k be large enough that this expansion factor β is greater than $1/\alpha$. Then the map $q \circ g \circ f \circ f_k^{-1} : f_k(\Omega_k) \to q(g(\Omega^*))$ has an expansion factor $\beta\alpha > 1$ at $f_k(x)$. But this contradicts the Schwarz–Pick lemma, Theorem 0.6, since the image of this map is contained in U, and q and g preserve circles. This contradiction completes the proof of the uniformization theorem in the hyperbolic case.

Parabolic Case ($f(\Omega) = \{0\}$). Define maps $g_k(z) = f_k(z)/f_k'(z_0)$, $g_k : \Omega_k \to \mathbb{C}$. These maps clearly form a normal family, since $g_k'(z_0) = 1$. Therefore we assume without loss of generality that the g_k converge uniformly on compact subsets of Ω to some map, say, g. Since $g'(z_0) = 1$, g cannot be a constant and therefore is a conformal homeomorphism of Ω onto a domain, say, Ω^*. As in the hyperbolic case, it follows from Corollary 6.2 that each boundary component of Ω^*, with the possible exception of $g^B(K_0)$, is a circle or a point. We will show that $K_0^* = g^B(K_0)$ is a point (the point ∞), and then the proof will be complete. Striving for a contradiction, assume that K_0^* consists of more than a single point and let F be the connected component of $\hat{\mathbb{C}} - \Omega^*$ containing K_0^*.

We first consider the case where F has interior points. Then there is a Möbius transformation m, which maps Ω^* into U, and with $\infty \in m(F)$. For each k the map

$$m \circ g \circ f_k^{-1} : f_k(\Omega_k) \to U$$

satisfies the hypotheses of the Schwarz–Pick lemma, Theorem 0.6, and is therefore a contraction in the hyperbolic metric. This implies that the inverse map $f_k \circ g^{-1} \circ m^{-1}$ is an expansion, which clearly contradicts our assumption that $f_k \to 0$ as $k \to \infty$. The contradiction shows that $\operatorname{int} F = \emptyset$.

The argument for the case where $\operatorname{int} F = \emptyset$ is more involved, but will still use the Schwarz–Pick lemma. Let $D \subset \hat{\mathbb{C}}$ be some open geometric disk, whose closure $\overline{D}$ contains F such that there are at least two distinct points, say, p and q, in $\partial D \cap F$. One can take for $\overline{D}$, for example, the closed disk with minimal radius, which contains F, in the spherical metric. (We allow $\infty \in \overline{D}$.) We also assume that at least one of the two relatively open arcs of ∂D with endpoints p and q is disjoint from F, as we may, without loss of generality. Let δ be such an arc.

For each angle θ let δ_θ be the circular arc whose endpoints are p and q, such that the oriented angle from δ to δ_θ at p is θ. We take δ_θ to be a relatively open arc; that is, $p, q \notin \delta_\theta$. For every $z \in \hat{\mathbb{C}} - \{p, q\}$ let $\theta(z)$ be that angle θ such that $z \in \delta_{\theta(z)}$. This defines $\theta(z)$ modulo 2π. However, since $\hat{\mathbb{C}} - F$ is simply connected, there is a continuous function $\hat{\theta} : \hat{\mathbb{C}} - F \to \mathbb{R}$ with $\hat{\theta}(z) = \theta(z)$ modulo 2π for $z \in \hat{\mathbb{C}} - F$.

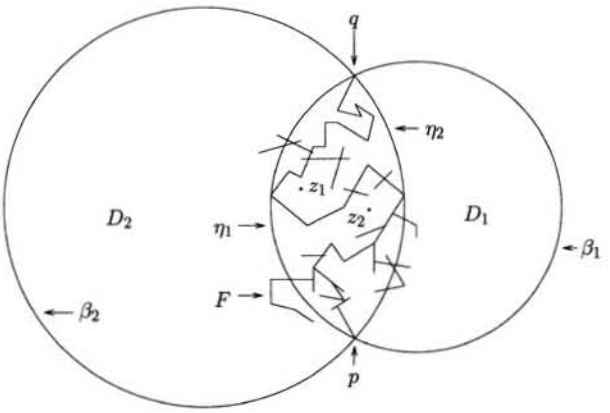


FIGURE 7.1

Now let $\theta_1 = \inf\{\hat{\theta}(z) : z \in \Omega^*\} = \inf\{\hat{\theta}(z) : z \in \hat{\mathbb{C}} - F\}$, and let $\theta_2 = \sup\{\hat{\theta}(z) : z \in \Omega^*\} = \sup\{\hat{\theta}(z) : z \in \hat{\mathbb{C}} - F\}$. From the fact that $F \subset \overline{D}$ it follows that θ_1 and θ_2 are finite. Let z_1 be a point in Ω^* with $\hat{\theta}(z_1) < \theta_1 + (\pi/4)$ and let $z_2 \in \Omega^*$ with $\hat{\theta}(z_2) > \theta_2 - (\pi/4)$. Let $\eta_1 = \delta_{\theta_1}$, $\eta_2 = \delta_{\theta_2}$, $\beta_1 = \delta_{\theta_1+\pi}$ and $\beta_2 = \delta_{\theta_2-\pi}$. Let D_1 and D_2 be the open disks

$$D_1 = \bigcup_{\theta_1<\theta<\theta_1+\pi} \delta_\theta, \qquad D_2 = \bigcup_{\theta_2-\pi<\theta<\theta_2} \delta_\theta.$$

Then we have $z_j \in D_j$, $\partial D_j = \beta_j \cup \eta_j \cup \{p, q\}$, for $j = 1, 2$. Let H_1 be the connected component of $D_1 - F$, which contains z_1, and let H_2 be the connected component of $D_2 - F$, which contains z_2. (See Figure 7.1.)

LEMMA 7.3. *$\hat{\beta}_1 = \partial H_1 - F$ and $\hat{\beta}_2 = \partial H_2 - F$ are subarcs of β_1 and β_2, respectively.*

Proof. Clearly $\partial H_1 - F \subset \partial D_1$. From the definition of η_1 it follows that $\partial H_1 \cap \eta_1 \subset F$, and this gives $\partial H_1 - F \subset \beta_1$.

Suppose now that $\partial D_1 \neq \partial D$. Then the arc η_1 is contained in either D or the complement of $\overline{D}$, because its endpoints p and q are in ∂D. But since there must be points $z \in F \subset \overline{D}$ with $\theta(z)$ arbitrarily close to θ_1, and since $\eta_1 = \delta_{\theta_1}$, it is impossible that η_1 is disjoint from $\overline{D}$. Therefore $\eta_1 \subset D$ and β_1 is disjoint from $\overline{D}$, which gives $\beta_1 \cap F = \emptyset$. This implies that $\hat{\beta}_1 = \beta_1$ in the case where $\partial D \neq \partial D_1$.

If $\partial D_1 = \partial D$, then we use the assumption that δ is an arc of ∂D with endpoints p and q, which is disjoint from F. Then δ must be equal to η_1 or β_1; and thus $F \cap \eta_1 = \emptyset$ or $F \cap \beta_1 = \emptyset$. If $F \cap \beta_1 = \emptyset$, then we get again $\hat{\beta}_1 = \beta_1$. Therefore suppose that $F \cap \eta_1 = \emptyset$. (A priori this can happen.) From the definitions of θ_1 and η_1 it then follows that F must intersect D_1.

Since $\partial D = \partial D_1$ and $F \subset \overline{D}$, this shows that $D = D_1$. This implies that $\hat{\beta}_1 = \partial H_1 - F$ is connected, because given 2 points $x, y \in \hat{\beta}_1$, one can connect them by a simple path γ in $H_1 \cup \{x, y\}$, and F, being connected, must be contained in one of the two connected components of $\overline{D_1} - \gamma$. This proves that $\hat{\beta}_1$ is a subarc of β_1; the proof for $\hat{\beta}_2$ is similar. □

Returning to the proof of the Uniformization Theorem 7.1, we will now show that $H_1 \cap H_2 = \emptyset$. Observe that $\theta_1 < \hat{\theta}(z) < \theta_1 + \pi$ for $z \in H_1$, because H_1 is connected, $z_1 \in H_1 \subset D_1 = \bigcup\{\delta_\theta : \theta_1 < \theta < \theta_1 + \pi\}$ and $\theta_1 < \hat{\theta}(z_1) < \theta_1 + \pi/2$. Similarly $\theta_2 - \pi < \hat{\theta}(z) < \theta_2$ for $z \in H_2$. So it will be sufficient to show that $\theta_1 + 2\pi \leqslant \theta_2$. For every $z \notin F$ we have $\theta_1 < \hat{\theta}(z) < \theta_2$. Thus if θ is a number satisfying $\theta_1 + 2\pi > \theta > \theta_2$, then θ cannot equal $\hat{\theta}(z)$ modulo 2π. Consequently $\bigcup\{\delta_\theta : \theta_1 + 2\pi > \theta > \theta_2\} \subset F$. However, since F has empty interior, this gives $\theta_1 + 2\pi \leqslant \theta_2$, which establishes $H_1 \cap H_2 = \emptyset$.

Let k be some integer such that $g^{-1}(z_1), g^{-1}(z_2) \in \Omega_k$. Let A be a Jordan domain, which contains $g(\Omega_k)$ and whose boundary ∂A is contained in Ω^*. We also require that ∂A intersect $\hat{\beta}_1$ in precisely 2 points p_1, q_1 and intersect $\hat{\beta}_2$ in precisely 2 points p_2, q_2. Thanks to Lemma 7.3, it is not difficult to see that such a Jordan domain exists. (One can just take some Jordan curve in Ω^*, which circles around F very close to F, and then modify it, if necessary, to avoid access intersections with $\hat{\beta}_1$ and $\hat{\beta}_2$. The domain disjoint from F bounded by this curve is taken as A.)

By the inductive hypothesis, there is a conformal map $f_A : A \cap \Omega^* \to U$, which takes each boundary component of $A \cap \Omega^*$ to a circle or a point, and with $f_A(\partial A) = \partial U$. Let $\alpha_1 = \partial A \cap H_1$, $\alpha_2 = \partial A \cap H_2$, $\beta = \hat{\beta}_1 \cap A$ and $\eta = \partial D_1 - \beta$. Now we can apply the Angle Lemma 6.4, with $f_A, A \cap \Omega^*, z_1, D_1, \alpha_1$ in place of $f, \Omega, z_0, D, \alpha$, respectively. We then conclude that

$$\operatorname{ang}\big(f_A(z_1), f_A(\alpha_1)\big) \geqslant \operatorname{ang}(z_1, \eta) > \operatorname{ang}(z_1, \eta_1).$$

By similar reasoning we also have

$$\operatorname{ang}\big(f_A(z_2), f_A(\alpha_2)\big) > \operatorname{ang}(z_2, \eta_2).$$

Because $\theta_1 < \hat{\theta}(z_1) < \theta_1 + (\pi/4)$, it follows from Lemma 6.5 that $\operatorname{ang}(z_1, \eta_1) > 3\pi/2$. Likewise $\operatorname{ang}(z_2, \eta_2) > 3\pi/2$. From the above we conclude that

$$\operatorname{ang}\big(f_A(z_1), f_A(\alpha_1)\big) > 3\pi/2 \tag{7.1}$$

and

$$\operatorname{ang}\big(f_A(z_2), f_A(\alpha_2)\big) > 3\pi/2. \tag{7.2}$$

But $\alpha_1 \cap \alpha_2 = \emptyset$, because $H_1 \cap H_2 = \emptyset$; and therefore, $f_A(\alpha_1) \cap f_A(\alpha_2) = \emptyset$. It thus follows from (7.2) that

$$\operatorname{ang}\bigl(f_A(z_2), f_A(\alpha_1)\bigr) < \pi/2.$$

When compared with (7.1), this gives a lower bound $C > 0$ for the hyperbolic distance from $f_A(z_1)$ to $f_A(z_2)$ in the hyperbolic metric on U. The number C is an absolute constant, which can be described as the hyperbolic distance between two distinct circular arcs in U that have common endpoints in ∂U, each having an angle of $\pi/4$ with ∂U (again by Lemma 6.5).

The domain $A \cap \Omega^*$ contains $g(\Omega_k)$, and so we can consider the map $f_A \circ g \circ f_k^{-1} : f_k(\Omega_k) \to U$. By the Schwarz–Pick lemma, Theorem 0.6, it follows that this map is a contraction in the hyperbolic metric. This tells us that

$$d_{\text{hyp}}\bigl(f_k(g^{-1}(z_1)), f_k(g^{-1}(z_2))\bigr) \geqslant d_{\text{hyp}}\bigl(f_A(z_1), f_A(z_2)\bigr) > C,$$

and $d_{\text{hyp}}(f_k(g^{-1}(z_1)), f_k(g^{-1}(z_2))) > C > 0$ holds for every sufficiently large k. That is a contradiction to our assumption that $f_k \to 0$ as $k \to \infty$. And this contradiction completes the proof of the theorem. □

8. Domains in Riemann surfaces

In this section we prove Theorem 0.2. We will make use of the space of ends $\mathcal{E}(\Omega)$ of a Riemann surface Ω, which is defined exactly as in Section 1. The term *closed surface* means a compact surface with no boundary and an *open surface* means a surface without boundary. We use the term "surface" to mean "connected surface".

Proof of Theorem 0.2. We start with the proof of existence. If the genus of Ω is 0, then Ω can be conformally embedded in the Riemann sphere $\hat{\mathbb{C}}$, and existence follows from Theorem 0.1. Assume therefore that the genus is nonzero.

Because Ω has finite genus, there is some compact subset $F \subset \Omega$ such that each connected component of $\Omega - F$ is a 0-genus (planar) surface, which has one boundary component in Ω. It follows that there is a topological embedding $i : \Omega \to S$ of Ω into a closed topological surface such that every boundary component of $i(\Omega)$ in S is contained in some topological disk in S. (One can "fill in the holes" in each connected component of $\Omega - F$.) Let $\tilde{S}$ be the universal cover of S, let $p : \tilde{S} \to S$ be the covering map and let $\tilde{\Omega} = p^{-1}(i(\Omega))$. Then $\tilde{\Omega}$ is a covering surface for Ω, with covering map $p_\Omega = i^{-1} \circ p$, and thus has the structure of a Riemann surface. Moreover $\tilde{\Omega}$ is planar, since $\tilde{S}$ is a topological disk. From this it follows that $\tilde{\Omega}$ can be conformally embedded in the plane and, therefore, by Theorem 0.1, $\tilde{\Omega}$ is conformally homeomorphic to a circle domain $\tilde{\Omega}^* \subset \mathbb{C}$.

Let Γ be the group of deck transformations for the covering $p : \tilde{S} \to S$. The group Γ acts by conformal homeomorphisms on $\tilde{\Omega}$ and, via the conformal homeomorphism of $\tilde{\Omega}$ and $\tilde{\Omega}^*$, it also acts by conformal homeomorphisms on $\tilde{\Omega}^*$. Furthermore the uniqueness part in Theorem 0.1 shows that the action of Γ on $\tilde{\Omega}^*$ is by Möbius transformations.

Suppose that K is some boundary component of $\tilde{\Omega}$ in $\tilde{S}$. Since $p(K)$ is contained in a topological disk in S, it follows that no nontrivial element of Γ stabilizes K. Let e_∞ be the only end of $\tilde{S}$. Consequently no nontrivial element of Γ fixes an end of $\tilde{\Omega}^*$, except for the end e_∞^*, which corresponds to e_∞. Let E_∞ be the connected component of $\hat{\mathbb{C}} - \tilde{\Omega}^*$ corresponding to e_∞^* and let $\tilde{R} = \hat{\mathbb{C}} - E_\infty$. Then Γ acts freely, co-compactly and discretely on $\tilde{R}$. The pair $(\tilde{R}/\Gamma, \tilde{\Omega}^*/\Gamma)$ is then the required pair.

To prove uniqueness let R_1 and R_2 be closed Riemann surfaces, let Ω_1, Ω_2 be circle domains in R_1 and R_2, respectively, and suppose that $h : \Omega_1 \to \Omega_2$ is a conformal homeomorphism. Let $\tilde{R}_1, \tilde{R}_2$ be the universal covers of R_1, R_2, with covering maps p_1, p_2, respectively. We think of $\tilde{R}_1 \subset \hat{\mathbb{C}}$ and $\tilde{R}_2 \subset \hat{\mathbb{C}}$ as being the unit disk, the plane or the sphere. Set $\tilde{\Omega}_1 = p_1^{-1}\Omega_1$, $\tilde{\Omega}_2 = p_2^{-1}\Omega_2$; these are circle domains in $\hat{\mathbb{C}}$. From the fact that Ω_1 and Ω_2 are circle domains in R_1 and R_2, respectively, it follows that the homeomorphism h lifts to a homeomorphism $\tilde{h} : \tilde{\Omega}_1 \to \tilde{\Omega}_2$. From Theorem 0.1 we know that $\tilde{h}$ is a Möbius transformation. Therefore $\tilde{h}$ extends to a conformal homeomorphism $\tilde{H} : \tilde{R}_1 \to \tilde{R}_2$. From the fact that $\tilde{h}$ conjugates the deck transformations of the cover $p_1 : \tilde{\Omega}_1 \to \Omega_1$ to the deck transformations of the cover $p_2 : \tilde{\Omega}_2 \to \Omega_2$, it follows that $\tilde{H}$ conjugates the deck transformations of the cover $p_1 : \tilde{R}_1 \to R_1$ to the deck transformations of the cover $p_2 : \tilde{R}_2 \to R_2$ and therefore descends to a conformal homeomorphism $H : R_1 \to R_2$. The restriction of H to Ω_1 is obviously h. This shows uniqueness, completing the proof of the theorem. □

9. Uniformizations of circle packings

Surely circle packings and circle domains are closely related. The following definitions give a common generalization of these two concepts.

Definitions. Let Ω be any domain in $\hat{\mathbb{C}}$ (or, more generally, in a Riemann surface). A D-*packing in* Ω is an indexed collection $P = \{P_i : i \in V\}$ of compact topological disks in Ω with disjoint interiors. The *nerve*, or *graph*, of the packing P is the abstract graph $G = (V, E)$, whose vertex set is V and where an edge $\langle i, j \rangle$ occurs in E precisely when the disks P_i and P_j intersect. An *interstice* of the packing is a connected component of $\Omega - \bigcup\{P_v : v \in V\}$, a *finite interstice* is an interstice whose boundary is contained in finitely many of the packed disks, and the *carrier* of the packing is the union of all the packed

sets and all the finite interstices. A *decent packing* is a D-packing in which the intersection of any two sets is at most a single point, and the intersection of any three sets is empty. The *limit points* of a packing are the set of all points p in $\hat{\mathbb{C}}$ with the property that every neighborhood of p intersects infinitely many of the packed sets. A packing P in a domain Ω is said to be an *acceptable packing in* Ω if P is decent and has no limit points in Ω. If P is an acceptable packing in Ω, then $\Omega_P = \Omega - \bigcup\{\text{int}(P_i) : i \in V\}$ is a *generalized domain*; specifically it is the generalized domain associated with P and Ω.

Note that, in general, a generalized domain is not a domain, since it is not open. However it is connected and is the closure of its interior.

Given a D-packing P in $\hat{\mathbb{C}}$, one can get a planar embedding of its graph. To do that choose a point in the interior of each packed set to be the image of the associated vertex. Then the image of each edge $\langle i, j\rangle$ can be chosen as a simple path that lies in $P_i \cup P_j$ and connects the images of the vertices. There is no problem in making the images of the edges disjoint, except at the vertices. Thus nerves of planar D-packings are planar.

Of particular interest are nerves that are maximal with respect to being planar; that is, the introduction of one additional edge to the graph would make it nonplanar. These are the 1-skeletons of a triangulation of an open planar surface. We will call them *planar triangulations*, for short, and when we use the term triangulation, it will be implicitly assumed that the triangulation is connected. Planar triangulations have another important property: their embedding in the sphere $\hat{\mathbb{C}}$ is topologically unique (or unique up to reflection, if the orientation of the sphere is taken into account). This means that any two embeddings of a planar triangulation in $\hat{\mathbb{C}}$ are related by a self-homeomorphism of $\hat{\mathbb{C}}$. Thus a decent packing in $\hat{\mathbb{C}}$, whose nerve is a triangulation, is topologically determined by its nerve. We will see below that when the packed sets are geometric disks and the nerve is a given planar triangulation, then under certain conditions the packing is also geometrically determined.

One can study either a packing or the generalized domain associated to it. The difference is like the proverbial difference between looking at the half-full glass or at the half-empty glass. Of course, when there are few edges in the nerve of the packing, there is little to work with, and one must look at the domain. Historically both approaches to the subject are present. For example, Koebe looked at the domain and achieved the circle packing theorem as a consequence of his uniformization theorem. On the other hand, Thurston mostly looked at the circles, while the Rodin–Sullivan work [RoSu] and the paper [He2] adopt a mixture of the two views. In retrospect, the incompatibility theorem, which is the main tool in [Sch2] and [Sch3], can be seen as some kind of fixed-point theorem for packings. There, two finite topological

packings satisfy some boundary conditions analogous to the condition that the associated map on the boundary of the domain have a fixed-point index -1. The conclusion is that there exists a "fixed point of negative index" between the packings.

Definitions. A *conformal homeomorphism* between generalized domains $h : \Omega \to \Omega^*$ is a homeomorphism that is conformal in the interior of Ω, while an *anticonformal homeomorphism* is a homeomorphism that is anticonformal in the interior of Ω. If such an h exists, then Ω and Ω^* are said to be conformally or anticonformally homeomorphic, respectively.

Example 9.1. Let P and P^* be decent packings in $\hat{\mathbb{C}}$ having carriers Ω and Ω^* and planar triangulations $\mathcal{T}$ and $\mathcal{T}^*$ as nerves, respectively. Then P and P^* are acceptable packings in Ω and Ω^*, respectively; and the associated generalized domains Ω_P and $\Omega^*_{P^*}$ are conformally homeomorphic or anticonformally homeomorphic if and only if $\mathcal{T}$ and $\mathcal{T}^*$ are combinatorially isomorphic. To see this, note that all of the interstices in the packings must be triangular interstices; that is, their boundary lies on three of the packed sets. Thus all one has to do to show that Ω_P and $\Omega^*_{P^*}$ are conformally or anticonformally homeomorphic is to construct the conformal, or anticonformal, maps between combinatorically corresponding triangular interstices and glue them properly. Since there is a freedom of choice of the image of three points on the boundary for Riemann maps between Jordan domains, one can do this while maintaining the continuity at the points of contact between any two interstices. This shows that Ω and Ω^* are conformally or anticonformally homeomorphic. The other direction is obvious.

When we speak of a *circle packing*, we will mean a D-packing of geometric, rather than topological, closed disks. If a circle packing has no limit points in a domain Ω, then it follows that it is acceptable in Ω. A circle packing, whose nerve is a triangulation, is always acceptable in its carrier.

Definition. Let Ω be a circle domain and let P be an acceptable circle packing in Ω. The associated generalized domain Ω_P will be called a *generalized circle domain.*

We can now state a generalization of Theorem 0.1, which is applicable to circle packings.

THEOREM 9.2. *Any generalized domain Ω in $\hat{\mathbb{C}}$ that has at most countably many ends is conformally homeomorphic to a generalized circle domain $\Omega^* \subset \hat{\mathbb{C}}$. Moreover Ω^* is unique up to Möbius transformations, and every conformal automorphism of Ω^* is the restriction of a Möbius transformation.*

Theorem 0.3 follows immediately as a corollary:

Proof of Theorem 0.3. We start with existence. One easily constructs a decent planar packing P^* with nerve $\mathcal{T}$. Since P^* is acceptable in its carrier Ω^*, one can form the generalized domain $\Omega^*_{P^*}$. From Theorem 9.2 we conclude that there is a generalized circle domain $\Omega \subset \hat{\mathbb{C}}$ that is conformally homeomorphic to $\Omega^*_{P^*}$. The circle packing associated to Ω is then the required packing P. This proves existence. The uniqueness follows from Example 9.1 and Theorem 9.2. □

To prove Theorem 9.2 one must essentially adapt, for generalized domains, the proof of Theorem 0.1 and the proofs of all the theorems that precede it. There are two minor difficulties, which require some changes in the proofs. The first has to do with the fact that the interior of a generalized domain is not connected. When working with domains, we used the fact that if a conformal function has nonisolated fixed points, then it is the identity. This is no longer true when the domain of the function is not connected and, therefore, in the proof of Theorem 9.2, one must take special care to avoid nonisolated fixed points.

In the proofs above, quite often we have chosen a Jordan curve γ in the domain Ω to cut and isolate a part of the domain we wanted to examine from other parts. This can still be done in generalized domains, but the resulting two pieces that Ω breaks into may no longer be generalized domains. An example of this phenomenon can be seen in the generalized domain Ω_H obtained from the plane by the deletion of the interiors of disks that form an infinite hexagonal circle packing. The bounded part of Ω_H determined by any Jordan curve $\gamma \subset \Omega_H$ will not be a generalized domain, unless it is contained in one interstice, because γ would have to touch some boundary circles more than once. This forces us to further broaden the class of "domains" under discussion.

Definitions. Let $\hat{\Omega}$ be some domain in $\hat{\mathbb{C}}$ and let P be a D-packing in $\hat{\Omega}$. Suppose that the intersection of any three sets in the packing P is empty and the intersection of any two contains at most a finite number of points. Further suppose that at most one of the connected components of the complement of any pair of sets in P intersects with other sets in the packing. Then $\Omega = \hat{\Omega} - \bigcup\{\text{int}(P_v) : v \in V\}$ will be called a *degenerated generalized domain.*

A *bi-gon* in a degenerated generalized domain is a finite interstice whose boundary lies in two of the sets in the packing. Thus a degenerate generalized domain without bi-gons is a generalized domain.

A *morphism* of degenerate generalized domains is a continuous map $f : \Omega \to \Omega^*$ between degenerate generalized domains that is conformal and

injective in $\text{int}(\Omega) - B$, where B is some union of bi-gons of Ω and f is constant in each bi-gon contained in B.

As explained above, the advantage of working with degenerated generalized domains over generalized domains is that it is easy to cut a degenerated generalized domain along a Jordan curve and get two degenerated generalized domains.

Proof of Theorem 9.2. Let $\hat{\Omega}$ and $P = (P_i : i \in V)$ be the domain and the packing, which define Ω; that is, $\Omega = \hat{\Omega} - \bigcup\{\text{int}(P_i) : i \in V\}$.

We need a construct for degenerated generalized domains analogous to the space of boundary components of domains. The elements of this space will be called the *boundary elements*. These come in two flavors: the *elements at infinity* are just the boundary components of $\hat{\Omega}$, and the *border elements* are the sets of the form ∂P_i. The topology on the set of boundary elements $B^e(\Omega)$ is such that any neighborhood in $B^e(\Omega)$ around any element at infinity $K \in B(\hat{\Omega})$ is the set of all boundary elements that intersect a neighborhood of K in $\hat{\mathbb{C}}$, and $\{\partial P_i\}$ is a neighborhood of any border element ∂P_i. (Thus the set of border elements is discrete in $B^e(\Omega)$, and the inclusion of $B(\hat{\Omega})$ in $B^e(\Omega)$ is a homeomorphic embedding.) As with $B(\hat{\Omega})$, the set $B^e(\Omega)$ is compact, Hausdorff and countable. The type and rank are then defined for $B^e(\Omega)$ as for $B(\hat{\Omega})$.

In the following paragraphs we outline the modifications needed in the theorems and lemmas leading to Theorem 0.1 to get corresponding statements for degenerate generalized domains.

Note that Theorem 2.1 also holds if we allow the boundary of A to be a finite union of Jordan curves, which may intersect at finitely many points.

We will need a slightly more general version of Lemma 2.2. In the more general version we do not require f to be a homeomorphism, but we do require the preimage of any point in K to be a point or an arc on J. The hypothesis that f is orientation preserving should then be weakened to the requirement that the image of a positively oriented arc from a point x to a point y in J be a positively oriented arc from $f(x)$ to $f(y)$ in J, or a single point. In this situation part (1) of Lemma 2.2 is dropped, the proof for the case $J \subset \hat{K}$ in part (2) remains unchanged and the proof for the case $K \subset \hat{J}$ is done similarly as the proof for $J \subset \hat{K}$. The proofs of parts (3) and (4) remain unchanged.

Some adjustments are needed in Corollary 2.4 as well. First Ω and Ω^* are permitted to be possibly degenerated generalized domains, with f a morphism between them. Obviously any mention of boundary components is replaced by boundary elements. (This same change is needed in all of the lemmas and theorems we discuss here and will not be mentioned, unless there is some

special need.) Of course we no longer require that F, the continuous extension of f, be a homeomorphism, since f does not have to be a homeomorphism, but only that the restriction of F to each boundary element at infinity be a homeomorphism. The conclusions of the corollary also need some revision. The new conclusions are that f has at most n *isolated* fixed points and the number of fixed points in any set S of isolated fixed points is at most n, counting multiplicities. This change is needed, since f may fix whole connected components of the interior of Ω. The proof remains almost unchanged. One only needs to note that $z \to F(z) + c$ has only isolated fixed points if it has no fixed points on the boundary.

The description of Ω in the statement of the Schwarz–Pick lemma, Theorem 0.6, is modified to the following: Ω is a possibly degenerate generalized domain contained in $\overline{A}$, and every boundary element of it is a circle or a point, except possibly for ∂A, which is also a boundary element of Ω. A similar change is done for Ω^*. The statement of Lemma 4.1 changes only in that the continuous extension of f to $\overline{\Omega}$ is not required to be a homeomorphism, only its restriction to each boundary element at infinity needs to be a homeomorphism. A change is needed in the proof of this lemma, since when one of the fixed points p, q is a nonisolated fixed point, one cannot immediately conclude that f is the identity. Suppose this to be the case. Then f must fix a connected component of the interior of Ω that is not a bi-gon. Let H be the union of the connected components that f fixes. If H is the interior of Ω, the lemma follows; if not, then there is some border boundary element K that has nontrivial arcs in $\overline{H}$ and in the closure of some connected component B of the interior of Ω which is not in H. Since f fixes an arc of K, then $f(K) = K$, because both are circles. If f is the identity on K, then f must fix B, which contradicts our assumptions. If not, then there will be two points, $x, y \in K$, such that $d_{\text{hyp}}(x, y) < d_{\text{hyp}}(f(x), f(y))$. The same would hold for some points x', y' in the interior of Ω sufficiently close to x, y, respectively. Then one can postcompose f with a Möbius transformation m, which contracts distances in the hyperbolic metric on U and takes $f(x')$ and $f(y')$ to x' and y', respectively. Since one has 2 dimensions of freedom in choosing m, one easily arranges that x' and y' will be isolated fixed points of $m \circ f$. Then the contradiction follows as in the original proof of Lemma 4.1.

The statement and proof of Lemma 5.1 for possibly degenerated generalized domains remain essentially unchanged.

In the formulation of Lemma 5.2, again the requirement that the continuous extension of f to $\overline{\Omega} - K$ be a homeomorphism needs to be changed to the requirement that its restriction to each boundary element at infinity $H \in B^e_\infty(\Omega)$ be a homeomorphism. In the proof, the only modification is that

one must make sure that p is an isolated fixed point for g. It is easy to see that p can be chosen so that this is the case. (Note that p is chosen before g, and a different choice of p may give a different choice of g.)

The statement and proof of the Boundary Extension Theorem remain essentially the same, and the generalized form of Theorem 0.6 follows from it and Lemma 4.1. The uniqueness part of Theorem 9.2 now clearly follows from Theorem 0.6.

Except for the obvious modifications, similar to those in the Schwarz–Pick lemma (Theorem 0.6), no change is needed in the formulation of the Maximum Modulus Theorem 6.1. In the proof, one must take care of the possibility that z_0 is a nonisolated fixed point of g. In that case, let H be the union of the connected components of the interior of Ω that are fixed by g. Obviously H cannot contain all of the interior of Ω and, therefore, there is a connected component B of the interior of Ω that is disjoint from H, but whose closure intersects the closure of H. If p is in the intersection of the closures, then in B near p one can find a point q with $g(q) - q$ very close to 0. Furthermore q can be chosen to also satisfy $g'(q) \neq 1$. Then q is an isolated fixed point for $\hat{g}(z) = g(z) - g(q) + q$, and a contradiction follows.

Corollary 6.2 requires substantial changes, and it will be replaced by a discussion below.

The changes necessary in the formulation and proof of the Angle Lemma 6.4 are similar to those made in the previous lemmas and theorems above. These changes are left to the reader.

In the modified proof of existence, the inductive claim is that given a degenerate generalized domain $\Omega \subset \hat{\mathbb{C}}$ with $\mathrm{tp}(\Omega) < (\lambda, n)$ there exists a morphism of it onto a generalized circle domain. It is not very well known, but Koebe [Ko4] also proved the existence statement in the case of degenerate generalized domains with finitely many boundary elements by taking limits of complements of packings, where the packed disks almost touch. This covers the base of the induction.

The maps f_k are defined as in the original version, but one has to work a little harder to argue that their limits are either a constant or a morphism. First it is clear that a subsequence of the f_k converges uniformly on compact subsets of the interior of Ω. Using the reflection principle, one easily concludes that a subsequence of $\{f_k\}$ converges uniformly on compact subsets of Ω. Then exactly the same argument as in Corollary 6.2 shows that $f^B(K)$ is a circle or a point whenever $K \in B^e(\Omega) - K_0$, where f is the limit of the f_k.

We will now show that f is either a constant or a morphism. Restricted to each connected component of the interior of Ω, f is either a constant or a conformal homeomorphism. Also f is clearly injective where it is not a constant. Suppose that f is constant on some interstice L, which is not a

bi-gon. Say it takes the value c there. Let M be the connected component of $f^{-1}(c)$ containing L and let $\partial_\Omega M$ be the relative boundary of M in Ω. If K is some border boundary element contained in M, then clearly every connected component of the interior of Ω whose boundary has an arc on K, is also contained in M. This shows that $\partial_\Omega M$ consists of contact points; that is, points in the intersection of the closures of two distinct interstices. The number of points in $\partial_\Omega M$ cannot be 1; on the other hand, if there are 2 or more contact points in $\partial_\Omega M$, then there are at least 3 border boundary elements intersecting $\partial_\Omega M$. These elements will have the property that their images under f contain c, but also contain other points. However this is impossible, since at most two circles can touch at any given point. This implies that $\partial_\Omega M$ is empty and thus shows that f is a constant if it is constant on some connected component of the interior of Ω that is not a bi-gon.

This same argument is repeated for the hyperbolic and parabolic cases. Except for this, the proof remains intact. This completes the proof of Theorem 9.2. □

Proof of Theorem 0.4. Theorem 0.4 follows from Theorem 9.2 exactly as Theorem 0.2 followed from Theorem 0.1. □

Addendum: Almost circular domains with uncountably many boundary components

We now state and outline the proof of a generalization to the existence part of our main result, Theorem 0.1, which was obtained after this article was accepted. The details will appear in a forthcoming paper.

THEOREM 10.1. *Let Ω be a domain in $\mathbb{C}$ and let $B_*(\Omega) \subset B(\Omega)$ be the collection of all boundary components that are not circles or points. If the closure of $B_*(\Omega)$ in $B(\Omega)$ is countable, then Ω is conformally homeomorphic to a circle domain.*

The proof of Theorem 10.1 proceeds by induction in much the same way as the proof of existence presented earlier. The sticky point is, however, that the rigidity results used, primarily the Schwarz–Pick lemma, require some extra hypotheses when there are uncountably many boundary components. For this purpose, the theory of quasiconformal maps is useful. More specifically the Schwarz–Pick lemma, and most of the other results here, are applications of Corollary 2.4, which in general fails when there are uncountably many boundary components. The first step in the proof of Theorem 10.1 is a variation of that corollary, as follows:

LEMMA 10.2. *Let $f : \Omega \to \Omega^*$ be a conformal homeomorphism between bounded plane domains and let $B_0 \subset B(\Omega)$ be a finite collection of boundary components of Ω. Suppose that f extends to a quasiconformal homeomorphism $F : \mathbb{C} \to \mathbb{C}$ and that the following conditions hold:*

(1) *the conformal dilatation of F on $\partial\Omega$ is 0 a.e. (this is automatically satisfied if $\partial\Omega$ has measure 0);*

(2) *all of the boundary components in $B(\Omega) - B_0$ and $B(\Omega^*) - f^B(B_0)$ are circles and points, and all of the boundary components in B_0 are Jordan curves;*

(3) *B_0 contains the boundary component of Ω, which is contained in the unbounded component of $\mathbb{C} - \Omega$ and, similarly, $f^B(B_0)$ contains the boundary component of Ω^*, which is contained in the unbounded component of $\mathbb{C} - \Omega^*$;*

(4) *F has no fixed points in any of the boundary components in B_0.*

Let n be the index of the restriction of F to the boundary components in B_0. Then f has at most n fixed points in Ω. Furthermore, if S is a set of fixed points of f, then the total number of fixed points for f in S, counting multiplicity, is at most n.

The proof of this lemma is based on approximations by finitely connected domains, on Corollary 2.4 for finitely connected domains and on a rigidity result of Sullivan [Su] to show that the approximations converge to f.

The rest of the proof of Theorem 10.1 is like the proof of the existence part of Theorem 0.1, but with the extra burden that whenever Corollary 2.4 is directly or indirectly applied, it can be arranged that the hypotheses of Lemma 10.2 are satisfied.

PRINCETON UNIVERSITY, PRINCETON, NEW JERSEY
WEIZMANN INSTITUTE, REHOVOTH, ISRAEL

REFERENCES

[A] D. AHARONOV, The hexagonal packing lemma and discrete potential theory, Can. Math. Bull. **33** (1990), 247–252.

[AhBe] L. AHLFORS and L. BERS, Riemann's mapping theorem for variable metrics, Ann. of Math. **72** (1960), 385–404.

[An1] E.M. ANDREEV, On convex polyhedra in Lobačevskiĭ spaces, Mat. Sb. (N.S.) **81** (1970), 445–478; English transl. in Math. USSR Sb. **10** (1970), 413–440.

[An2] ———, On convex polyhedra of finite volume in Lobačevskiĭ space, Mat. Sb. (N.S.) **83** (1970), 256–260; English transl. in Math. USSR Sb. **12** (1970), 255–259.

[BaFP] I. BÁRÁNY, Z. FÜREDI and J. PACH, Discrete convex functions and proof of the six circle conjecture of Fejes Tóth, Can. J. Math. **36** (1984), 569–576.

[BSte1] A.F. BEARDON and K. STEPHENSON, The uniformization theorem for circle packings, Indiana Math. J. **39** (1990), 1383–1425.

[BSte2] ———, Circle packings in different geometries, Tôhoku Math. J. **43** (1991), 27–36.

[BSte3] ———, The Schwarz–Pick lemma for circle packings, Illinois J. of Math. **35** (1991), 577–606.

[Be] L. BERS, Uniformization by Beltrami equations, Comm. Pure Appl. Math. **14** (1961), 215–228.

[BoS] P.L. BOWERS and K. STEPHENSON, The set of circle packing points in the Teichmüller space of a surface of finite conformal type is dense, Math. Proc. Cambridge Phil. Soc. **111** (1992), 487–513.

[Br1] R. BROOKS, On the deformation theory of classical Schottky groups, Duke Math. J. **52** (1985), 1009–1024.

[Br2] ———, Circle packings and co-compact extensions of Kleinian groups, Invent. Math. **86** (1986), 461–469.

[CR] I. CARTER and B. RODIN, An inverse problem for circle packing and conformal mapping, Preprint.

[CdV1] Y. COLIN DE VERDIÈRE, Empilements de cercles: Convergence d'une méthode de point fixe, Forum Math. **1** (1989), 395–402.

[CdV2] ———, Un principe variationnel pour les empilements de cercles, Invent. Math. **104** (1991), 655–669.

[De] R. DENNEBERG, Konforme Abbildung einer Klasse unendlich vielfach zusammenhangender schlichter Bereiche auf Kreisbereiche, Diss., Leipziger Berichte **84** (1932), 331–352.

[Gr] H. GRÖTZSCH, Eine Bemerkung zum Koebeschen Kreisnormierungsprinzip, Leipziger Berichte **87** (1935), 319–324.

[GPo] V. GUILLEMIN and A. POLLACK, *Differential Topology*, Prentice–Hall, Englewood Cliffs, 1974, p. 222.

[Haa] A. HAAS, Linearization and mappings onto pseudocircle domains, Trans. A.M.S. **282** (1984), 415–429.

[Hau] F. HAUSDORFF, *Set Theory*, 2nd ed., Chelsea Pub. Co., New York, 1962.

[He1] Z.X. HE, Solving Beltrami equations by circle packings, Trans. A.M.S. **322** (1990), 657–670.

[He2] ———, An estimate for hexagonal circle packings, J. Diff. Geom. **33** (1991), 395–412.

[HeSch] Z.X. HE and O. SCHRAMM, A hyperbolicity criterion for circle packings, Preprint.

[Ko1] P. KOEBE, Uber die Uniformisierung beliebiger analytischer Kurven, III, Nachr. Ges Wiss. Gott. (1908), 337–358.

[Ko2] ———, Abhandlungen zur Theorie der Konformen Abbildung: VI. Abbildung mehrfach zusammenhängender Bereiche auf Kreisbereiche, etc., Math. Z. **7** (1920), 235–301.

[Ko3] ———, Uber die konforme Abbildung endlich- und unendlich-vielfach zusammenhängender symmetrischer Bereiche auf Kreisbereiche, Acta Math. **43** (1922), 263–287.

[Ko4] ———, Kontaktprobleme der konformen Abbildung, Berichte Verhande. Sächs. Akad. Wiss. Leipzig, Math.-Phys. Klasse **88** (1936), 141–164.

[LV] O. LEHTO and K.I. VIRTANEN, *Quasiconformal Mappings in the Plane*, Springer–Verlag, Berlin, 1973, p. 258.

[MRo] A. MARDEN and B. RODIN, On Thurston's formulation and proof of Andreev's theorem, in *Computational Methods and Function Theory*, Ruscheweyh, Saff, Salinas, Varga, eds., Lecture Notes in Math. **1435**, Springer–Verlag, 1989, pp. 103–115.

[Me1] H. MESCHOWSKI, Uber die konforme Abbildung gewisser Bereiche von unendlich hohen Zusammenhang auf Vollkreisbereiche, I, Math. Ann. **123** (1951), 392–405.

[Me2] ———, Uber die konforme Abbildung gewisser Bereiche von unendlich hohen Zusammenhang auf Vollkreisbereiche, II, Math. Ann. **124** (1952), 178–181.

[Mi] J.W. MILNOR, *Topology from the Differential Viewpoint*, Univ. Press of Virginia, 1965, p. 64.

[Ro1] B. RODIN, Schwarz's lemma for circle packings, Invent. Math. **89** (1987), 271–289.

[Ro2] ———, Schwarz's lemma for circle packings, II, J. Diff. Geom. **30** (1989), 539–554.

[Ro3] ———, On a problem of A. Beardon and K. Stephenson, Indiana Math. J. **40** (1991), 271–275.

[RoSu] B. RODIN and D. SULLIVAN, The convergence of circle packings to the Riemann mapping, J. Diff. Geom. **26** (1987), 349–360.

[Sa] L. SARIO, Uber Riemannsche Flachen mit hebbarem Rand, Ann. Acad. Sci. Fenn., Ser. AI **50** (1948), 1–79.

[Sch1] O. SCHRAMM, *Packing Two-dimensional Bodies with Prescribed Combinatorics and Applications to the Construction of Conformal and Quasiconformal Mappings*, Ph.D. Thesis, Princeton Univ., 1990.

[Sch2] ———, Rigidity of infinite (circle) packings, J. A.M.S. **4** (1991), 127–149.

[Sch3] ———, Existence and uniqueness of packings with specified combinatorics, Israel J. Math. **73** (1991), 321-341.

[Si1] R.J. SIBNER, Uniformizations of symmetric Riemann surfaces by Schottky groups, Trans. A.M.S. **116** (1965), 79–85.

[Si2] ———, Remarks on the Koebe Kreisnormierungsproblem, Comm. Math. Helv. **43** (1968), 289–295.

[Si3] ———, 'Uniformizations' of infinitely connected domains, in *Advances in the Theory of Riemann Surfaces*, Proc. 1969 Stony Brook Conf., Ann. of Math. Studies **66**, Princeton Univ. Press, 1971, pp. 407–419.

[Ste] K. STEPHENSON, Circle packings in the approximation of conformal mappings, Bull. A.M.S. **23** (1990), 407–415.

[Str1] K.L. STREBEL, Uber das Kreisnormierungsproblem der konformen Abbildung, Ann. Acad. Sci. Fenn. **1** (1951), 1–22.

[Str2] ———, Uber die konforme Abbildung von Gebieten unendlich hohen Zusammenhangs, Comm. Math. Helv. **27** (1953), 101–127.

[Su] D. SULLIVAN, On the ergodic theory at infinity of an arbitrary discrete group of hyperbolic motions, Ann. of Math. Studies **97**, Princeton Univ. Press, 1981, pp. 465–496.

[Th1] W.P. THURSTON, The geometry and topology of 3-manifolds, Lecture notes, Princeton Univ., 1980, Chapter 13.

[Th2] ———, The finite Riemann mapping theorem, invited address, International Symposium in Celebration of the Proof of the Bieberbach Conjecture, Purdue University, 1985.

(Received July 8, 1991)

Discrete Comput Geom 14:123–149 (1995)

Discrete & Computational
Geometry
© 1995 Springer-Verlag New York Inc.

Hyperbolic and Parabolic Packings*

Zheng-Xu He[1] and O. Schramm[2]

[1]University of California at San Diego,
La Jolla, CA 92093, USA
zhe@math.ucsd.edu

[2]Mathematics Department, The Weizmann Institute,
Rehovot 76100, Israel
schramm@wisdom.weizmann.ac.il

Abstract. The contacts graph, or nerve, of a packing, is a combinatorial graph that describes the combinatorics of the packing. Let G be the 1-skeleton of a triangulation of an open disk. G is said to be CP parabolic (resp. CP hyperbolic) if there is a locally finite disk packing P in the plane (resp. the unit disk) with contacts graph G. Several criteria for deciding whether G is CP parabolic or CP hyperbolic are given, including a necessary and sufficient combinatorial criterion. A criterion in terms of the random walk says that if the random walk on G is recurrent, then G is CP parabolic. Conversely, if G has bounded valence and the random walk on G is transient, then G is CP hyperbolic.

We also give a new proof that G is either CP parabolic or CP hyperbolic, but not both. The new proof has the advantage of being applicable to packings of more general shapes. Another new result is that if G is CP hyperbolic and D is any simply connected proper subdomain of the plane, then there is a disk packing P with contacts graph G such that P is contained and locally finite in D.

1. Introduction

We consider packings of compact connected sets in the plane $\mathbb{C} = \mathbb{R}^2$ or in the Riemann sphere $\hat{\mathbb{C}} = S^2$.

Given an indexed packing $P = (P_v: v \in V)$, its *contact graph*, or *nerve* $G = G(P)$, is defined as follows. The set of vertices of G is V, the indexing set for P, and an

* Both authors acknowledge support by NSF grants. The first author was also supported by the A. Sloan Research Fellowship.

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_7, © Springer Science+Business Media, LLC 2011

edge $[v, u]$ appears in G precisely when the sets P_v and P_u intersect. Thus G encodes some of the combinatorics of P. If all the sets P_v are smooth disks[1] in $\mathbb{C}$, then it is easy to see that the contacts graph is planar.

The circle-packing theorem [16] says that for any finite planar graph G there is some packing of (geometric) disks in the plane whose contacts graph is G. This fantastic theorem has received much attention since Thurston conjectured that the Riemann map from a simply connected domain to the unit disk can be approximated using circle packings with prescribed nerves. The conjecture was later proved by Rodin and Sullivan [20]. Some proofs of the circle-packing theorem appear in [1], [2], [28, Chapter 13], [18], [10], [4], [13], [6], [7], [21], [24], and [23].

Here, we are concerned with infinite packings. Suppose, for example, that G is a *disk triangulation graph*; that is, the 1-skeleton of a triangulation of an open topological disk. By taking a Hausdorff limit of packings corresponding to finite subgraphs of G, an infinite packing P of disks in $\mathbb{C}$ whose contacts graph is G can be obtained. A few questions then naturally arise about the properties of P. Can P be bounded? Can P be locally finite in the plane? (This means that every compact subset of the plane intersects finitely many of the sets in the packing.) To what extent is P unique?

It is not hard to see that (still assuming G to be a disk triangulation graph) there is a unique open topological disk $D \subset \hat{\mathbb{C}}$ such that P is contained in D and is locally finite in D. The boundary of D is just the set of accumulation points of P.[2] This D is called the *carrier* of P, and is denoted $\mathrm{carr}(P)$.

It was proved in [15] that P can be chosen such that $\mathrm{carr}(P)$ is the plane or the unit disk $U = \{z \in \mathbb{C}: |z| < 1\}$. Beardon and Stephenson [3] have obtained this result under the additional assumption that G has bounded valence.[3] There is a strong uniqueness statement valid when $\mathrm{carr}(P) = \mathbb{C}$: any other disk packing $P' \subset \hat{\mathbb{C}}$ with nerve G is the image of P under a Möbius transformation [22], [15]. (The Möbius group is the group generated by inversions in circles. It is six dimensional.) In particular, it follows that there cannot be two disk packings P, P' with $\mathrm{carr}(P) = \mathbb{C}$, $\mathrm{carr}(P') = U$, and $G = G(P) = G(P')$. If $\mathrm{carr}(P) = U$, then there is a weaker form of uniqueness: any disk packing P' with $\mathrm{carr}(P') = U$ that has nerve G is the image of P under a Möbius transformation.

All this parallels neatly with the analytic theory. The existence of a locally finite packing in U or $\mathbb{C}$ is a discrete analog of the uniformization theorem, which says that any simply connected noncompact Riemann surface is conformally equivalent to $\mathbb{C}$ or U. The parallels of the uniqueness statements are that any conformal map from the plane into the sphere or from U onto U is a Möbius transformation.

Let us say that a disk triangulation graph G is *CP parabolic* (resp. *CP hyperbolic*)

---

[1] The term *disk* means a geometric disk, a *topological disk* means a set homeomorphic to a compact disk, and a *smooth disk* is a topological disk with C^1 boundary.

[2] A point z is an accumulation point of P if every neighborhood of z intersects infinitely many sets in P.

[3] The *valence* or *degree* of a vertex is the number of neighbors it has. "G has bounded valence" means that there is some $C < \infty$ such that every vertex has valence less than C.

if there is a disk packing P with contacts graph G and carrier $\text{carr}(P) = \mathbb{C}$ (resp. $\text{carr}(P) = U$).

We introduce the notion of a VEL parabolic graph. VEL parabolicity is a combinatorial property, which is defined using Cannon's vertex extremal length [8]. The precise definitions appear later. A graph which is not VEL parabolic is called VEL hyperbolic. We prove that a disk triangulation graph is CP parabolic iff it is VEL parabolic. This gives a complete combinatorial characterization of the "CP type" of any disk triangulation graph.

Using this equivalence of CP parabolic and VEL parabolic, we prove:

1.1. Theorem. *Let G be a disk triangulation graph. If the random walk on G is recurrent, then G is CP parabolic. Conversely, if the degrees of the vertices in G are bounded and the random walk on G is transient, then G is CP hyperbolic.*

It will be shown that there are CP parabolic disk triangulation graphs on which the random walk is transient.

We also give new proofs to the above-quoted results that every disk triangulation graph is either CP parabolic, or CP hyperbolic, but not both. The results here actually generalize these theorems, since the proofs apply not only to packings by geometric disks, but to more general sets. In order to state some of our results, we introduce the notion of fat sets. Heuristically, a set is fat if its area is roughly proportional to the square of its diameter, and this property also holds locally. The precise definition is:

Definitions [26]. The open disk with center x and radius r is denoted $D(x, r)$. Let $\tau > 0$. A measurable set $X \subset \hat{\mathbb{C}}$ is *τ-fat* if, for every $x \in X$, $x \neq \infty$, and for every $r > 0$ such that $D(x, r)$ does not contain X, the inequality

$$\text{area}(X \cap D(x, r)) \geq \tau \, \text{area}(D(x, r))$$

holds. A packing $P = (P_v : v \in V)$ is *fat* if there is some $\tau > 0$ such that each P_v is τ-fat.

For example, any smooth disk is τ-fat for some $\tau > 0$, and it is not hard to see that K-quasi-disks are $\tau(K)$-fat [26].

We can now state:

1.2. Theorem. *Let $G = (V, E)$ be a disk triangulation graph, and for each $v \in V$ let $Q_v \subset \mathbb{C}$ be a smooth compact topological disk. Suppose that there is some $\tau > 0$ such that each Q_v is τ-fat. Let $D \subset \mathbb{C}$ be a simply connected domain, and suppose that $D \neq \mathbb{C}$ (resp. $D = \mathbb{C}$) if G is VEL hyperbolic (resp. VEL parabolic). Then there is a packing $P = (P_v : v \in V)$ in D, which is locally finite in D, whose contacts graph is G, and such that P_v is homothetic to Q_v for each $v \in V$.*[4]

Conversely, suppose that $P = (P_v : v \in V)$ is a fat packing in $\hat{\mathbb{C}}$ of smooth disks whose nerve is G. Then G is VEL parabolic if and only if $\hat{\mathbb{C}}$-carr(P) consists of a single point.

[4] "P_v is homothetic to Q_v" means that there are $a_v > 0$ and $b_v \in \mathbb{C}$ such that $P_v = a_v Q_v + b_v$.

From [24] we know that given a finite planar graph $G^* = (V^*, E^*)$ and a smooth disk $Q_v^* \subset \mathbb{C}$ for each $v \in V^*$, there is a packing $P^* = (P_v^*: v \in V^*)$, with $G(P^*) = G^*$ and P_v^* homothetic to Q_v^* for each $v \in V^*$. This constitutes the "finite case" for the existence part in Theorem 1.2. The basic innovation here is the control one gets on carr(P). The situation where the Q_v are disks, G is VEL hyperbolic, and D is an arbitrary simply connected proper subdomain of $\mathbb{C}$ seems interesting in itself.

Although the equivalence of CP parabolicity to VEL parabolicity gives a complete characterization for disk triangulation graphs, it is quite natural to ask for other criteria. It has been shown by Beardon and Stephenson [5] that if every vertex in G has degree greater than 7, then G is CP hyperbolic, while if every vertex has degree at most 6, G is CP parabolic. We show that if finitely many vertices in G have valence greater than 6, then G is CP parabolic, while if the lower average valence (see Section 10 for the definition) in G is greater than 6, G is CP hyperbolic.

From Rodin and Sullivan's proof of the length–area lemma [20], it follows that if $\gamma_1, \gamma_2, \ldots$ is a sequence of nested simple closed paths in G and $\sum_j 1/|\gamma_j| = \infty$, then G is not CP hyperbolic. This can be seen as a criterion for CP parabolicity. In Section 9 we present a criterion of CP hyperbolicity based on a perimetric inequality in G. There will also be a somewhat restricted converse to this criterion, which is in the spirit of Rodin and Sullivan's length–area lemma.

The interested reader may wish to consult Soardi's paper [27], which studies problems related to those discussed here.

2. Discrete Extremal Length

In this section, we define discrete extremal length. Later, a brief discussion of the history of these definitions appears. We have chosen to start with an abstract notion, and then specialize to more geometric situations.

Combinatorial Extremal Length. Let Γ be a nonempty collection of nonempty subsets of some set X. A (discrete) *metric* on X is a function $m: X \to [0, \infty)$. The *area* of m is just the square of the L^2 norm of m:

$$\text{area}(m) = \|m\|^2 = \sum_{x \in X} m(x)^2.$$

The collection of all metrics m on X with $0 < \text{area}(m) < \infty$ is denoted $\mathcal{M}(X)$. Given a set $A \subset X$, we define the *length* of A in the metric m to be

$$L_m(A) = \sum_{x \in A} m(x).$$

This is also called the m-length of A. If Γ is a collection of subsets of X, we define its m-length to be the least m-length of a set in Γ:

$$L_m(\Gamma) = \inf_{A \in \Gamma} L_m(A).$$

Finally, the *extremal length* of Γ is defined as

$$\mathrm{EL}(\Gamma) = \sup\left\{ \frac{L_m(\Gamma)^2}{\mathrm{area}(m)} : m \in \mathcal{M}(X) \right\}.$$

This is a number in $[0, \infty]$. Note that the ratio $L_m(\Gamma)^2/\mathrm{area}(m)$ does not change if we multiply m by a positive constant. Also note that $\mathrm{EL}(\Gamma)$ does not depend on X; that is, the value of $\mathrm{EL}(\Gamma)$ does not change if we replace X with any other set that contains every $A \in \Gamma$.

The verification of the following simple monotonicity property of extremal length is left to the reader.

2.1. Monotonicity Property. *If each* $\gamma \in \Gamma$ *contains some* $\gamma' \in \Gamma'$, *then* $\mathrm{EL}(\Gamma) \geq \mathrm{EL}(\Gamma')$.

At least when X is finite, a geometric interpretation can be given to $\mathrm{EL}(\Gamma)$. Consider the Euclidean space $\mathbb{R}^X$ of all functions $f\colon X \to \mathbb{R}$. For each subset γ of X, let $\chi_\gamma \in \mathbb{R}^X$ be defined by $\chi_\gamma(x) = 1$ for $x \in \gamma$ and $\chi_\gamma(x) = 0$ otherwise. Now let Γ be, as before, a collection of subsets of X.

2.2. Theorem (Geometric Description of Extremal Length). *Let* m_0 *be the point of least norm in the convex hull of* $\{\chi_\gamma\colon \gamma \in \Gamma\}$. *Then*

$$\mathrm{EL}(\Gamma) = \frac{L_{m_0}(\Gamma)^2}{\mathrm{area}(m_0)} = \|m_0\|^2.$$

We do not use this theorem. The simple proof is left to the reader.

Extremal Length in Graphs. In the following, $G = (V, E)$ is a locally finite connected graph. It will always be a simple graph; that is, each edge has two distinct vertices, and there is at most one edge joining any two vertices.

A *path* γ in G is a finite or infinite sequence $(v_0, v_1, \ldots)$ of vertices such that $[v_i, v_{i+1}] \in E$ for every $i = 0, 1, \ldots$. The edges and vertices of γ are denoted by $E(\gamma) = \{[v_i, v_{i+1}]\colon i = 0, 1, \ldots\}$ and $V(\gamma) = \{v_0, v_1, \ldots\}$, respectively. Likewise, for Γ a set of paths in G, we set $V(\Gamma) = \{V(\gamma)\colon \gamma \in \Gamma\}$ and $E(\Gamma) = \{E(\gamma)\colon \gamma \in \Gamma\}$. A set $A \subset V$ of vertices is said to be *connected* if, for every $v, w \in A$, there is a path γ in G from v to w with $V(\gamma) \subset A$. (We allow trivial paths, paths that contain only one vertex.)

Given subsets $A, B \subset V$, we let $\Gamma(A, B) = \Gamma_G(A, B)$ denote the set of all paths in G with initial point in A and terminal point in B. We let $\Gamma_V(A, B)$ (resp. $\Gamma_E(A, B)$) denote the sets of vertices (resp. edges) of such paths:

$$\Gamma_V(A, B) = \{V(\gamma)\colon \gamma \in \Gamma(A, B)\},$$

$$\Gamma_E(A, B) = \{E(\gamma)\colon \gamma \in \Gamma(A, B)\}.$$

A function $m: V \to [0,\infty)$ is called a *v-metric* on G, and a function $m: E \to [0,\infty)$ is called an *e-metric*. When m is a v-metric (resp. an e-metric) we use $L_m(\gamma)$ as a shorthand for $L_m(V(\gamma))$ (resp. $L_m(E(\gamma))$).

The *vertex extremal length* VEL and *edge extremal length* EEL between A and B are defined by

$$\mathrm{VEL} = \mathrm{VEL}_G(A,B) = \mathrm{EL}(\Gamma_V(A,B)),$$

$$\mathrm{EEL} = \mathrm{EEL}_G(A,B) = \mathrm{EL}(\Gamma_E(A,B)).$$

To make the definition of $\mathrm{VEL}(A,B)$ more explicit, we have

$$\mathrm{VEL}(A,B) = \sup_m \inf_\gamma \frac{L_m(\gamma)^2}{\mathrm{area}(m)} = \sup_m \inf_\gamma \frac{\left(\sum_{v\in V(\gamma)} m(v)\right)^2}{\sum_{v\in V} m(v)^2}.$$

Here m runs over $\mathcal{M}(V)$ and γ runs over $\Gamma_G(A,B)$.

These definitions give two discrete analogs for the classical notion of extremal length. (Reference [17] is a good introduction to continuous extremal length.) As we will see below, both are useful. The edge extremal length was introduced by Duffin, who showed in [12] that $\mathrm{EEL}(A,B)$ is equal to the electrical resistance between A and B, if each edge in G is considered to be a resistor with unit resistance. The vertex extremal length was introduced by Cannon [8]. Cannon's motivation was to obtain criteria for deciding when a group can be made to act conformally on the Riemann sphere $\hat{\mathbb{C}}$. Later it was discovered [9], [25] that extremal metrics of vertex extremal length (that is, metrics realizing the supremum in the definition of the extremal length) give square tilings of rectangles with prescribed contacts.

An infinite path γ in G is *transient* if it contains infinitely many distinct vertices. The set of transient paths in G that have an initial point in A is denoted by $\Gamma(A,\infty)$. The edge and vertex extremal length from A to ∞ are defined as

$$\mathrm{EEL}(A,\infty) = \mathrm{EL}(\{E(\gamma): \gamma \in \Gamma(A,\infty)\}),$$

$$\mathrm{VEL}(A,\infty) = \mathrm{EL}(\{V(\gamma): \gamma \in \Gamma(A,\infty)\}).$$

Of course, this makes sense only for infinite G.

For a v-metric or e-metric m, we let $d_m(A,B)$ (resp. $d_m(A,\infty)$) denote the distance from A to B (resp. to ∞) in the metric m; that is,

$$d_m(A,B) = L_m(\Gamma(A,B)) = \inf\{L_m(\gamma): \gamma \in \Gamma(A,B)\},$$

$$d_m(A,\infty) = L_m(\Gamma(A,\infty)) = \inf\{L_m(\gamma): \gamma \in \Gamma(A,\infty)\}.$$

An infinite graph G is *VEL parabolic* if $\mathrm{VEL}(\{v\},\infty) = \infty$ for some $v \in V$. Otherwise, G is *VEL hyperbolic*. Similarly, G is *EEL parabolic* if $\mathrm{EEL}(\{v\},\infty) = \infty$ for some $v \in V$, and is *EEL hyperbolic*, otherwise.

2.3. Remark. If $\mathrm{VEL}(\{v\}, \infty) = \infty$, then a finite area v-metric $m \in \mathcal{M}(V)$ exists such that $d_m(\{v\}, \infty) = \infty$. To see this, just take $m(v) = \sum_{j=1}^{\infty} m_j(v)$, where the metrics m_j satisfy $A(m_j) < 2^{-j}$ and $L_{m_j}(\Gamma(\{v\}, \infty)) = 1$.

2.4. Exercise. Try to determine $\mathrm{VEL}(A, B), \mathrm{EEL}(A, B)$, and whether G is EEL or VEL parabolic for examples of your choice.

2.5. Exercise. Let G be an infinite connected graph, and let $A \subset V$ be finite and nonempty. Show that $\mathrm{EEL}(A, \infty) = \infty$ iff G is EEL parabolic, and that $\mathrm{VEL}(A, \infty) = \infty$ iff G is VEL parabolic.

While the VEL type (whether parabolic or hyperbolic) is more relevant to packings, the EEL type is closely related to random walks and electricity. We do not introduce the terminology of electrical networks here, but remark that a graph G is electrically parabolic if the electric resistance to infinity in the graph is infinite. (See [11].)

The following theorem is known.

2.6. Theorem. *Let $G = (V, E)$ be a locally finite connected graph. The following are equivalent:*

(1) *G is EEL parabolic.*
(2) *G is electrically parabolic.*
(3) *The simple random walk on G is recurrent.*

The equivalence of (1) and (2) is essentially contained in [12], while the equivalence of (2) and (3) is given in [11]. Also see Section 4 of [29] regarding Theorem 2.6 and further equivalent properties.

In Section 8 we see that VEL and EEL parabolicity are closely related.

3. The Packing Type and Vertex Extremal Length

3.1. Type Characterization Theorem. *Let $P = (P_v: v \in V)$ be a fat packing of (compact connected) sets in the Riemann sphere $\hat{\mathbb{C}}$, and let $G = (V, E)$ be the contacts graph of P. Assume that G is locally finite and connected.*

(1) *If P is locally finite in $\hat{\mathbb{C}} - \{p\}$, where p is some point in $\hat{\mathbb{C}}$, then G is VEL parabolic.*
(2) *Conversely, suppose that each P_v is a smooth disk and that G is a disk triangulation graph, which is VEL parabolic. Then P is locally finite in $\hat{\mathbb{C}} - \{p\}$ for some point $p \in \hat{\mathbb{C}}$.*

The following results about fat sets prove useful.

3.2. Observation. *Let F be a τ-fat set, $\tau > 0$. Then*

$$\operatorname{area}(D(z,3r) \cap F) \geq \pi\tau \operatorname{diameter}(D(z,r) \cap F)^2$$

holds for every $z \in \mathbb{C}$, $r > 0$.

Proof. Let $x, y \in D(z,r) \cap F$. It is clear that $D(x, |y-x|) \subset D(z,3r)$. By the τ-fatness of F, we then have

$$\operatorname{area}(D(z,3r) \cap F) \geq \operatorname{area}(D(x,|y-x|) \cap F) \geq \pi\tau|y-x|^2.$$

The observation follows. □

The following lemma appears in [26].

3.3. Lemma. *There is a positive function τ^*: $(0,\infty) \to (0,\infty)$ such that for every $\tau > 0$, for every τ-fat set $A \subset \hat{\mathbb{C}}$, and for every Möbius transformation φ: $\hat{\mathbb{C}} \to \hat{\mathbb{C}}$ the set $\varphi(A)$ is $\tau^*(\tau)$-fat.*

A central ingredient in the proof of 3.1 is the following lemma, which will also be useful later.

3.4. Lemma. *Let $P = (P_v: v \in V)$ be a fat packing in $\hat{\mathbb{C}}$. Let $G = (V, E)$ denote the contacts graph of P, and assume that G is locally finite. Suppose that z is an accumulation point of the packing P that does not belong to $\bigcup_{v \in V} P_v$. Let K be a compact set in $\hat{\mathbb{C}}$ that does not contain z. For every $A \subset \hat{\mathbb{C}}$ let $V(A)$ denote the set of vertices $v \in V$ such that P_v intersects A. Then*

$$\sup\{\operatorname{VEL}_G(V(K), V(W)): W \text{ is open and } z \in W\} = \infty.$$

In the following, $C(z,r) = \partial D(z,r)$ denotes the circle with center z and radius r.

Proof. Let $\tau > 0$ be such that all the sets P_v are τ-fat. Suppose first that $z \neq \infty$. We now establish that a neighborhood of z is disjoint from $\bigcup_{v \in V(K)} P_v$. Let $R > 0$ be smaller than the distance from z to K, and let V' be the set of $v \in V$ such that P_v intersects both circles $C(z,R)$ and $C(z,R/2)$. Since the distance from $C(z,R)$ to $C(z,R/2)$ is $R/2$, for each $v \in V(C(z,R)) \cap V(C(z,R/2))$, we have $\operatorname{diameter}(D(z,R) \cap P_v) > R/2$. Therefore, Observation 3.2 shows that $\operatorname{area}(D(z,3R) \cap P_v) \geq \pi\tau R^2/4$. In particular, we see that $V(C(z,R)) \cap V(C(z,R/2))$ is finite. This implies that there is an $r_1 \in (0, R/2)$ such that $V(D(z,r_1))$ is disjoint from $V(C(z,R)) \cap V(C(z,R/2))$. Then it follows that $D(z,r_1)$ is disjoint from $\bigcup_{v \in V(K)} P_v$.

We define inductively a sequence $r_1 > r_2 > \cdots$ of positive numbers. The first number in this sequence, r_1, has been defined already. Suppose that $n > 1$, and that $r_1, \ldots, r_{n-1}$ have been defined. Let $r_n \in (0, r_{n-1}/2)$ be sufficiently small so that $V(C(z,r_n)) \cap V(C(z,r_{n-1}/2)) = \varnothing$. The argument above shows that such an r_n exists.

For each n let A_n be the closed annulus bounded by $C(z, r_n)$ and $C(z, r_n/2)$. Define a v-metric m on G by setting

$$m(v) = \sum_{n=1}^{\infty} \frac{\text{diameter}(P_v \cap A_n)}{nr_n},$$

for each $v \in V$. By the construction of the sequence r_n, at most one term in this sum is nonzero. Using this and Observation 3.2, we get an estimate for the area of m, as follows:

$$\begin{aligned}
\text{area}(m) &= \sum_{v \in V} \left(\sum_{n=1}^{\infty} \frac{\text{diameter}(P_v \cap A_n)}{nr_n} \right)^2 \\
&= \sum_{n=1}^{\infty} \sum_{v \in V} \frac{\text{diameter}(P_v \cap A_n)^2}{n^2 r_n^2} \\
&\le \sum_{n=1}^{\infty} \sum_{v \in V} \frac{\text{diameter}(P_v \cap D(z, r_n))^2}{n^2 r_n^2} \\
&\le \sum_{n=1}^{\infty} \sum_{v \in V} \frac{\pi^{-1}\tau^{-1} \text{ area}(P_v \cap D(z, 3r_n))}{n^2 r_n^2} \\
&\le \pi^{-1}\tau^{-1} \sum_{n=1}^{\infty} \frac{\text{area}(D(z, 3r_n))}{n^2 r_n^2} \\
&= 9\tau^{-1} \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty.
\end{aligned}$$

Fix a positive integer N, and consider some path γ from $V(K)$ to $V(D(z, r_N))$. For each integer $n \in [1, N)$ the union $\bigcup_{v \in \gamma} P_v$ is a connected set that intersects the two circles $C(z, r_n)$ and $C(z, r_n/2)$ forming the boundary of A_n. Therefore, for such n, $\Sigma_{v \in \gamma} \text{diameter}(P_v \cap A_n) \ge r_n/2$. This then implies $L_m(\gamma) \ge \frac{1}{2} \Sigma_{n=1}^{N-1} 1/n$, which tends to infinity as $N \to \infty$. Since $\text{area}(m) < \infty$, we get $\text{VEL}(V(K), V(D(z, r_N))) \to \infty$, as $N \to \infty$, which proves the lemma in case $z \neq \infty$.

It is easy to modify the above argument to deal with the case $z = \infty$. The numbers $r_1, r_2, \ldots$ must satisfy in this case $D(0, r_1) \supset \bigcup \{P_v : v \in V(K)\}$, $r_{n+1} > 2r_n$, and $V(C(0, 2r_n)) \cap V(C(0, r_{n+1})) = \emptyset$. The annulus A_n is defined as the annulus whose boundary is $C(0, r_n) \cup C(0, 2r_n)$. The rest of the proof remains essentially the same. Alternatively, using Lemma 3.3, the case $z = \infty$ can be reduced to the case $z = 0$. □

With this lemma, the proof of the first part of 3.1 is easy.

Proof of 3.1(1). Suppose that P is locally finite in $\hat{\mathbb{C}} - \{p\}$. Pick some $v_0 \in V$. Applying Lemma 3.4 with $K = P_{v_0}$, $z = p$, we see that G is VEL parabolic. □

4. Some Topological Lemmas

In this section we gather a few elementary topological lemmas, which will be needed below. The reader is advised to skip the proofs, and perhaps return to them later.

The following two lemmas will enable us to infer topological information of a packing from the combinatorics of the contacts graph.

4.1. Neighbors Separation Lemma. *Let $P = (P_v: v \in V)$ be a packing of smooth disks in $\hat{\mathbb{C}}$, and suppose that the contacts graph $G = (V, E)$ of P is a disk triangulation graph. Let $v_0 \in V$ be some vertex, and let $N \subset V - \{v_0\}$ be the set of neighbors of v_0. Then there is a Jordan curve $\gamma \subset \bigcup_{v \in N} P_v - P_{v_0}$ that separates P_{v_0} from $\bigcup_{v \in V-(N \cup \{v_0\})} P_v$ in $\hat{\mathbb{C}}$.*[5]

The same conclusion holds if G is the (finite) 1-skeleton of a triangulation of a closed disk that has v_0 and all its neighbors as interior vertices.

Proof. Note that the assumptions that each P_v is smooth imply that the intersection of any three sets in P is empty. Let f be an embedding of G in $\hat{\mathbb{C}}$ such that the image of any edge $[v_1, v_2]$ is contained in $P_{v_1} \cup P_{v_2}$ and is disjoint from all the other sets in the packing. (To make an explicit construction of such an embedding, for each $v \in V$ let $h_v: \overline{U} \to P_v$ be a homeomorphism from the closed unit disk $\overline{U} \subset \mathbb{C}$ onto P_v, and for every edge $[v_1, v_2] \in E$ let p_{v_1, v_2} be some point in the intersection $P_{v_1} \cap P_{v_2}$. We may then take $f([v_1, v_2]) = \{h_{v_1}(th_{v_1}^{-1}(p_{v_1, v_2})): 0 \le t \le 1\} \cup \{h_{v_2}(th_{v_2}^{-1}(p_{v_1, v_2})): 0 \le t \le 1\}$.)

In the following we think of G as the 1-skeleton of a triangulation T of an (open or closed) disk. Let $[v_1, v_2, v_3]$ be any triangle in the interior of T. For $j = 1, 2, 3$, let V_j be the neighbors of v_j in G. Clearly, for each $j = 1, 2, 3$ the set $V_j' = V_j - \{v_1, v_2, v_3\}$ is connected as a set of vertices. Since any two of the sets V_1', V_2', V_3' intersect, the union $V' = V_1' \cup V_2' \cup V_3'$ is connected. Since G is connected and any path from a vertex $v \in V - \{v_1, v_2, v_3\}$ to a vertex in $\{v_1, v_2, v_3\}$ must intersect V', it follows that any two vertices in $G - \{v_1, v_2, v_3\}$ can be connected by a path in $G - \{v_1, v_2, v_3\}$. If $(u_1, u_2, u_3, \ldots, u_n)$ is a path in $G - \{v_1, v_2, v_3\}$, then the path $\bigcup_{j=1}^{n-1} f([u_j, u_{j+1}])$ is disjoint from $P_{v_1} \cup P_{v_2} \cup P_{v_3}$, and intersects both P_{u_1} and P_{u_n}. We therefore conclude that every $P_v, v \in V - \{v_1, v_2, v_3\}$, is contained in the same connected component of $\hat{\mathbb{C}} - (f([v_1, v_2]) \cup f([v_2, v_3]) \cup f([v_3, v_1]))$. The set $f([v_1, v_2]) \cup f([v_2, v_3]) \cup f([v_3, v_1])$ is a simple closed curve; we let D_{v_1, v_2, v_3} denote the component of $\hat{\mathbb{C}} - (f([v_1, v_2]) \cup f([v_2, v_3]) \cup f(v_3, v_1]))$ that is disjoint from $\bigcup_{v \in V - \{v_1, v_2, v_3\}} P_v$. Now consider two distinct triangles $[v_1, v_2, v_3], [w_1, w_2, w_3]$ in T. The intersection of the two Jordan curves $\partial D_{v_1, v_2, v_3}, \partial D_{w_1, w_2, w_3}$ is either empty, or consists of a single point, or a single arc. Therefore, the sets $D_{v_1, v_2, v_3}, D_{w_1, w_2, w_3}$ are either disjoint, or one is contained in the other. Suppose, without loss of generality, that $w_1 \notin \{v_1, v_2, v_3\}$. Then $\partial D_{w_1, w_2, w_3}$ intersects P_{w_1}, which is disjoint from the closure of D_{v_1, v_2, v_3}. We conclude that D_{w_1, w_2, w_3} is not contained in D_{v_1, v_2, v_3}. Similarly, D_{v_1, v_2, v_3} is not contained in D_{w_1, w_2, w_3}. Hence $D_{w_1, w_2, w_3} \cap D_{v_1, v_2, v_3} = \emptyset$.

[5] A Jordan curve is a simple closed curve.

Let $n_0, n_2, \ldots, n_{k-1}$ be the neighbors of v_0 in circular order around v_0, and let γ be the Jordan curve $\gamma = \bigcup_{j=0}^{k-1} f([n_j, n_{j+1}])$, where we take $n_k = n_0$. The curve γ is contained in $\bigcup_{v \in N} P_v$ and is disjoint from P_{v_0}. We say that two distinct triangles $[v_1, v_2, v_3], [w_1, w_2, w_3]$ in T *neighbor* if they share an edge. If $[v_1, v_2, v_3]$ is a triangle of T that does not contain v_0 but neighbors with a triangle containing v_0, say with $[v_0, n_j, n_{j+1}]$, then D_{v_1, v_2, v_3} and $D_{v_0, n_j, n_{j+1}}$ lie on opposite sides of the arc $f([n_j, n_{j+1}])$. Consequently, D_{v_1, v_2, v_3} is not in the same connected component of $\hat{\mathbb{C}} - \gamma$ as P_{v_0}. If $[v_1, v_2, v_3]$ and $[w_1, w_2, w_3]$ are two neighboring triangles that do not contain v_0, then it is clear that D_{v_1, v_2, v_3} and D_{w_1, w_2, w_3} are in the same connected component of $\hat{\mathbb{C}} - \gamma$. Hence it easily follows that for every triangle $[v_1, v_2, v_3]$ that does not contain v_0 the set D_{v_1, v_2, v_3} is disjoint from the connected component of $\hat{\mathbb{C}} - \gamma$ that contains P_{v_0}. This implies that γ separates P_{v_0} from $\bigcup_{v \in V - (N \cup \{v_0\})} P_v$, and the lemma follows since $\gamma \subset \bigcup_{v \in N} P_v - P_{v_0}$. □

4.2. Corollary. *Let G be a disk triangulation graph, and let P be a packing of smooth disks in $\hat{\mathbb{C}}$ with $G(P) = G$. Let Z be the set of accumulation points of P. Then there is a connected component D of $\hat{\mathbb{C}} - Z$ that contains P, P is locally finite in D, and D is a topological disk.*

This D is called the *carrier* of P, $D = \mathrm{carr}(P)$.
The verification of Corollary 4.2 is left to the reader.

4.3. Lemma. *Let $P = (P_v : v \in V)$ and $G = (V, E)$ be as in Lemma* 4.1, *and let $v \in V$, $C \subset V - \{u\}$. Suppose that C is finite and u is contained in a finite component of $G - C$. Then $\bigcup_{v \in C} P_v$ separates P_u from the set of accumulation points of P.*

Proof. Let V_0 be the set of vertices that are contained in the same connected component of $G - C$ as u is, and let $K \subset \hat{\mathbb{C}} - \bigcup_{v \in C} P_v$ be a connected set that intersects P_u. For $w \in V$, let $N(w) \subset V - \{w\}$ denote the neighbors of w in G. From Lemma 4.1 we know that for each $w \in V_0$ there is a Jordan curve $\gamma_w \subset \bigcup_{v \in N(w)} P_v - P_w$ that separates P_w from $\bigcup_{v \in V - (N(w) \cup \{w\})} P_v$. Let Q_w denote the component of $\hat{\mathbb{C}} - \gamma_w$ that contains P_w, and let $Q = \bigcup_{v \in V_0} Q_v$. Suppose that $p \in K \cap \partial Q_w$, where $w \in V_0$. Then $p \in K \cap \gamma_w$. Since K is disjoint from $\bigcup_{v \in C} P_v$ and $\gamma_w \subset \bigcup_{v \in N(w)} P_v$, we conclude that $p \in Q_{w'}$ with $w' \in V_0$. Thus $\partial Q_w \cap K \subset Q$, for every $w \in V_0$. Since V_0 is finite, we have $\partial Q \subset \bigcup_{v \in V_0} \partial Q_v$. The above implies that $\partial Q \cap K \subset Q$, and because Q is open, $\partial Q \cap K = \emptyset$. Hence $Q \cap K$ is a relatively open and relatively closed subset of K. As $Q \cap K \neq \emptyset$ and K is connected, we conclude that $K \subset Q$. Because each Q_v intersects finitely many of the sets in the packing P, the lemma follows. □

4.4. Connected Cut Lemma. *Let $G = (V, E)$ be the* 1*-skeleton of a triangulation T of a simply connected surface S. Let $A, B \subset V$ be two disjoint connected sets of vertices. Suppose that $X \subset V$ intersects every path joining A and B. Then there is a connected subset of X that intersects every such path.*

The lemma is surely known, though we have not been able to locate a reference. Since the proof of Alexander's lemma in [19] can be modified to establish 4.4, we do not include a proof here.

5. Duality

The following theorem appears in [25].

5.1. Duality Theorem. *Let $G = (V, E)$ be a finite connected graph, and let A, B be two nonempty subsets of V. Let $\Gamma = \Gamma(A, B)$ denote the set of all paths from A to B in G. Let Γ^* denote the collection of all sets $C \subset V$ with the property that each $\gamma \in \Gamma$ intersects C. Then*

$$\mathrm{EL}(\Gamma^*) = \mathrm{EL}(\Gamma)^{-1}.$$

A related duality theorem can be found in [9].

We need only the inequality $\mathrm{EL}(\Gamma^*) \le \mathrm{EL}(\Gamma)^{-1}$, but in the following slightly more general setting, where the graph is infinite.

5.2. Proposition. *Let $G = (V, E)$ be a connected graph, possibly infinite, let $A, B \subset V$ be two nonempty subsets. Let Γ be either $V(\Gamma(A, B))$ or $V(\Gamma(A, \infty))$. Denote by Γ^* the collection of all subsets $\gamma^* \subset V$ such that γ^* intersects every $\gamma \in \Gamma$. Then*

$$\mathrm{EL}(\Gamma^*)\,\mathrm{EL}(\Gamma) \le 1.$$

Proof. If $\mathrm{EL}(\Gamma) = 0$, there is nothing to prove. So assume that $m \in \mathcal{M}(V)$ satisfies $L_m(\gamma) \ge L > 0$, for some L and every $\gamma \in \Gamma$. For $v \in V$ let the *height* of v be defined as

$$h(v) = \inf\{L_m(\gamma)\colon \gamma \text{ is a path from } A \text{ to } v\}.$$

For $t \in \mathbb{R}$, let V_t denote the set of vertices $v \in V$ such that $h(v) - m(v) \le t \le h(v)$. Since the m-length of every path in Γ is at least L, it is easy to see that $V_t \in \Gamma^*$ for $t \in [0, L)$.

Now let $m^* \in \mathcal{M}(V)$, and set $L^* = L_{m^*}(\Gamma^*) = \inf\{L_{m^*}(\gamma^*)\colon \gamma^* \in \Gamma^*\}$. Since $V_t \in \Gamma^*$ for $t \in [0, L)$, we have

$$L^* L \le \int_0^L L_{m^*}(V_t)\,dt = \int_0^L \sum_{v \in V_t} m^*(v)\,dt.$$

For any $v \in V$, the set of t such that $v \in V_t$ is an interval of length $m(v)$. Therefore, the above inequality yields

$$L^* L \le \sum_{v \in V} m^*(v) m(v) \le \|m^*\|\,\|m\|.$$

This gives

$$\frac{L^{*2}}{\text{area}(m^*)}\,\frac{L^2}{\text{area}(m)} \leq 1,$$

which proves the proposition. □

Suppose now that Γ is some finite nonempty collection of finite nonempty subsets of some set X. Let Γ^* denote the collection of all subsets of X that intersect each $\gamma \in \Gamma$. It is not difficult to see that $\text{EL}(\Gamma)\,\text{EL}(\Gamma^*) \geq 1$. (Consider the geometric interpretation, Theorem 2.2, of combinatorial extremal length.) The example $\Gamma = \{\{1,2\}, \{2,3\}, \{3,1\}\}$, $\Gamma^* = \Gamma$, shows that the inequality may be strict. Hence duality fails in the purely combinatorial setting.

6. CP Hyperbolic Implies VEL Hyperbolic

Proof of 3.1 (continued). It remains to prove the second part of the theorem. We now adopt the assumptions of 3.1(2). Let Z be the set of accumulation points of P. Our immediate goal is to verify that Z is connected. Let $V_1 \subset V_2 \subset \cdots$ be a sequence of finite subsets of V such that $V = \bigcup_n V_n$. For each n, let Q_n be the set of vertices in the infinite connected component of $G - V_n$, and let $\hat{Q}_n$ denote the closure of $\bigcup_{v \in Q_n} P_v$. Clearly, we have $\hat{Q}_1 \supset \hat{Q}_2 \supset \cdots$, and each set $\hat{Q}_n$ is compact and connected. Note that $Z = \bigcap_n \hat{Q}_n$. Since a nested intersection of compact connected sets is connected, it follows that Z is connected.

Let $u \in V$ be some vertex. Normalizing with a Möbius transformation, we assume that $\{z \in \mathbb{C}: |z| \geq 1\} \cup \{\infty\}$ is contained in P_u. Lemma 3.3 shows that this does not involve any loss of generality.

Let m be the v-metric on G defined by

$$m(v) = \begin{cases} \text{diameter}(P_v) & \text{for } v \neq u, \\ 0 & \text{for } v = u. \end{cases}$$

Let $\tau > 0$ be such that each P_v is τ-fat. Since $P_v \subset D(0,1)$ for $v \neq u$, we have

$$\begin{aligned} \text{area}(m) &= \sum_{v \in V - \{u\}} \text{diameter}(P_v)^2 \\ &\leq \pi^{-1}\tau^{-1} \sum_{v \in V - \{u\}} \text{area}(P_v) \\ &\leq \pi^{-1}\tau^{-1}\text{area}(D(0,1)) < \infty. \end{aligned}$$

Let C be any finite subset of $V - \{u\}$ such that u is disjoint from the infinite component of $G - C$. From Lemma 4.3 it follows that the union $\bigcup_{v \in C} P_v$ separates P_u from Z. This clearly implies that

$$\sum_{v \in C} m(v) \geq \text{diameter}(Z). \tag{6.1}$$

Since G is VEL parabolic and $A(m) < \infty$, Proposition 5.2 implies that

$$\inf_S \sum_{v \in S} m(v) = 0,$$

where the infimum runs over all sets $S \subset V$ such that u is not in an infinite component of $G - S$. However, every such S contains a finite $C \subset S$ such that u is not in the infinite component of $G - C$ (for example, the neighbors of the component of $G - S$ containing u). Therefore, (6.1) shows that diameter$(Z) = 0$, as required. □

We easily get the following generalization of 3.1(2).

6.1. Theorem. *Let $\tau > 0$, let $P = (P_v: v \in V)$ be a packing of τ-fat sets in the Riemann sphere $\hat{\mathbb{C}}$, and let $G = (V, E)$ be the contacts graph of P. Assume that G is connected, locally finite, and VEL parabolic. Also suppose that for each $u \in V$ there is a Jordan curve $\gamma \subset \bigcup_{v \in N(u)} P_v - P_u$ that separates P_u from $\bigcup_{v \in V - (N(u) \cup \{u\})} P_v$ in $\hat{\mathbb{C}}$. Here $N(u)$ denotes the set of neighbors of u. Then the set of accumulation points of P has zero length. If G has one end this set consists of a single point.*

We recall that a graph $G = (V, E)$ has one end iff $G - K$ has one infinite component for every finite $K \subset V$.

For example, if P is a locally finite tiling of a domain $\Omega \subset \mathbb{C}$ by compact squares, and if the contacts graph is connected, then $\partial\Omega$ has zero length. In this case the contacts graph does not have to be planar, since four squares may meet at a point.

Proof. The proof is essentially the same as for 3.1(2). Note that the assumptions that the sets P_v are smooth and that G is a disk triangulation graph were used only in the proof of Lemma 4.1. Since we are assuming the conclusion of this lemma here, these assumptions are not needed. If G has one end, then it is easy to see that the set of accumulation points of P is connected. The proof of 3.1(2) shows that in our present case for every $\varepsilon > 0$ the set of accumulation points of P is covered by a finite collection of sets such that the sum of their diameters is less than ε. This clearly implies Theorem 6.1. □

7. Uniformizations of Packings

7.1. Uniformization Theorem. *Let $G = (V, E)$ be a disk triangulation graph, for each $v \in V$ let $Q_v \subset \mathbb{C}$ be a smooth disk, and let $D \subset \mathbb{C}$ be a simply connected domain.*

Assume that there is a $\tau > 0$ such that Q_v is τ-fat for each $v \in V$. Also suppose that $D \neq \mathbb{C}$ (resp. $D = \mathbb{C}$) if G is VEL hyperbolic (resp. parabolic). Then there is a packing $P = (P_v: v \in V)$ with $\mathrm{carr}(P) = D$ whose contacts graph is G, and such that P_v is homothetic to Q_v for each $v \in V$.

We note that the continuous analogue of this theorem appears in [26]. The proof is also similar.

Proof. Let T be the triangulation of a disk that has G as its 1-skeleton. Let $T^1 \subset T^2 \subset T^3 \subset \cdots$ be an exhaustion of T. By this we mean that $T = \bigcup_j T^j$, and each T^j is a finite triangulation of a disk (with boundary). It is easy to see that such an exhaustion exists. We also require, without loss of generality, that T^1 has some interior vertex, say v_0. For each $j = 1, 2, \ldots$, let $G^j = (V^j, E^j)$ denote the 1-skeleton of T^j.

Suppose, without loss of generality, that $0 \in D$ and 0 is in the interior of Q_{v_0}. Let D^j be a sequence of smooth Jordan domains[6] in $\mathbb{C}$ such that $0 \in D^1 \subset D^2 \subset \cdots$ and $D = \bigcup_j D^j$. From the packing theorems of [24] we know that for each $j = 1, 2, \ldots$ there is a packing $P^j = (P_v^j: v \in V^j)$ in the closure of D^j, such that each P_v^j is homothetic to Q_v, the sets P_v^j are tangent to ∂D^j when v is a boundary vertex of T^j, and $P_{v_0}^j$ has the form $t_j Q_{v_0}$ for some $t_j > 0$. Let $\{j(k)\}$ be a subsequence of $\{1, 2, \ldots\}$ such that the Hausdorff limit

$$\tilde{P}_v = \lim_{k \to \infty} \frac{1}{\mathrm{diameter}\left(P_{v_0}^{j(k)}\right)} P_v^{j(k)} \tag{7.1}$$

exists for every $v \in V$. The Hausdorff limit is taken in $\hat{\mathbb{C}}$; that is, *a priori* we must allow for the possibility that ∞ is contained in some $\tilde{P}_v$.

We show now that the sets $\tilde{P}_v$ do not degenerate to single points and do not contain ∞. The set $\tilde{P}_{v_0}$ certainly is OK, since it contains 0, has diameter 1, and is homothetic to Q_{v_0}, by construction. Let u be any neighbor of v_0. Since $\tilde{P}_u$ is a Hausdorff limit of sets homothetic to Q_u, which is smooth, $\tilde{P}_u$ is either homothetic to Q_u, or is a single point, or a half-plane, or $\tilde{P}_u = \hat{\mathbb{C}}$. The last case is clearly impossible, since the interior of $\tilde{P}_u$ does not intersect $\tilde{P}_{v_0}$. It is also clear that $\tilde{P}_u$ intersects $\tilde{P}_{v_0}$ but does not intersect its interior.

Let $u_1, u_2, \ldots, u_n$ be the neighbors of v_0, in circular order. For every j such that v_0 and all its neighbors are in the interior of T^j, the set $P_{u_1}^j \cup \cdots \cup P_{u_n}^j$ contains a Jordan curve that separates $P_{v_0}^j$ from ∞. (This follows from Lemma 4.1.) Therefore, for at least two neighbors u of v_0 the sets $\tilde{P}_u$ contain more than a single point. Suppose, for example, that $\tilde{P}_{u_1}$ is a single point p, and that $\tilde{P}_{u_n}$ is not a single point. Let m be the largest number in $\{1, 2, \ldots, n\}$ such that $\tilde{P}_{u_r} = \{p\}$ for each $r < m$ in $\{1, 2, \ldots, n\}$. Since at least two $\tilde{P}_{u_i}$ do not degenerate to points, $m < n$. It is clear that each $\tilde{P}_{u_i}$ intersects $\tilde{P}_{u_{i+1}}$ and that $\tilde{P}_{u_1}$ intersects $\tilde{P}_{u_n}$. Therefore, the three smooth sets $\tilde{P}_{v_0}, \tilde{P}_{u_n}, \tilde{P}_{u_m}$ contain the point p. This implies that the interiors of two

[6] A smooth Jordan domain is the interior of a smooth disk.

of these sets must intersect, which is clearly impossible. Thus we conclude that none of the sets $\tilde{P}_{u_i}$ consists of a single point, and that the ratios diameter$(P^j_{v_0})$/ diameter$(P^j_{u_i})$ are bounded from above. (The reader may wish to compare the above argument with the Ring Lemma of [20].)

Is it possible that $\tilde{P}_{u_1}$ is a half-plane? To see that it is not, consider a Hausdorff limit of the packings $(h_j(P^j_v)\colon v \in V^j)$, where h_j is the homothety that takes $P^j_{u_1}$ to Q_{u_1}. The same argument as above, but with the roles of u_1 and v_0 switched, shows then that the ratios diameter$(P^j_{u_1})$/diameter$(P^j_{v_0})$ are bounded from above. Similarly, for every edge $[u,w]$ the ratio diameter(P^j_u)/diameter(P^j_w) is bounded independently of j. Since G is connected, this also holds when $u,w \in V$ are not neighbors. Therefore, each set $\tilde{P}_v$ is not a half-plane, nor a point, and thus is homothetic to Q_v.

If G is VEL parabolic, then part (2) of 3.1 implies that $\tilde{P}$ is locally finite in $\hat{\mathbb{C}} - \{p\}$ for some $p \in \hat{\mathbb{C}}$. It is easy to see that $p = \infty$, and thus $\tilde{P}$ is locally finite in $\mathbb{C}$. This completes the proof in the case that G is VEL parabolic.

Now suppose that G is VEL hyperbolic. The set $P^j_{v_0}$ is contained in $D \subsetneq \mathbb{C}$ and has the form $t_j Q_{v_0}$, $t_j > 0$. Since 0 is an interior point of Q_{v_0}, this implies that the sequence t_j is bounded from above, and hence diameter$(P^j_{v_0})$ is bounded from above. By passing to a subsequence of $j(k)$, if necessary, assume that $t = \lim_{k\to\infty}$ diameter$(P^j_{v_0}) \in [0,\infty)$ exists. We have established above that for any $v,w \in V$ the ratios diameter(P^j_v)/diameter(P^j_w) remain bounded as $j \to \infty$. Consider the Hausdorff limits

$$P_v = \lim_{k\to\infty} P^{j(k)}_v. \tag{7.2}$$

If $t = 0$, then, because G is connected, it follows that $P_v = \{0\}$ for each v, and in particular the limits (7.2) exist. If $t > 0$, then comparing with (7.1), we conclude again that these limits exist, and that each P_v is homothetic to Q_v.

We now prove that each P_v is contained in D. Consider some vertex $v \in V$, and let $N(v)$ denote the neighbors of v. By Lemma 4.1, for each j sufficiently large (so that $N(v)$ is contained in the interior of T^j) there is a Jordan curve in $\bigcup_{u\in N(v)} P^j_u - P^j_v$ that separates P^j_v from ∂D^j. Assuming that $t > 0$, since for any fixed u the sets P^j_u vary within a compact collection of homotheties of Q_u, the above implies that the distance from P^j_v to ∂D^j is bounded from below independently of j. Therefore, $P_v \subset D$. The same conclusion is true, of course, if $t = 0$, because then $P_v = \{0\}$. So we have established that the packing P is contained in D.

Clearly, the interiors of the sets P_v are disjoint, and $P_v \cap P_w \neq \emptyset$ whenever $[v,w] \in E$. Therefore, the proof will be complete once we show that the packing $P = (P_v\colon v \in V)$ is locally finite in D. (This will also rule out the possibility $t = 0$, $P_v = \{0\}$.) That is actually the most significant part of the proof. It turns out that the packing $\tilde{P}$ is useful to proving this property of P.

Let F be some compact connected subset of D that contains 0. We prove that F intersects finitely many sets in P, and this shows that P is locally finite in D. Let F' be any compact connected subset of D that contains F in its interior. Let z be some accumulation point of $\tilde{P}$. From Lemma 4.3 we know that $\tilde{P}$ is disjoint from its accumulation points. By Theorem 3.1, z is not the only accumulation point of $\tilde{P}$.

Therefore, there is a compact connected set K that intersects $\tilde{P}_{v_0}$, contains an accumulation point of $\tilde{P}$, and is disjoint from z. In the following, for a set $X \subset \hat{\mathbb{C}}$, let $\tilde{V}(X)$ denote the set of $v \in V$ such that $\tilde{P}_v$ intersects X. Since K is connected and contains an accumulation point of $\tilde{P}$, it is clear that each component of $\tilde{V}(K)$ is infinite. (This follows from Lemma 4.3.)

We let ε be a small positive number whose value is determined below. By Lemma 3.4, there is some open set $W = W(z, K, \varepsilon)$ containing z such that

$$\mathrm{VEL}_G(\tilde{V}(K), \tilde{V}(W)) > \frac{1}{\varepsilon}.$$

Without loss of generality, we assume that W is connected. Then every component of $\tilde{V}(W)$ is infinite.

Assume for the moment that D has finite area. We show that if ε is chosen sufficiently small, then P_v is disjoint from F for every $v \in \tilde{V}(W)$. Let C be some component of $\tilde{V}(W)$. Let j be sufficiently large so that C intersects V^j, and let C^j be any component of $C \cap V^j$. Since every component of $\tilde{V}(W)$ is infinite, C is infinite, and therefore C^j must contain boundary vertices of T^j. Let H^j be the component of $\tilde{V}(K) \cap V^j$ that contains v_0. The above argument tells us that H^j contains boundary vertices of T^j. Let Γ^{*j} denote the family of all subsets of V^j that intersect every path in $\Gamma_{G^j}(H^j, C^j)$. Proposition 5.2 now implies that

$$\mathrm{EL}(\Gamma^{*j}) \leq \mathrm{VEL}_{G^j}(H^j, C^j)^{-1} \leq \mathrm{VEL}_G(\tilde{V}(K), \tilde{V}(W))^{-1} < \varepsilon. \tag{7.3}$$

Consider the v-metric $m = m_j$ that assigns to each $v \in V^j$ the diameter of P_v^j. By the τ-fatness of the sets Q_v, we have

$$\mathrm{area}(P_v^j) \geq \tau\pi m(v)^2.$$

This implies

$$\mathrm{area}(m) \leq \tau^{-1}\pi^{-1}\,\mathrm{area}(D) < \infty.$$

Inequality (7.3) now implies that there is some $\gamma^* \in \Gamma^{*j}$ such that

$$L_m(\gamma^*) < \sqrt{\frac{\varepsilon\,\mathrm{area}(D)}{\tau\pi}}.$$

We now choose ε to be sufficiently small so that the right-hand side of the above inequality is smaller than $d(F', \partial D)/2$, half the distance from F' to ∂D. So we have

$$L_m(\gamma^*) < \frac{d(F', \partial D)}{2}. \tag{7.4}$$

Since the sets C^j and H^j are connected, Lemma 4.4 implies that there is a $\gamma_1^* \in \Gamma^{*j}$ that is connected and is contained in γ^*. Let $Y^j = \bigcup_{v \in \gamma_1^*} P_v^j$. Because Y^j is connected, we may estimate its diameter as follows:

$$\text{diameter}(Y^j) \leq \sum_{v \in \gamma^*} \text{diameter}(P_v^j) = L_m(\gamma^*) < \frac{d(F', \partial D)}{2}.$$

Recall that C^j and H^j both contain boundary vertices of T^j. Since γ_1^* separates H^j from C^j, it too must contain boundary vertices. This implies that Y^j intersects ∂D^j. We now assume that j is sufficiently large so that $d(F', \partial D^j) > d(F', \partial D)/2$. Since $\text{diameter}(Y^j) < d(F', \partial D)/2 < d(F', \partial D^j)$, and Y^j intersects ∂D^j, it is clear that Y^j does not intersect F'. Since γ_1^* separates H^j from C^j in G^j, it is clear that $Y^j \cup \partial D^j$ separates $\bigcup_{v \in C^j} P_v^j$ from $P_{v_0}^j$. Since $Y^j \cup \partial D^j$ does not intersect F', which is connected and intersects $P_{v_0}^j$, it follows that $\bigcup_{v \in C^j} P_v^j$ is disjoint from F'. Recall that C^j is any component of $C \cap V^j$, and C is any component of $\tilde{V}(W)$. Therefore, for any $v \in \tilde{V}(W)$, if j is sufficiently large so that $v \in V^j$ and $d(F', \partial D^j) > d(F', \partial D)/2$, then $P_v^j \cap F' = \emptyset$. Taking limits, it follows that P_v is disjoint from the interior of F', which contains F, and so $P_v \cap F = \emptyset$.

We summarize our conclusions as follows. For every accumulation point z of $\tilde{P}$ there is a neighborhood W_z of z such that $P_v \cap F = \emptyset$ for every $v \in \tilde{V}(W_z)$. Let W^* be the union of all W_z, over all accumulation points z of $\tilde{P}$. Then W^* is an open set that contains the accumulation points of $\tilde{P}$. Consequently, $V - \tilde{V}(W^*)$ is finite. Since $F \cap P_v = \emptyset$ for all $v \in \tilde{V}(W^*)$, only finitely many sets in the packing P intersect F. Hence P is locally finite in D.

This concludes the proof in the case that D has finite area. When D has infinite area, the same proof is valid when the spherical metric of $\hat{\mathbb{C}}$ is used in place of the flat metric of $\mathbb{C}$. The only fact to note is that there is some τ_1, which depends only on τ, such that the spherical area of P_v^j is at least τ_1 times the square of the spherical diameter of P_v^j. This follows easily from Lemma 3.3. Thus the proof is complete. □

We can now prove:

7.2. Theorem. *A disk triangulation graph is CP parabolic iff it is VEL parabolic. A disk triangulation graph is CP hyperbolic iff it is VEL hyperbolic.*

Proof of Theorems 1.2 and 7.2. These follow immediately from 3.1 and 7.1. □

8. VEL Parabolicity, EEL Parabolicity, and Recurrence

We have seen that the VEL type of a disk triangulation graph is equal to its CP type, and now we establish the connection between VEL and EEL type. Through Theorem 2.6, this relates the CP type of graph to well-studied notions.

8.1. Theorem. *Let $G = (V, E)$ be a locally finite graph. If G is EEL parabolic, then it is also VEL parabolic. Conversely, if G has bounded valence and is VEL parabolic, then it is EEL parabolic.*

Proof. Suppose that M is an e-metric on G. Define a v-metric m on G by

$$m(v) = \max\{M([v,u]) : [v,u] \in E\}.$$

If γ is any transient path in G, then a simple diagonalization argument shows that there is a path $\gamma' = (v_1, v_2, \ldots)$ with distinct vertices that are all in γ. Thus

$$L_m(\gamma) \geq L_m(\gamma') = \sum_{j=1}^{\infty} m(v_j) \geq \sum_{j=1}^{\infty} M([v_j, v_{j+1}]) = L_M(\gamma'). \tag{8.1}$$

For each $v \in V$ let $e(v)$ denote an edge e of G containing v that maximizes $M(e)$ among such edges. Clearly, each $e \in E$ is equal to $e(v)$ for at most two vertices v. Using this, we get

$$\begin{aligned} \operatorname{area}(m) &= \sum_{v \in V} m(v)^2 = \sum_{v \in V} M(e(v))^2 \\ &\leq 2 \sum_{e \in E} M(e)^2 = 2 \operatorname{area}(M). \end{aligned} \tag{8.2}$$

Together with (8.1) this establishes that an EEL parabolic graph is VEL parabolic.

To prove the opposite implication, assume that there is a global bound k on the valence of any vertex $v \in V$. Let m be some v-metric on G. Define an e-metric M by $M([u,v]) = \max(m(u), m(v))$. It is easy to establish that for any path γ we have $L_M(\gamma) \geq L_m(\gamma)$. Moreover, since each vertex is incident with at most k edges, a calculation similar to (8.2) gives $\operatorname{area}(M) \leq k \operatorname{area}(m)$. These inequalities show that a bounded valence VEL parabolic graph is EEL parabolic, and the proof of the theorem is complete. □

8.2. Theorem. *There is a disk triangulation graph which is CP and VEL parabolic, but EEL hyperbolic and transient.*

This shows that the bounded valence requirement in the second part of Theorem 8.1 is essential.

Proof. Let T be a triangulation of an open disk. It is not hard to see that by adding vertices and edges inside the triangular faces of T a new triangulation T^* whose 1-skeleton G^* is transient can be obtained. On the other hand, G^* is VEL parabolic iff the 1-skeleton of T is VEL parabolic. The details are left to the reader. □

Proof (of 1.1). Follows immediately from Theorems 8.1 and 7.2. □

9. Perimetric Inequalities and the Type

9.1. Theorem. *Let $G = (V, E)$ be a locally finite, infinite, connected graph, let W_0 be a finite nonempty set of vertices of G, and let $g: [0, \infty) \to (0, \infty)$ be some nondecreasing function.*

(1) *If G is VEL parabolic and satisfies the perimetric inequality*

$$|\partial W| \geq g(|W|) \tag{9.1}$$

for every finite connected vertex set $W \supset W_0$, then

$$\sum_{n=1}^{\infty} \frac{1}{g(n)^2} = \infty. \tag{9.2}$$

Here ∂W denotes the set of vertices that are not in W but neighbor with some vertex in W, and $|A|$ denotes the cardinality of a set A.

(2) *If (9.2) holds, and*

$$|\partial W_k| \leq g(|W_k|) \tag{9.3}$$

is valid for every $k = 0, 1, 2, \ldots$, where W_k is defined inductively by $W_{k+1} = W_k \cup \partial W_k$, then G is VEL parabolic.

We remark that part (1) and its proof are analogous to a criterion of Grigor'yan for the hyperbolicity of a Riemannian manifold [14]. Part (2) can be viewed as a generalization of the Rodin–Sullivan length–area lemma [20].

Proof. Assume that G is VEL parabolic. Let m be some v-metric on G with $\text{area}(m) < \infty$ and $d_m(W_0, \infty) = \infty$. (See Remark 2.3.) We also assume, without loss of generality, that $m(v) > 0$ for each $v \in V$. For each $v \in V$, let I_v be the interval

$$I_v = [d_m(W_0, v) - m(v), d_m(W_0, v)].$$

For $h \in [0, \infty)$ set

$$V_h = \{v \in V: h \in I_v\},$$

$$w(h) = \sum \{m(v): v \in V_h\},$$

$$Y_h = \{v \in V: I_v \subset [0, h]\},$$

$$n(h) = |Y_h|.$$

It is easy to see that $V_h = \partial Y_h$, and therefore

$$|V_h| \geq g(n(h)). \tag{9.4}$$

It turns out that $n(h)$ is not convenient to work with, since it is not smooth enough. We therefore define

$$s_v(h) = \frac{\text{length}(I_v \cap [0,h])}{m(v)} \qquad \text{for} \quad v \in V,$$

$$s(h) = \sum_{v \in V} s_v(h).$$

Note that $s_v(h)$ is equal to 0 for $h \leq \min I_v$, $s_v(h) = 1$ for $h \geq \max I_v$, and s_v is linear in I_v. Since $d_m(W_0, \infty) = \infty$, it follows that for every $h \in [0,\infty)$ there are finitely many v such that I_v intersects $[0,h]$. Therefore $s(h)$ is a piecewise linear function. It should be thought of as a smoothed version of $n(h)$.

Now set

$$f(x) = \min\left(g\left(\frac{x}{2}\right), \frac{x}{2}\right). \tag{9.5}$$

Let $h \in [0,\infty)$. If $|V_h| \geq s(h)/2$, then

$$|V_h| \geq f(s(h)). \tag{9.6}$$

Suppose that $|V_h| < s(h)/2$. Then we have $n(h) = |Y_h| \geq s(h) - |V_h| > s(h)/2$. Consequently

$$|V_h| \geq g(n(h)) \geq g\left(\frac{s(h)}{2}\right) \geq f(s(h)),$$

and we conclude that (9.6) holds in any case.

We are now ready to do some real work. At points h where $s(h)$ is differentiable, we have

$$\frac{ds}{dh}(h) = \sum_{v \in V_h} s_v'(h) = \sum_{v \in V_h} \frac{1}{m(v)}.$$

Therefore, using the Cauchy–Schwarz inequality (or the inequality between the arithmetic and harmonic means) and (9.6), we get

$$\frac{ds}{dh} \geq \frac{|V_h|^2}{\sum \{m(v): v \in V_h\}} \geq \frac{f(s(h))^2}{w(h)}.$$

This gives

$$\frac{ds}{f(s)^2} \geq \frac{dh}{w(h)}.$$

Integrating for h in some interval $[a, b]$, $0 < a < b < \infty$, and using Cauchy–Schwarz again, we get

$$\int_{s(a)}^{s(b)} \frac{ds}{f(s)^2} \geq \int_a^b \frac{dh}{w(h)} \geq \frac{(b-a)^2}{\int_a^b w(h)\, dh}. \tag{9.7}$$

Note that

$$\int_0^\infty w(h)\, dh = \int_0^\infty \sum_{v \in V_h} m(v)\, dh = \sum_{v \in V} \int_{I_v} m(v)\, dh = \sum_{v \in V} m(v)^2 = \text{area}(m) < \infty.$$

Therefore, letting $b \to \infty$ in (9.7), we get

$$\int_{s(a)}^\infty \frac{ds}{f(s)^2} = \infty.$$

Since

$$\frac{1}{f(s)^2} = \max\left(\frac{1}{g(s/2)^2}, \frac{4}{s^2}\right) \leq \frac{1}{g(s/2)^2} + \frac{4}{s^2},$$

we find that $\int_0^\infty g(s)^{-2}\, ds = \infty$, which implies (9.2). This proves part (1).

To establish part (2), now set $n_k = |W_k|$, and assume that (9.3) holds. Let N be some positive integer, and define a v-metric m on G by

$$m(v) = \begin{cases} g(n_k)^{-1} & \text{for } v \in \partial W_k, \quad k \leq N, \\ 0 & \text{otherwise.} \end{cases}$$

We have

$$d_m(W_0, \infty) \geq d_m(W_0, \partial W_N) \geq \sum_{k=0}^N g(n_k)^{-1}.$$

On the other hand,

$$\text{area}(m) \leq \sum_{k=0}^N \frac{|\partial W_k|}{g(n_k)^2} \leq \sum_{k=0}^N g(n_k)^{-1}.$$

Since the above are valid for each N, we get

$$\text{VEL}(W_0, \infty) \geq \sum_{k=0}^\infty g(n_k)^{-1}. \tag{9.8}$$

Note that

$$n_{k+1} = |W_{k+1}| = |W_k \cup \partial W_k| \leq |W_k| + |\partial W_k| \leq n_k + g(n_k).$$

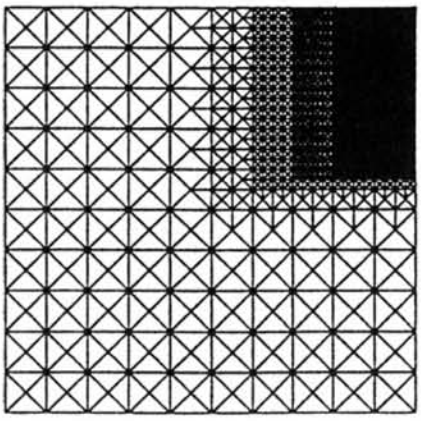


Fig. 9.1. A parabolic graph with exponential growth.

Using this and the monotonicity of g, we obtain

$$\frac{1}{g(n_k)} \geq \frac{1}{n_{k+1}-n_k} \sum_{n=n_k}^{n_{k+1}-1} \frac{1}{g(n)} \geq \sum_{n=n_k}^{n_{k+1}-1} \frac{1}{g(n_k)} \frac{1}{g(n)} \geq \sum_{n=n_k}^{n_{k+1}-1} \frac{1}{g(n)^2}.$$

This implies

$$\sum_{k=0}^{\infty} g(n_k)^{-1} \geq \sum_{n=n_0}^{\infty} \frac{1}{g(n)^2}.$$

Now part (2) follows from (9.8). □

There is a certain asymmetry in the two parts of Theorem 9.1. While the first part examines the relation between the size of the boundary of W and the size of W for every finite connected vertex set $W \supset W_0$, the second part does this only for the sets W_k. This difference is essential; that is, part (1) fails if (9.1) is only assumed for the sets W_k. Figure 9.1 gives a disk triangulation graph, which is essentially equivalent to a graph constructed by Soardi [27], with the following properties:

(1) G is VEL parabolic.
(2) The maximum degree in G is 8.
(3) $|\partial W_k| \geq C|W_k|$ for some $C > 0$ and all $k = 0, 1, \ldots$. The constant C does depend on the choice of W_0, but not on k.

Property (3) clearly implies that G has exponential growth; i.e., $|W_k| \geq (1 + C)^k$.

10. Determining the Type from the Valences

A natural question is, can the type of a graph be determined from the valences of its vertices. Suppose for a moment that G is a disk triangulation graph. It is known [5] that if all the vertices of G have degree greater than 6, then G is not CP parabolic,

and if all the vertices of G have degree at most 6, then G is CP parabolic. The following two theorems generalize these results.

10.1. Theorem. *Let G be the 1-skeleton of an infinite triangulation of a surface, and suppose that at most finitely many vertices in G have degree greater than 6. Then G is VEL parabolic, EEL parabolic, and recurrent.*

Due to Theorem 7.2 this implies that a disk triangulation graph with finitely many vertices of degree greater than 6 is CP parabolic.

Proof. The method of proof is to show that the rate of growth of G is too slow for G to be hyperbolic.

Given a set $W \subset V$, we let ∂W denote the set of $v \in V - W$ that neighbor with some vertex in W, and let $\tilde{\partial} W$ be the set of vertices in ∂W that neighbor with some vertex in $V - (W \cup \partial W)$. If K is finite, then ∂K is a finite set of vertices, and K is disjoint from the infinite components of $V - \partial K$.

Let $K_0 \subset V$ be a finite nonempty set of vertices that contains all the vertices in V of degree greater than 6. We define the sequence $K_1, K_2, \ldots$ inductively by setting

$$K_{n+1} = K_n \cup \partial K_n .$$

Note that each K_n is finite, and each vertex in $\tilde{\partial} K_n$ has at least one neighbor in K_n and at least one neighbor outside of K_{n+1}. Let C_n denote the set of vertices in ∂K_n that have precisely one neighbor in K_n, and let $D_n = \partial K_n - C_n$. Consider some $v \in \tilde{\partial} K_n$. Let $u_1, \ldots, u_m$ be its neighbors, in circular order, and set $u_0 = u_m$. Since v has a neighbor in K_n, we assume without loss of generality that it is $u_0 = u_m$. Let $j \in \{1, \ldots, m-1\}$ be such that $u_j \notin K_{n+1}$. Since $v \in \tilde{\partial} K_n$, such a j exists. Let a be the least index in $\{1, \ldots, j\}$ such that $u_a \notin K_{n+1}$, and let b be the maximal index in $\{j, \ldots, m-1\}$ such that $u_b \notin K_{n+1}$. Since u_{a-1} neighbors with v and with u_a, and $u_{a-1} \in K_{n+1}$, it is clear that $u_a \in D_{n+1}$ and $u_{a-1} \in \partial K_n$. Similarly, $u_{b+1} \in \partial K_n$ and $u_b \in D_{n+1}$. By construction, $\{u_a, \ldots, u_b\}$ contains all the neighbors of v in ∂K_{n+1}.

Suppose for a moment that $v \in D_n \cap \tilde{\partial} K_n$. We know that v has at most six neighbors. Of these, at least two are in K_n, and at least two are in ∂K_n, namely, u_{a-1} and u_{b+1}. If $a \neq b$, then v has at least two neighbors in D_{n+1}, namely, u_a, u_b. As $2 + 2 + 2 = 6$, we see that when $a \neq b$, v has precisely two neighbors in D_{n+1} and no neighbors in C_{n+1}. If $a = b$, then $u_a = u_b$ is the only neighbor of v in ∂K_{n+1}, and this neighbor is in D_{n+1}. We conclude that a vertex in D_n neighbors with at most two vertices in D_{n+1} and with no vertices in C_{n+1}.

The above reasoning also shows that a vertex in C_n neighbors with at most three vertices in ∂K_{n+1}, of which at most one is in C_{n+1}. One conclusion that we get is

$$|C_{n+1}| \leq |C_n|. \tag{10.1}$$

Let m_{n+1} denote the number of edges between K_{n+1} and D_{n+1}. On the one hand, $m_{n+1} \geq 2|D_{n+1}|$, because every vertex in D_{n+1} has at least two neighbors in K_{n+1}.

On the other hand, the only vertices in K_{n+1} that neighbor with D_{n+1} are in $D_n \cup C_n$, the vertices in D_n have at most two neighbors in D_{n+1}, and the vertices in C_n have at most three neighbors in D_{n+1}. Therefore,

$$2|D_{n+1}| \le m_{n+1} \le 2|D_n| + 3|C_n|,$$

which gives

$$|D_{n+1}| \le |D_n| + \frac{3|C_n|}{2}. \tag{10.2}$$

Using induction and inequalities (10.1) and (10.2), we see that

$$|C_n| \le |C_0|, \qquad |D_n| \le |D_0| + \frac{3n|C_0|}{2}.$$

Therefore,

$$|\partial K_n| = |C_n \cup D_n| \le |D_0| + (2n+1)|C_0|. \tag{10.3}$$

Let m be the v-metric on G defined by $m(v) = 1/(n \log n)$ for $v \in \partial K_n$, $n > 1$, and $m(v) = 0$ for $v \notin \bigcup_{n>1} \partial K_n$. Since ∂K_n intersects every transient path meeting K_0, we see that $d_m(K_0, \infty) \ge \sum_{n>1} 1/(n \log n) = \infty$. On the other hand, (10.3) implies that area$(m) < \infty$. Hence G is VEL parabolic. From Theorems 8.1 and 2.6 it follows that G is EEL parabolic and recurrent. □

Let G be a disk triangulation graph. For $v \in V$, let $\deg(v)$ denote the degree of v in G. The average valence of a finite nonempty set of vertices W is just

$$\operatorname{av}(W) = \frac{1}{|W|} \sum_{v \in W} \deg(v).$$

The *lower average valence* of G is defined to be

$$\operatorname{lav}(G) = \sup_{W_0} \inf_{W \supset W_0} \operatorname{av}(W);$$

where W and W_0 are nonempty finite connected sets of vertices. (The authors do not know if this notion appears in the literature.)

10.2. Theorem. *Let G be a locally finite connected planar graph, and suppose that* $\operatorname{lav}(G) > 6$. *Then G is VEL hyperbolic, and therefore EEL hyperbolic and transient.*

Note that the lower average valence of the hexagonal grid is 6.

Beardon and Stephenson [5] have shown that if every vertex of G has degree at least 7, then G is not CP parabolic. The above theorem is a generalization of this result.

Proof. In any finite planar graph G^* with vertex set V^*, the average valence satisfies

$$\mathrm{av}(V^*) < 6. \tag{10.4}$$

This is a well-known fact but for the convenience of the nonexpert readers, we give the proof here. Let n, e, f be the number of vertices, edges, and faces of the graph (which is embedded in the plane). The Euler formula gives $n + f = e + 2$, and the inequality $3f \le 2e$ holds if $f > 1$, since every face must have at least three edges on its boundary, and each edge is on the boundary of at most two faces. From these it follows that $n > e/3$ (actually it is this inequality which we need later). However, $\mathrm{av}(V^*) = 2e/n$, since every edge is counted exactly twice in the sum $\Sigma_{v \in V^*} \deg(v)$. This establishes $\mathrm{av}(V^*) < 6$.

We now return to the infinite graph G. Let W_0 be a finite connected nonempty set of vertices such that $\mathrm{av}(W) > C > 6$ for some constant C and every finite connected set of vertices $W \supset W_0$. Consider such a W, and let G^* be the restriction of G to $W \cup \partial W$; that is, the vertices of G^* are $W \cup \partial W$, and an edge of G appears in G^* iff both its endpoints are in $W \cup \partial W$. Denote by n and e the number of vertices and edges in G^*, respectively. Then, clearly, $2e \ge |W|\,\mathrm{av}(W)$, and therefore, by the previous paragraph,

$$|W| + |\partial W| = |W \cup \partial W| = n > \frac{e}{3} \ge \frac{|W|\mathrm{av}(W)}{6} > \left(\frac{C}{6}\right)|W|.$$

This gives

$$|\partial W| > g(|W|)$$

with $g(x) = (C - 6)x/6$. Now, since $\Sigma_{n=1}^{\infty} g(n)^{-2} < \infty$, part (1) of Theorem 9.1 shows that G must be VEL hyperbolic, and the proof is complete. □

It would be interesting to narrow the wide gap between Theorems 10.1 and 10.2. Suppose, for example, that G is a bounded valence disk triangulation graph and that v_0 is some vertex in G. Let $k_n = \Sigma_v (6 - \deg(v))$, where the sum extends over all vertices v at distance at most n from v_0. Can criteria for the type of G based on the sequence $\{k_n\}$ be given? For example, if k_n is bounded, does it follow that G is VEL parabolic?

Acknowledgments

We thankfully acknowledge fruitful discussions with Peter Doyle and Burt Rodin. We also thank the anonymous referees for helpful comments, and especially for the improvement of the statement and proof of Theorem 9.1.

References

1. E. M. Andreev, On convex polyhedra in Lobačevskiĭ spaces, *Mat. Sb.* (*N.S.*) **81** (**123**) (1970), 445–478; English transl. *Math. USSR-Sb.* **10** (1970), 413–440.

2. E. M. Andreev, On convex polyhedra of finite volume in Lobačevskiĭ space, *Mat. Sb. (N.S.)* **83 (125)** (1970), 256–260; English transl. *Math. USSR-Sb.* **12** (1970), 255–259.
3. A. F. Beardon and K. Stephenson, The uniformization theorem for circle packings, *Indiana Univ. Math. J.* **39** (1990), 1383–1425.
4. A. F. Beardon and K. Stephenson, The Schwarz–Pick lemma for circle packings, *Illinois J. Math.* **141** (1991), 577–606.
5. A. F. Beardon and K. Stephenson, Circle packings in different geometries, *Tôhoku Math. J.* **43** (1991), 27–36.
6. P. L. Bowers, The upper Perron method for labeled complexes with applications to circle packings, *Math. Proc. Cambridge Philos. Soc.* **114** (1993), 321–345.
7. G. R. Brightwell and E. R. Scheinerman, Representations of planar graphs, *SIAM J. Discrete Math.* **6** (1993), 214–229.
8. J. W. Cannon, The combinatorial Riemann mapping theorem, *Acta Math.* **173** (1994), 155–234.
9. J. W. Cannon, W. J. Floyd, and W. R. Parry, Squaring rectangles: the finite Riemann mapping theorem, *The Mathematical Heritage of Wilhelm Magnus—Groups, Geometry and Special Functions*, Contemporary Mathematics Series, American Mathematical Society, Providence, RI, 1994.
10. Y. Colin De Verdiére, Un principe variationnel pour les empilements de cercles, *Invent. Math.* **104** (1991), 655–669.
11. P. G. Doyle and J. L. Snell, *Random Walks and Electric Networks*, The Carus Mathematical Monographs, Vol. 22, Mathematical Association of America, Washington, DC, 1984.
12. R. J. Duffin, The extremal length of a network, *J. Math. Anal. Appl.* **5** (1962), 200–215.
13. B. T. Garrett, Circle packings and polyhedral surfaces, *Discrete Comput. Geom.* **8** (1992), 429–440.
14. A. A. Grigor'yan, On the existence of positive fundamental solutions of the laplace equation on Riemannian manifolds, *Math. USSR-Sb.* **56** (1987), 349–358.
15. Z.-X. He and O. Schramm, Fixed points, Koebe uniformization and circle packings, *Ann. of Math.* **137** (1993), 369–406.
16. P. Koebe, Kontaktprobleme der konformen Abbildung, *Ber. Ver. Sächs. Akad. Wiss. Leipzig Math.-Phys. Klasse* **88** (1936), 141–164.
17. O. Lehto and K. I. Virtanen, *Quasiconformal Mappings in the Plane*, Springer-Verlag, Berlin, 1973.
18. A. Marden and B. Rodin, *On Thurston's Formulation and Proof of Andreev's Theorem*, Lecture Notes in Mathematics, Vol. 1435, Springer-Verlag, Berlin, 1989, pp. 103–115.
19. M. H. A. Newman, *Elements of the Topology of Plane Sets of Points*, 2nd, edn., Cambridge University Press, Cambridge, 1939, Dover, New York, 1992.
20. B. Rodin and D. Sullivan, The convergence of circle packings to the Riemann mapping, *J. Differential Geom.* **26** (1987), 349–360.
21. O. Schramm, Existence and uniqueness of packings with specified combinatorics, *Israel J. Math.* **73** (1991), 321–341.
22. O. Schramm, Rigidity of infinite (circle) packings, *J. Amer. Math. Soc.* **4** (1991), 127–149.
23. O. Schramm, How to cage an egg, *Invent. Math.* **107** (1992), 543–560.
24. O. Schramm, Conformal uniformization and packings, *Israel J. Math.* (to appear).
25. O. Schramm, Square tilings with prescribed combinatorics, *Israel J. Math.* **84** (1993), 97–118.
26. O. Schramm, Transboundary extremal length, *J. Analyse Math.* (to appear).
27. P. M. Soardi, Recurrence and transience of the edge graph of a tiling of the euclidean plane, *Math. Ann.* **287** (1990), 613–626.
28. W. P. Thurston, *The Geometry and Topology of 3-Manifolds*, Princeton University Notes, Princeton University Press, Princeton, NJ, 1982.
29. W. Woess, Random walks on infinite graphs and groups—a survey on selected topics, *Bull. London Math. Soc.* **26** (1994), 1–60.

Received December 9, 1991, and in revised form July 19, 1994.

An electronic version is available at doi: 10.1215/S0012-7094-97-08611-7

Acta Math., 180 (1998), 219–245
© 1998 by Institut Mittag-Leffler. All rights reserved

The C^∞-convergence of hexagonal disk packings to the Riemann map

by

ZHENG-XU HE and ODED SCHRAMM

University of California, San Diego
La Jolla, CA, U.S.A.

The Weizmann Institute of Science
Rehovot, Israel

1. Introduction

Let $\Omega \subsetneq \mathbf{C}$ be a simply-connected domain. The Rodin–Sullivan Theorem states that a sequence of disk packings in the unit disk U converges, in a well-defined sense, to a conformal map from Ω to U. Moreover, it is known that the first and second derivatives converge as well. Here, it is proven that for hexagonal disk packings the convergence is C^∞. This is done by studying Möbius invariants of the disk packings that are discrete analogs of the Schwarzian derivative.

As a consequence, the first n derivatives of the conformal map can be approximated by quantities which depend on the positions of the centers of some $n+1$ consecutive disks in the packing.

Suppose that $P=(P_v : v \in V)$ is an indexed disk packing in the plane. The *nerve* (or *tangency graph*) of P is the graph $G=(E,V)$ on V, the indexing set of P, such that $[v,u] \in E$ if and only if P_v and P_u are tangent. The Disk Packing Theorem [10] says that given a graph G which is the 1-skeleton of a topological triangulation of $\bar{U}=\{z \in \mathbf{C} : |z| \leqslant 1\}$, there is some disk packing in $\bar{U}$ with nerve G, such that the boundary disks (i.e., those disks corresponding to the boundary vertices) are all tangent to the unit circle ∂U. Moreover, the disk packing is unique up to (possibly orientation-reversing) Möbius transformations of $\bar{U}$ (see, for example, [18, Chapter 13], [12], [3] or [4] for other proofs and extensions).

For each $\varepsilon > 0$, let H^ε denote the hexagonal grid with mesh ε. The vertices of H^ε form the hexagonal lattice:

$$V^\varepsilon = \{n\varepsilon + m\omega\varepsilon : (n,m) \in \mathbf{Z} \times \mathbf{Z}\},$$

The first named author was supported by NSF Grant DMS 96-22068, and the second named author by NSF Grant DMS 94-03548.

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_9, © Springer Science+Business Media, LLC 2011

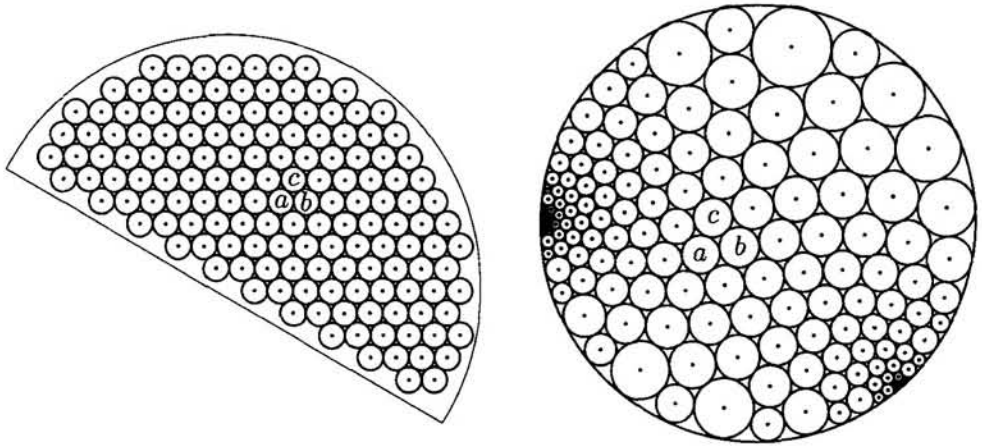


Fig. 1.1. The packings R^ε and P^ε, and the map f^ε. Several points have been marked to aid in grasping the correspondence.

where ω is the cube root of -1,

$$\omega = \exp \tfrac{1}{3}\pi i = \tfrac{1}{2}(\sqrt{3}\,i+1),$$

and an edge connects any two vertices of H^ε at distance ε.

Let Ω be a simply-connected domain in $\mathbf{C}$ with $\Omega \neq \mathbf{C}$. Then there is a subgrid H^ε_Ω of H^ε, equal to the 1-skeleton of a triangulation of a closed topological disk contained in Ω that approximates Ω. Let V^ε_Ω denote the set of vertices in H^ε_Ω. From the Disk Packing Theorem it follows that there is a disk packing $P^\varepsilon = (P^\varepsilon_v : v \in V^\varepsilon_\Omega)$ in $\overline{U}$ whose nerve is H^ε_Ω and with the property that boundary disks are tangent to ∂U. For each $v \in V^\varepsilon_\Omega$, let $f^\varepsilon(v)$ denote the center of the disk P^ε_v (see Figure 1.1). Then the Rodin–Sullivan Theorem [15] tells us that, assuming that the packings P^ε are suitably normalized (by a Möbius transformation, possibly orientation-reversing), the discrete functions $f^\varepsilon : V^\varepsilon_\Omega \to U$ converge "locally uniformly" in Ω to the (similarly normalized) Riemann map $f : \Omega \to U$, when $\varepsilon \to 0$.

There is a natural definition for the discrete partial derivatives of functions defined on the lattice V^ε_Ω. Given discrete functions $g^\varepsilon : V^\varepsilon_\Omega \to \mathbf{C}$, where $\varepsilon \in (0,1)$, we say that g^ε *converge in* $C^\infty(\Omega)$ to a smooth function $g : \Omega \to \mathbf{C}$ as $\varepsilon \to 0$, if g^ε converge locally uniformly in Ω to g, and the discrete partial derivatives of g^ε of any order converge locally uniformly to the corresponding partial derivatives of g. The precise definitions will be given in §2.

For each $v \in V^\varepsilon_\Omega$, let $r^\varepsilon(v)$ denote the radius of P^ε_v. Our main theorem is:

C^∞-CONVERGENCE THEOREM 1.1. *The discrete functions $f^\varepsilon : V^\varepsilon_\Omega \to U$ converge in $C^\infty(\Omega)$ to the Riemann mapping $f : \Omega \to U$. The discrete functions $2r^\varepsilon(v)/\varepsilon$ converge in $C^\infty(\Omega)$ to $|f'|$.*

As a consequence, the derivatives of the conformal map up to order n can be approximated by quantities which depend on the positions of the centers of some $n+1$ consecutive disks in the packing.

W. Thurston [19] constructed the packings P^ε, and conjectured that f^ε converges to the Riemann mapping. The locally uniform convergence (i.e., C^0-convergence) was proved by Rodin and Sullivan [15], using a rigidity theorem for quasi-conformal deformations of Schottky groups and a length-area argument. The C^1-convergence was proved in [6], using an area estimate; and C^2-convergence in [5], using the same area estimate. The result of [5] says that the "intersticial maps" converge in C^2 to f. In [8] it is shown that the method of [6] works well for general disk packings with bounded valence.

Recently, the C^2-convergence of the intersticial maps has been generalized in [9] to disk packings of arbitrary combinatorial pattern, even without the assumption of bounded valence. This is done using methods of discrete extremal length and fixed point index arguments. We intend to write a paper which further simplifies [9], avoids the discrete extremal length methods, and gives estimates for the convergence rates.

Several other interesting results also appeared in [13], [14], [1] and [2]. These works may be combined together to yield an alternative proof of the C^1-convergence. We note that K. Stephenson [17] has also given a proof of the C^0-convergence using Markov chains and electrical networks "with leaks"; and recently, L. Carleson has found a different proof based on Rodin's equation [13], namely,

$$\pi = \sum_{k=1}^{6} \tan^{-1}\left(\sqrt{\frac{r_k r_{k+1}}{R(r_k + r_{k+1} + R)}}\right), \tag{1.1}$$

where r_k, $k \in \mathbf{Z}_6 = \mathbf{Z}/6\mathbf{Z}$, are the radii of six disks which surround a disk of radius R.

Following is a brief description of our method. First some Möbius invariants for finite hexagonal disk packings are defined and their elementary equations are derived. Similar invariants and equations were introduced [16] in the setting of circle patterns with the combinatorics of the square grid. We will then use the Möbius invariants to define (discrete) Schwarzians (or Schwarzian derivatives). Roughly, the Schwarzians are some appropriately scaled measure of the deformation of pairs of neighboring interstices from their standard positions and, as the continuous Schwarzian, they are unchanged if the packing is replaced by a Möbius image. For a regular hexagonal packing (in which all disks have identical size), the Schwarzians are tailored to be 0. The results of [6] will be used to show that the Schwarzians of the packings P^ε are bounded. On the other hand, from the equations satisfied by the Möbius invariants, we will show that the (discrete) Laplacian of the Schwarzians is a polynomial function in ε and the Schwarzians themselves. It is then a consequence of a regularity theorem of discrete elliptic equations

that all discrete derivatives of the Schwarzians, of any given order, are locally bounded. The C^∞-convergence of f^ε and $2r^\varepsilon/\varepsilon$ will then follow.

It is also proved that a linear combination of the discrete Schwarzians of P^ε converges in C^∞ to the Schwarzian derivative of the conformal mapping. Another interesting result is that some natural "contact transformations" defined by the positions of triplets of mutually tangent disks converge in C^∞. In fact, any reasonable, natural discrete function associated to P^ε can be shown to be convergent in $C^\infty(\Omega)$.

It is possible to carry out our proof with an investigation of the radii in place of the Schwarzians, and with Rodin's equation (1.1) taking the place of the equation which the discrete Schwarzians satisfy. But perhaps some of the details become more complicated.

The rest of the paper is organized as follows. §2 introduces some discrete partial differential operators on the lattices V^ε, and defines precisely the meaning of locally uniform convergence and C^∞-convergence. §3 gives the construction of the subgrid H^ε_Ω and fixes a normalization for P^ε, and hence for f^ε and f. §4 introduces the Schwarzians, and derives the equations satisfied by them. These equations will be used in §5 to obtain the formula for the discrete Laplacian of the Schwarzians. A consequence of the formula is that if the Schwarzians are all bounded, then so are their Laplacians. In §6, we will prove that the Schwarzians are uniformly bounded in any compact subset of Ω. In §7 the Lipschitz norm of a discrete derivative of a function is estimated in terms of the L^∞-norms of the function and its discrete Laplacian. At that point, one can conclude that the Schwarzians are uniformly Lipschitz, and therefore convergent in $C^0(\Omega)$. In §8, local uniform bounds on the partial derivatives of any order for the Schwarzians are obtained. In §9, the contact transformations of disk patterns are defined, and it is shown that they converge in $C^\infty(\Omega)$. Then, in §10, the main theorem is obtained as a consequence of the C^∞-convergence of the contact transformations. Finally, disk patterns of more general combinatorics are discussed in §11.

We thank the anonymous referee for valuable suggestions, which lead us to the current formulation of the main theorem.

2. Preliminaries on discrete differential operators and the continuous limit of discrete functions

We will need to introduce some discrete differential operators. For each $\varepsilon>0$, recall that the set V^ε is the set of vertices in the hexagonal grid H^ε, i.e., the points $n\varepsilon+m\omega\varepsilon$, where $\omega=\exp\left(\frac{1}{3}\pi i\right)$, and n, m are integers.

For any $k\in\mathbf{Z}_6$, let $L^\varepsilon_k: V^\varepsilon\to V^\varepsilon$ denote the translation

$$L^\varepsilon_k v = v+\varepsilon\omega^k. \tag{2.1}$$

Let $W \subseteq V^\varepsilon$ be a subset. A vertex $v \in W$ is called an *interior vertex* of W if for each k, $k \in \mathbf{Z}_6$, the neighboring vertex $L_k^\varepsilon v$ is contained in W. Let $W^{(0)} = W$, and for each integer $l \geqslant 1$, let $W^{(l)}$ denote the set of interior vertices of $W^{(l-1)}$.

Given a function $\eta: W \to \mathbf{R}$, the (discrete) directional derivative $\partial_k^\varepsilon \eta: W^{(1)} \to \mathbf{R}$ is defined by

$$\partial_k^\varepsilon \eta(v) = \varepsilon^{-1}(\eta(L_k^\varepsilon v) - \eta(v)) = \varepsilon^{-1}(\eta(v + \varepsilon\omega^k) - \eta(v)). \tag{2.2}$$

Let $L_k^\varepsilon \eta$ denote the function which differs from η by the translation L_k^ε:

$$L_k^\varepsilon \eta(v) = \eta(L_k^\varepsilon v) = \eta(v + \varepsilon\omega^k). \tag{2.3}$$

Then we have

$$\partial_k^\varepsilon = \varepsilon^{-1}[L_k^\varepsilon - I],$$

where $I\eta = \eta$.

The (discrete) Laplacian of a function $\eta: W \to \mathbf{R}$ is a function in $W^{(1)}$ defined by the formula

$$\Delta^\varepsilon \eta(v) = \tfrac{2}{3}\varepsilon^{-2} \sum_{k \in \mathbf{Z}_6} (\eta(L_k^\varepsilon v) - \eta(v)) = \tfrac{2}{3}\varepsilon^{-2} \sum_{k \in \mathbf{Z}_6} (\eta(v + \varepsilon\omega^k) - \eta(v)). \tag{2.4}$$

In other words,

$$\Delta^\varepsilon = \tfrac{2}{3}\varepsilon^{-2} \sum_{k \in \mathbf{Z}_6} (L_k^\varepsilon - I).$$

Note that the Laplacian of x^2 restricted to V^ε is 2. That is the reason for the factor $\frac{2}{3}\varepsilon^{-2}$.

Clearly, the operators I, L_j^ε, ∂_k^ε and Δ^ε commute with each other.

For any function g defined on a subset $W \subseteq V^\varepsilon$, we will use $\|g\|_W$ to denote its $L^\infty(W)$-norm:

$$\|g\|_W = \sup_{v \in W} |g(v)|.$$

For any differentiable function $G: \Omega \to \mathbf{R}$ and any $k \in \mathbf{Z}_6$, let $\partial_k G$ denote the directional derivative

$$\partial_k G(z) = \lim_{t \to 0} \frac{G(z + t\omega^k) - G(z)}{t}. \tag{2.5}$$

Definitions. Let $f: \Omega \to \mathbf{C}^d$ be some function defined in some domain $\Omega \subset \mathbf{C}$. For each $\varepsilon > 0$, let f^ε be some function defined on some set of vertices $V_0^\varepsilon \subset V^\varepsilon$, with values in $\mathbf{C}^d$. Suppose that for each $z \in \Omega$ there are some $\delta_1, \delta_2 > 0$ such that $\{v \in V^\varepsilon : |v - z| < \delta_2\} \subset V_0^\varepsilon$ whenever $\varepsilon \in (0, \delta_1)$.

If for every $z \in \Omega$ and every $\sigma > 0$ there are some $\delta_1, \delta_2 > 0$ such that $|f(z) - f^\varepsilon(v)| < \sigma$, for every $\varepsilon \in (0, \delta_1)$ and every $v \in V^\varepsilon$ with $|v - z| < \delta_2$, then we say that f^ε converges to f, locally uniformly in Ω.

Let $n \in \mathbf{N}$, and suppose that f is C^n-smooth. If for every sequence $k_1, ..., k_j \in \mathbf{Z}_6$ with $j \leqslant n$ we have $\partial^\varepsilon_{k_j} \partial^\varepsilon_{k_{j-1}} ... \partial^\varepsilon_{k_1} f^\varepsilon \to \partial_{k_j} \partial_{k_{j-1}} ... \partial_{k_1} f$ locally uniformly in Ω, then we say that f^ε converges to f in $C^n(\Omega)$. If that holds for all $n \in \mathbf{N}$, then the convergence is C^∞.

The functions f^ε are said to be *uniformly bounded* in $C^n(\Omega)$ provided that for every compact $K \subset \Omega$ there is some constant $C(K, n)$ such that

$$\|\partial^\varepsilon_{k_j} \partial^\varepsilon_{k_{j-1}} ... \partial^\varepsilon_{k_1} f^\varepsilon\|_{K \cap V^\varepsilon} < C(K, n)$$

whenever $j \leqslant n$, and ε is sufficiently small. The functions f^ε are *uniformly bounded* in $C^\infty(\Omega)$, if they are uniformly bounded in $C^n(\Omega)$ for every $n \in \mathbf{N}$.

Following are some simple lemmas about C^∞-convergence of such functions.

LEMMA 2.1. *Let n be a positive integer. Suppose that the functions f^ε are uniformly bounded in $C^n(\Omega)$. Then for every sequence of $\varepsilon \to 0$ there is a $C^{n-1}(\Omega)$-function f and a subsequence of $\varepsilon \to 0$ such that $f^\varepsilon \to f$ in $C^{n-1}(\Omega)$ along that subsequence.*

Proof. Since f^ε are uniformly bounded in $C^1(\Omega)$, the standard proof of the Arzela–Ascoli Theorem shows that there is a continuous function f defined in Ω and a subsequence of $\varepsilon \to 0$ such that $f^\varepsilon \to f$ locally uniformly along that subsequence.

We apply the same argument to the discrete derivatives of the functions f^ε with order at most $n-1$, and conclude that for some subsequence all the discrete derivatives of order at most $n-1$ of the functions f^ε will converge locally uniformly in Ω. It is then easy to verify that $f^\varepsilon \to f$ in $C^{n-1}(\Omega)$. The details are left to the reader. □

LEMMA 2.2. *Suppose that $f^\varepsilon, g^\varepsilon, h^\varepsilon$ converge in $C^\infty(\Omega)$ to functions $f, g, h: \Omega \to \mathbf{C}$, and suppose that $h \neq 0$ in Ω. Then the following convergences are in $C^\infty(\Omega)$:*

(1) $f^\varepsilon + g^\varepsilon \to f + g$,
(2) $f^\varepsilon g^\varepsilon \to fg$,
(3) $1/h^\varepsilon \to 1/h$,
(4) *if* $h^\varepsilon > 0$ *then* $\sqrt{h^\varepsilon} \to \sqrt{h}$,
(5) $|h^\varepsilon| \to |h|$.

Proof. The local uniform convergence is obvious in each of these cases. Hence, by Lemma 2.1, it is sufficient to show that these discrete functions are uniformly bounded in $C^\infty(\Omega)$. Part (1) is easy. Part (2) follows from the identity

$$\partial^\varepsilon_k (f^\varepsilon g^\varepsilon) = (\partial^\varepsilon_k f^\varepsilon) g^\varepsilon + (L^\varepsilon_k f^\varepsilon) \partial^\varepsilon_k g^\varepsilon$$

and induction. Similarly, parts (3) and (4) follow from the identities,

$$\partial_k^\varepsilon(1/h^\varepsilon)=\frac{-\partial_k^\varepsilon h^\varepsilon}{h^\varepsilon L_k^\varepsilon h^\varepsilon},$$

$$\partial_k^\varepsilon\sqrt{h^\varepsilon}=\frac{\partial_k^\varepsilon h^\varepsilon}{\sqrt{L_k^\varepsilon h^\varepsilon}+\sqrt{h^\varepsilon}}.$$

The details are left to the reader. The convergence of $|h^\varepsilon|$ is C^∞, because $|h^\varepsilon|=\sqrt{h^\varepsilon\overline{h^\varepsilon}}$. □

3. The setup

This section will describe the construction of the packings P^ε. The notations and assumptions introduced here will be adopted throughout the following.

Let Ω be a simply-connected domain in $\mathbf{C}$ with $\Omega\neq\mathbf{C}$. Fix two arbitrary distinct points z_0, z_0' in Ω. We now describe a standard construction of a triangulation of a simply-connected subregion of Ω that approximates Ω. For every small $\varepsilon>0$, consider the subset of vertices of $V^\varepsilon\cap\{z\in\mathbf{C}:|z|\leqslant 1/\varepsilon\}$ such that their distance to $\mathbf{C}-\Omega$ is bigger than ε; and let $\widetilde{V}_\Omega^\varepsilon$ be the set of vertices which are either an interior vertex of this subset, or shares an edge with an interior vertex. Let V_Ω^ε be the connected component of $\widetilde{V}_\Omega^\varepsilon$ which contains some vertex with distance at most ε to z_0, and let H_Ω^ε be the subgraph of H^ε spanned by the vertices in V_Ω^ε. Then for small ε, H_Ω^ε is equal to the 1-skeleton of a geometric triangulation (with equilateral triangles of side length ε) of a closed topological disk contained in Ω that approximates Ω.

As we stated in the introduction, it follows from the Disk Packing Theorem that there is a disk packing $P^\varepsilon=(P_v^\varepsilon:v\in V_\Omega^\varepsilon)$ in $\overline{U}$ whose nerve is H_Ω^ε and with the property that boundary disks are tangent to ∂U. We normalize P^ε by a (possibly orientation-reversing) Möbius transformation, so that (1) for a vertex $v_0\in V_\Omega^\varepsilon$ that is closest to z_0 the center of the disk $P_{v_0}^\varepsilon$ is 0, (2) for a vertex $v_0'\in V_\Omega^\varepsilon$ that is closest to z_0' the center of the disk $P_{v_0'}^\varepsilon$ is on the positive real ray, and (3) the six disks $P_{L_k^\varepsilon v}^\varepsilon$, $k=0,1,\ldots,5$, surround P_v^ε in positive circular order. Let $f^\varepsilon: V_\Omega^\varepsilon\to U$ be the function that maps every $v\in V_\Omega^\varepsilon$ to the center of the disk P_v^ε. Then the Rodin–Sullivan Theorem [15] tells us that f^ε converges locally uniformly in Ω to the Riemann map $f:\Omega\to U$ satisfying $f(z_0)=0$ and $f(z_0')>0$.

Let R^ε denote the regular hexagonal disk packing whose disks are centered at the vertices in the hexagonal grid H^ε. The disks of R^ε all have radius $\frac{1}{2}\varepsilon$. Let R_Ω^ε be the subpacking corresponding to the subset of vertices V_Ω^ε.

For any triplet of mutually tangent disks P_1, P_2, P_3 in P^ε or R^ε, there is a disk D whose boundary ∂D passes through the intersection points $P_1\cap P_2$, $P_2\cap P_3$ and $P_3\cap P_1$.

The disk D is called a *dual disk* of the packing. Note that ∂D is orthogonal to each of the circles ∂P_j, $j=1,2,3$.

4. The discrete Schwarzians

Let f be a complex analytic function defined on a domain in the plane $\mathbf{C}$, such that $f'(z)$ never vanishes. The Schwarzian derivative of f is defined by

$$\mathcal{S}f(z)=\left(\frac{f''(z)}{f'(z)}\right)'-\frac{1}{2}\left(\frac{f''(z)}{f'(z)}\right)^2=\frac{f'''(z)}{f'(z)}-\frac{3(f''(z))^2}{2(f'(z))^2}. \tag{4.1}$$

The Schwarzian derivative of f is itself a complex analytic function. It is elementary to check that for a Möbius transformation $T(z)=(az+b)/(cz+d)$ we have $\mathcal{S}(T\circ f)(z)=\mathcal{S}f(z)$. Moreover, $Sf=0$ if and only if f is equal to the restriction of a Möbius transformation. These and some further properties of the Schwarzian can be found in [11, §1], for example.

In the same spirit, we will define the Möbius invariants of hexagonal disk packings, and derive their immediate equations. Analogous invariants and equations were worked out in [16] in a similar way for circle patterns based on the square grid, where applications were found to the study of global properties of immersed patterns in $\widehat{\mathbf{C}}$. Here, we will use Möbius invariants as an intermediate means in the study of the convergence problem; and the Schwarzians will be defined as some suitably scaled measure of deformation of the Möbius invariants from their regular values. As an important step, we will derive a formula for the Laplacian of the Schwarzians in the next section.

Our method will yield the same result if Rodin's equation (1.1), which the radii satisfy, is used instead. Thus, the use of Möbius invariants is not essential. We have chosen to work with the Möbius invariants as they yield relatively easier equations.

For any edge $e=[v,u]$ in H^ε_Ω, we let p_e denote the point of tangency of the two disks P^ε_u, P^ε_v. Let $e=[v,u]$ be some edge in H^ε_Ω, let w_1,w_2 be the two vertices of V^ε that neighbor with both v and u, and suppose that $w_1,w_2\in V^\varepsilon_\Omega$. Let T be a Möbius transformation that sends the tangency point p_e to infinity. Then the two circles $\partial P_v, \partial P_u$ are mapped to lines. It follows that the four points $T(p_{[v,w_1]}), T(p_{[v,w_2]}), T(p_{[u,w_1]}), T(p_{[u,w_2]})$ are at the corners of a rectangle, see Figure 4.1.

Since T is well-determined up to post-composing by a similarity, the aspect ratio of the rectangle,

$$\frac{|T(p_{[v,w_1]})-T(p_{[v,w_2]})|}{|T(p_{[v,w_1]})-T(p_{[u,w_1]})|},$$

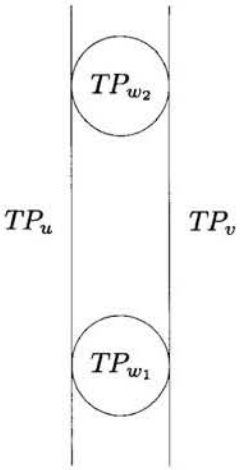


Fig. 4.1. The configuration after applying T.

is independent of the choice of T. It also does not change if we modify P^ε by a Möbius transformation. Set s_e to be this aspect ratio divided by $\sqrt{3}$,

$$s(e)=\frac{|T(p_{[v,w_1]})-T(p_{[v,w_2]})|}{\sqrt{3}\,|T(p_{[v,w_1]})-T(p_{[u,w_1]})|}, \tag{4.2}$$

and let

$$h(e)=\varepsilon^{-2}(s(e)-1) \tag{4.3}$$

be called the (discrete) *Schwarzian derivative*, or *Schwarzian* of P^ε at e. The factor $1/\sqrt{3}$ in (4.2) is justified by the fact that when the disks $P_u, P_v, P_{w_1}, P_{w_2}$ are all the same size, we get $s(e)=1$ and $h(e)=0$. The factor ε^{-2} is reasonable, because of the behavior of the Schwarzian derivative under rescaling, namely, $(\mathcal{S}g)(z)=\varepsilon^2(\mathcal{S}f)(\varepsilon z)$ when $g(z)=f(\varepsilon z)$. (This is also justified by the estimate of §6.)

For a vertex v in V^ε, denote $e_k(v)=[v, L_k^\varepsilon(v)]$. See Figure 4.2. Let $s_k, h_k: (V_\Omega^\varepsilon)^{(1)}\to\mathbf{R}$ be defined by $s_k(v)=s(e_k(v))$ and $h_k(v)=h(e_k(v))$. Clearly, $s_k(v)=s_{k+3}(L_k v)$ and $h_k(v)=h_{k+3}(L_k v)$.

LEMMA 4.1. *Let v be an interior vertex in V_Ω^ε. Then,*

$$s_k(v)+s_{k+2}(v)+s_{k+4}(v)=3s_k(v)s_{k+1}(v)s_{k+2}(v) \tag{4.4}$$

is valid for any $k\in\mathbf{Z}_6$.

Although we will not prove it here, the equations (4.4) are sufficient to guarantee that a positive function s on the edges of H^ε corresponds to an (immersed) hexagonal circle pattern (cf. [16]).

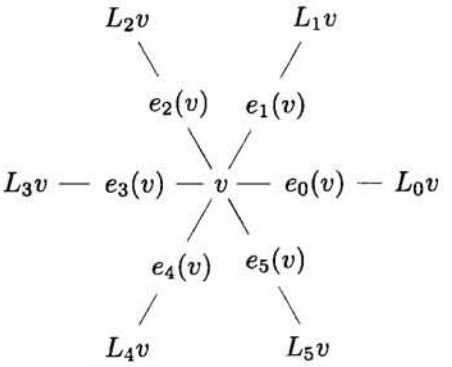


Fig. 4.2. The edges around a vertex.

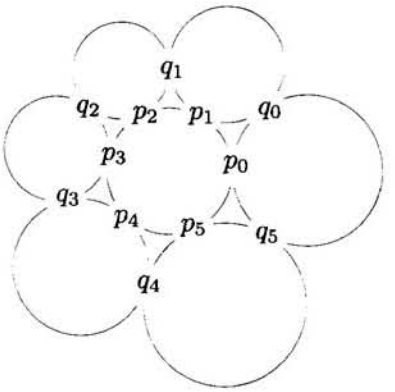


Fig. 4.3. The special points of a flower.

Proof. For any $k \in \mathbf{Z}_6$, let p_k be the point of tangency $P_v^\varepsilon \cap P_{L_k^\varepsilon v}^\varepsilon$, and let q_k be the point in $P_{L_k^\varepsilon v}^\varepsilon \cap P_{L_{k+1}^\varepsilon v}^\varepsilon$. See Figure 4.3. There is no loss of generality, because the packing P^ε may be reflected about a line.

Let $m_k = m_k(v)$ be the orientation-preserving Möbius transformation that takes p_k, p_{k-1}, q_{k-1} to $\infty, 0, 1$, respectively. Then, by the definition of the s_k's,

$$\begin{aligned} m_k(v)(p_{k+1}) &= -\sqrt{3}\, s_k i, \\ m_k(v)(q_k) &= 1 - \sqrt{3}\, s_k i, \\ m_k(v)(p_k) &= \infty. \end{aligned}$$

Consequently, setting $M_k = m_{k+1} \circ m_k^{-1}$, we get

$$\begin{aligned} M_k\big(-\sqrt{3}\, s_k i\big) &= \infty, \\ M_k\big(1 - \sqrt{3}\, s_k i\big) &= 1, \\ M_k(\infty) &= 0. \end{aligned}$$

Therefore,

$$M_k = \begin{pmatrix} 0 & i \\ i & -\sqrt{3}\, s_k \end{pmatrix}, \tag{4.5}$$

where the usual matrix notation for Möbius transformations is used. Note that the composition $M_5 \circ M_4 \circ M_3 \circ M_2 \circ M_1 \circ M_0$ is the identity. Hence, $M_5 \circ M_4 \circ M_3 = M_0^{-1} \circ M_1^{-1} \circ M_2^{-1}$, which evaluates to

$$\begin{pmatrix} \sqrt{3}\, s_4 & 3is_3s_4 - i \\ 3is_4s_5 - i & \sqrt{3}\,(s_3 + s_5 - 3s_3s_4s_5) \end{pmatrix} = \begin{pmatrix} \sqrt{3}\,(s_0 + s_2 - 3s_0s_1s_2) & i - 3is_0s_1 \\ i - 3is_1s_2 & \sqrt{3}\, s_1 \end{pmatrix}. \tag{4.6}$$

This is an equality of Möbius transformations, and is therefore valid up to a scalar factor. However, both sides have determinant 1, because the matrix in (4.5) has determinant 1, and therefore (4.6) is valid up to sign. From the upper left entries, we get

$$\pm s_4 = s_0 + s_2 - 3s_0s_1s_2. \tag{4.7}$$

Because s_4 is never zero, and since the set of configurations of 6 disks in a 'flower' around a given disk is connected, the sign in (4.7) does not depend on the configuration. When all circles have the same radius, $s_k = 1$, so the correct sign is minus. This proves (4.4) for $k=0$. The equations for the other values of k are valid by symmetry. □

5. The Laplacian of the discrete Schwarzians

In this section, we will use the equations of the s_k's of §4 to obtain the equations for the h_k's, and these will be used to show that $\Delta^\varepsilon h_k(v)$ is equal to a polynomial in $\varepsilon, h_{j_0}(v), L_{j_2} h_{j_1}(v)$; $j_0, j_1, j_2 \in \mathbf{Z}_6$. We consider only ε in the range $(0, 1)$, and therefore, if all h_k's are uniformly bounded in a vertex subset W, then so are the $\Delta^\varepsilon h_k$'s in the set $W^{(1)}$.

We substitute $s_k(v) = 1 + \varepsilon^2 h_k(v)$ in equation (4.4), simplify, and get

$$\begin{aligned} h_k(v) + h_{k+2}(v) + h_{k+4}(v) = {} & 3h_k(v) + 3h_{k+1}(v) + 3h_{k+2}(v) \\ & + 3\varepsilon^2 (h_k(v)h_{k+1}(v) + h_{k+1}(v)h_{k+2}(v) + h_{k+2}(v)h_k(v)) \\ & + 3\varepsilon^4 h_k(v)h_{k+1}(v)h_{k+2}(v). \end{aligned} \tag{5.1}$$

Set

$$\begin{aligned} \Psi_{k+1}(v) = {} & -(h_k(v)h_{k+1}(v) + h_{k+1}(v)h_{k+2}(v) + h_{k+2}(v)h_k(v)) \\ & - \varepsilon^2 h_k(v)h_{k+1}(v)h_{k+2}(v). \end{aligned} \tag{5.2}$$

Then equation (5.1) becomes

$$2h_k(v) + 3h_{k+1}(v) + 2h_{k+2}(v) - h_{k+4}(v) = 3\varepsilon^2 \Psi_{k+1}(v). \tag{5.3}$$

Replace k by $k+2$, multiply by 2, add to (5.3), and get

$$3h_{k+1}(v) + 6h_{k+2}(v) + 6h_{k+3}(v) + 3h_{k+4}(v) = 3\varepsilon^2 \Psi_{k+1}(v) + 6\varepsilon^2 \Psi_{k+3}(v). \tag{5.4}$$

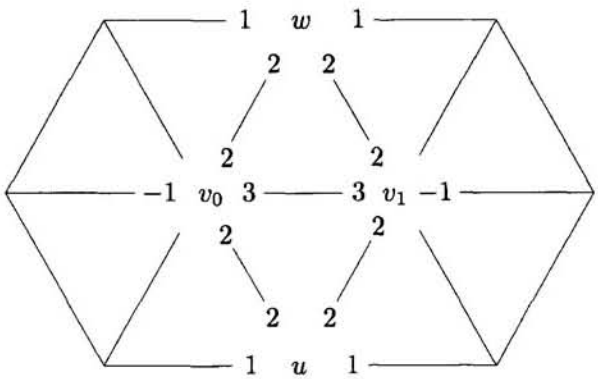


Fig. 5.1. The numbers mark the coefficients of the h-values in the identities (5.5) through (5.8).

LEMMA 5.1. *$\Delta^\varepsilon h_k(v)$ is equal to a polynomial in the variables ε, $h_{j_0}(v)$, $L^\varepsilon_{j_1} h_{j_2}(v)$; $j_0, j_1, j_2 \in \mathbf{Z}_6$. More specifically,*

$$\Delta^\varepsilon h_k = \tfrac{2}{3} L^\varepsilon_{k+1} \Psi_{k+3} + \tfrac{4}{3} L^\varepsilon_{k+1} \Psi_{k+5} + \tfrac{2}{3} L^\varepsilon_{k+5} \Psi_k + \tfrac{4}{3} L^\varepsilon_{k+5} \Psi_{k+2} - 2 L^\varepsilon_k \Psi_{k+3} - 2\Psi_k.$$

It is quite fortunate that $\Delta^\varepsilon h_k(v)$ is a polynomial in ε, $h_{j_0}(v)$, $L^\varepsilon_{j_1} h_{j_2}(v)$; $j_0, j_1, j_2 \in \mathbf{Z}_6$, since that simplifies many of the arguments that follow. However, the proof could be made to work if $\Delta^\varepsilon h_k(v)$ was only a C^∞-function of these variables. This might be important for generalizing the methods of this work to other settings. See §11 for further discussion.

Proof. We work on the case $k=0$. Fix some $v_0 \in (V^\varepsilon_\Omega)^{(2)}$. Let $v_1 = L_0 v_0$, $u = L_5 v_0$ and $w = L_1 v_0$. To understand the following computation, the reader is advised to consult Figure 5.1.

Apply equation (5.4) with v replaced by u and k replaced by 5, then use the relations $h_k(v) = h_{k+3}(L_k v)$, to get (after division by 3)

$$h_0(u) + 2h_4(v_1) + 2h_5(v_0) + h_0(L_4 v_0) = \varepsilon^2 \Psi_0(u) + 2\varepsilon^2 \Psi_2(u). \tag{5.5}$$

Similarly, apply equation (5.4) with v replaced by w and k replaced by 2, and obtain

$$h_0(L_2 v_0) + 2h_1(v_0) + 2h_2(v_1) + h_0(w) = \varepsilon^2 \Psi_3(w) + 2\varepsilon^2 \Psi_5(w). \tag{5.6}$$

The substitution v_0 for v and 5 for k in (5.3) gives

$$2h_5(v_0) + 3h_0(v_0) + 2h_1(v_0) - h_0(L_3 v_0) = 3\varepsilon^2 \Psi_0(v_0). \tag{5.7}$$

Similarly, the substitution v_1 for v and 2 for k in (5.3) gives

$$2h_2(v_1) + 3h_0(v_0) + 2h_4(v_1) - h_0(v_1) = 3\varepsilon^2 \Psi_3(v_1). \tag{5.8}$$

Now subtract the sum of equations (5.7) and (5.8) from the sum of equations (5.5) and (5.6), and get

$$\tfrac{3}{2}\varepsilon^2\Delta^\varepsilon h_0(v_0)=\varepsilon^2\Psi_3(w)+2\varepsilon^2\Psi_5(w)+\varepsilon^2\Psi_0(u)+2\varepsilon^2\Psi_2(u)-3\varepsilon^2\Psi_3(v_1)-3\varepsilon^2\Psi_0(v_0).$$

This proves the lemma for the case $k=0$. For other values of k the lemma is also valid, by symmetry. □

6. The discrete Schwarzians are bounded

In this section, we will recall the result of [6] which essentially says that the discrete Schwarzian derivative is uniformly bounded in any compact subset of the domain Ω. The bound is independent of ε. Precisely, we have the following lemma.

LEMMA 6.1. *Let v_0 be a vertex of V_Ω^ε, and suppose that the distance δ from v_0 to $\{z\in\mathbf{C}:|z|>1/\varepsilon\}-\Omega$ is greater than 2ε. Then*

$$|h_k(v_0)|=\varepsilon^{-2}|s_k(v_0)-1|\leqslant C,\quad k\in\mathbf{Z}_6, \tag{6.1}$$

for some constant $C=C(\delta)$, which depends only on δ.

The proof is quite similar to the proof of Lemma 1.5 in [5]. The boundedness of the Schwarzians is equivalent to the boundedness of the third order derivatives of f^ε. It is an open problem whether the above lemma can be proved directly using Rodin's equation (1.1) for the radii, or from the formula for the Laplacian of the Schwarzians.

Proof. By [6], there is a homeomorphism g^ε from the carrier of R_Ω^ε onto the carrier of P^ε with the following properties:

(1) For each $v\in V_\Omega^\varepsilon$, the image of the disk R_v^ε under g^ε is the corresponding disk P_v^ε.

(2) The restriction of g^ε to a dual disk of the packing R_Ω^ε is equal to a Möbius transformation.

(3) There is a universal constant $C_1>1$ such that for each $v\in(V_\Omega^\varepsilon)^{(2)}$, the map g^ε restricted to R_v^ε is C_1-quasiconformal.

(4) For each $v\in(V_\Omega^\varepsilon)^{(2)}$, there is a constant $C_2=C_2(\delta(v))>0$, which depends only on the distance $\delta(v)$ from v to $\{z\in\mathbf{C}:|z|>1/\varepsilon\}-\Omega$, such that the area of the subset of R_v^ε where g^ε fails to be conformal is bounded by $C_2\varepsilon^4$. (Note that the area of R_v^ε is $\frac{1}{4}\pi\varepsilon^2$.)

Consider the restriction of g^ε to $R_{v_0}^\varepsilon$. Let $\widetilde{D}_k$ be the dual disk bounded by the circle which passes through the tangency points of pairs of the disks $R_{v_0}^\varepsilon$, $R_{L_k^\varepsilon v_0}^\varepsilon$ and $R_{L_{k+1}^\varepsilon v_0}^\varepsilon$. See Figure 6.1.

Let z_1, z_2, z_3, z_4 be a quadruple of points on the circle $\partial R_{v_0}^\varepsilon$ which are: (1) sufficiently spaced out, say, $|z_{j_1}-z_{j_2}|\geqslant\frac{1}{100}\varepsilon$, for any $j_1\neq j_2$; (2) away from the points of tangency

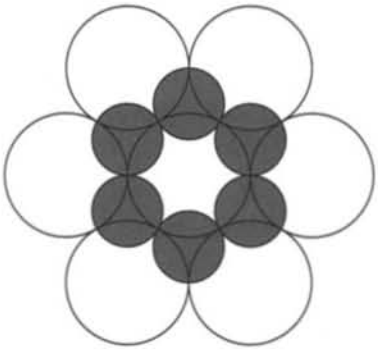


Fig. 6.1. The six dual disks.

$\{w_k\}=R^\varepsilon_{v_0}\cap R^\varepsilon_{L_k v_0}$, say, $|z_j-w_k|\geqslant\frac{1}{100}\varepsilon$; and (3) cyclicly ordered in the counter-clockwise direction. We claim that for any such quadruple,

$$|[g^\varepsilon(z_1),g^\varepsilon(z_2);g^\varepsilon(z_3),g^\varepsilon(z_4)]-[z_1,z_2;z_3,z_4]|\leqslant C_3\varepsilon^2, \tag{6.2}$$

where $[\cdot,\cdot;\cdot,\cdot]$ denotes the cross ratio, and C_3 depends only on δ.

In fact, there are positive numbers m and m^* such that the quadrilaterals $(R^\varepsilon_{v_0};z_1,z_2,z_3,z_4)$ and $(g^\varepsilon(R^\varepsilon_{v_0});g^\varepsilon(z_1),g^\varepsilon(z_2),g^\varepsilon(z_3),g^\varepsilon(z_4))$ are conformally homeomorphic to the standard rectangles

$$(Q_m=[0,m]\times[0,1];(0,0),(m,0),(m,1),(0,1))$$

and

$$(Q_{m^*}=[0,m^*]\times[0,1];(0,0),(m^*,0),(m^*,1),(0,1)),$$

respectively. The map g^ε will then be translated to a C_1-quasiconformal map F^ε between the standard rectangles

$$(Q_m;(0,0),(m,0),(m,1),(0,1)) \quad\text{and}\quad (Q_{m^*};(0,0),(m^*,0),(m^*,1),(0,1)).$$

The relations $|z_{j_1}-z_{j_2}|\geqslant\frac{1}{100}\varepsilon$, $j_1\neq j_2$, imply that m and $m^*\in[m/C_1,mC_1]$ are bounded from above and below by some universal positive constants. On the other hand, since $|z_j-w_k|\geqslant\frac{1}{100}\varepsilon$, by property (2) above we deduce that g^ε is conformal in the $\frac{1}{200}\varepsilon$-neighborhood of the points z_j, $1\leqslant j\leqslant 4$. Outside the $\frac{1}{200}\varepsilon$-neighborhoods of the z_j's, the conformal homeomorphism from R^ε onto Q_m is clearly Lipschitz, with Lipschitz constant bounded by $C_4\varepsilon^{-1}$, where C_4 is a universal constant. Thus, using property (4), it follows that F^ε is conformal except on a subset of area bounded by $C_4^2C_2\varepsilon^2$. It then

follows by a standard Grötsch argument (compare [6, §2]) that $|m^*/m-1|\leqslant C_5\varepsilon^2$, where C_5 depends on C_1 and C_2. Then (6.2) follows as m and m^* are related to the cross ratios $[z_1,z_2;z_3,z_4]$ and $[g^\varepsilon(z_1),g^\varepsilon(z_2);g^\varepsilon(z_3),g^\varepsilon(z_4)]$, respectively, by the same smooth function.

Let T_k be the Möbius transformation which agrees with g^ε on $\widetilde{D}_k$. Since a Möbius transformation is uniquely determined by its values at three points, it follows from (6.2) that for any $k\in\mathbf{Z}_6$,

$$|\varepsilon^{-1}T_k^{-1}\circ T_{k-1}(z)-\varepsilon^{-1}z|\leqslant C_6\varepsilon^2,\quad \forall z\in R^\varepsilon_{v_0}, \tag{6.3}$$

where C_6 depends only on C_3, and consequently, on δ.

It is then elementary to check that (6.3) implies that $|s_k(v_0)-1|\leqslant\varepsilon^2C_7$, with C_7 depending only on δ, which gives (6.1). □

7. Regularity of solutions of discrete elliptic equations

REGULARITY LEMMA 7.1. *Let W be a subset of V^ε, let $v_0\in W^{(1)}$, and let δ be the Euclidean distance from v_0 to $V^\varepsilon-W$. Let $\eta: W\to\mathbf{R}$ be any function. Then*

$$\delta|\partial_k^\varepsilon\eta(v_0)|\leqslant 7\|\eta\|_W+\tfrac{1}{2}\delta^2\|\Delta^\varepsilon\eta\|_{W^{(1)}} \tag{7.1}$$

holds for any $k\in\mathbf{Z}_6$.

This lemma is known in the continuous setting, and may certainly also be known in the discrete setting. The estimate is not sharp. For lack of a good reference, we include a proof.

Proof. Note that both sides of (7.1) are scale invariant; i.e., if we define $\hat\eta: a^{-1}W\to\mathbf{R}$ by $\hat\eta(v)=\eta(av)$, where $a>0$, then (7.1) for η is equivalent to (7.1) for $\hat\eta$ at $a^{-1}v_0$. Hence, we assume with no loss of generality that $\delta=1$. Also assume for convenience, $k=3$, $v_0=\varepsilon$. Again, there is no loss of generality.

Let $R(x+iy)=(\varepsilon-x)+iy$ be the reflection in the line $x=\frac{1}{2}\varepsilon$, and set

$$g(v)=\eta(v)-\eta(Rv).$$

As $L_3^\varepsilon v_0=v_0+(-1)\varepsilon=0=Rv_0$, what we need to prove is that

$$|g(v_0)|=|g(\varepsilon)|\leqslant\varepsilon\big(7\|\eta\|_W+\tfrac{1}{2}\|\Delta^\varepsilon\eta\|_{W^{(1)}}\big). \tag{7.2}$$

This is obvious if $2\leqslant 7\varepsilon$, so assume $7\varepsilon<2$. Also assume, with no loss of generality, that $g(\varepsilon)>0$.

Let

$$W_0=\left\{x+iy\in V^\varepsilon:\sqrt{\left(x-\tfrac{1}{2}\varepsilon\right)^2+y^2}<1-\tfrac{1}{2}\varepsilon\right\},$$
$$W_1=\left\{x+iy\in W_0:x\geqslant\tfrac{1}{2}\varepsilon\right\}.$$

Since $W_0=R(W_0)$ is contained in W, the function $g=\eta-\eta\circ R$ is well-defined in W_0. Denote $\tilde{x}=x-\frac{1}{2}\varepsilon$. Consider the "comparison" function $\hat{g}:W_0\to\mathbf{R}$ defined by

$$\hat{g}(x+iy)=\hat{g}\left(\tilde{x}+\tfrac{1}{2}\varepsilon+iy\right)=(7(\tilde{x}^2+y^2)+14(\tilde{x}-\tilde{x}^2))\|\eta\|_W+(\tilde{x}-\tilde{x}^2)\|\Delta^\varepsilon\eta\|_{W^{(1)}},\qquad(7.3)$$

and note that $\hat{g}\geqslant 0$ on W_1.

At any interior vertex of W_0, we have $\Delta^\varepsilon(g-\hat{g})\geqslant 0$, because

$$\Delta^\varepsilon g(v)=\Delta^\varepsilon\eta(v)-\Delta^\varepsilon\eta(Rv)\geqslant -2\|\Delta^\varepsilon\eta\|_{W^{(1)}},$$

and $\Delta^\varepsilon\hat{g}(v)=-2\|\Delta^\varepsilon\eta\|_{W^{(1)}}$.

In particular, $\Delta^\varepsilon(g-\hat{g})\geqslant 0$ at any interior vertex of W_1. Then it is elementary to see (from (2.4)) that the maximum of $g-\hat{g}$ on W_1 is attained at some boundary vertex, say $v_*=x_*+iy_*$, of W_1 (the maximum principle). We claim that

$$(g-\hat{g})(v_*)\leqslant\tfrac{7}{4}\varepsilon^2\|\eta\|_W+\tfrac{1}{4}\varepsilon^2\|\Delta^\varepsilon\eta\|_{W^{(1)}}.\qquad(7.4)$$

Clearly, the set of boundary vertices of W_1 is contained in the union of the subsets

$$B_1=\left\{x+iy\in W_1:x=\tfrac{1}{2}\varepsilon\right\},$$
$$B_2=\{x+iy\in W_1:x=\varepsilon\},$$
$$B_3=\left\{x+iy\in W_1:\sqrt{\left(x-\tfrac{1}{2}\varepsilon\right)^2+y^2}\geqslant 1-\tfrac{3}{2}\varepsilon\right\}.$$

If $v_*=\frac{1}{2}\varepsilon+iy_*\in B_1$, then $Rv_*=v_*$, $g(v_*)=0$, and $g(v_*)=0\leqslant\hat{g}(v_*)$, hence $(g-\hat{g})(v_*)\leqslant 0$. If $v_*\in B_3$ then

$$\hat{g}(v_*)\geqslant 7\left(\left(x_*-\tfrac{1}{2}\varepsilon\right)^2+y_*^2\right)\|\eta\|_W\geqslant 7\left(1-\tfrac{3}{2}\varepsilon\right)^2\|\eta\|_W$$
$$\geqslant 7\left(1-\tfrac{3}{7}\right)^2\|\eta\|_W\geqslant 2\|\eta\|_W\geqslant|g(v_*)|,$$

and therefore again $(g-\hat{g})(v_*)\leqslant 0$.

Now let $v_*=\varepsilon+iy_*\in B_2-B_3$. In this case, $\varepsilon+iy_*$ is an interior vertex of W_0. Since $\Delta^\varepsilon(g-\hat{g})(v_*)\geqslant 0$, and $(g-\hat{g})(\varepsilon+iy_*)$ is the maximum of $g-\hat{g}$ in W_1, and iy_* is the only neighbor of $v_*=\varepsilon+iy_*$ outside of W_1, it follows that

$$(g-\hat{g})(iy_*)\geqslant(g-\hat{g})(\varepsilon+iy_*).$$

As $g(iy_*)=g(R(\varepsilon+iy_*))=-g(\varepsilon+iy_*)$, we deduce that $\hat{g}(\varepsilon+iy_*)-\hat{g}(iy_*)\geqslant 2g(\varepsilon+iy_*)$, and then

$$(g-\hat{g})(\varepsilon+iy_*)\leqslant -\tfrac{1}{2}(\hat{g}(\varepsilon+iy_*)+\hat{g}(iy_*))\leqslant \tfrac{7}{4}\varepsilon^2\|\eta\|_W+\tfrac{1}{4}\varepsilon^2\|\Delta^\varepsilon\eta\|_{W^{(1)}}.$$

This proves estimate (7.4).

Since $(g-\hat{g})(v_*)$ is the maximum of $g-\hat{g}$ in W_1 and since $\varepsilon\in W_1$, we have $(g-\hat{g})(v_*)\geqslant (g-\hat{g})(\varepsilon)$, and then by (7.4),

$$\begin{aligned}|g(\varepsilon)|=g(\varepsilon)&=\hat{g}(\varepsilon)+(g-\hat{g})(\varepsilon)\leqslant \hat{g}(\varepsilon)+(g-\hat{g})(v_*)\\&\leqslant \hat{g}(\varepsilon)+\tfrac{7}{4}\varepsilon^2\|\eta\|_W+\tfrac{1}{4}\varepsilon^2\|\Delta^\varepsilon\eta\|_{W^{(1)}}=14\left(\tfrac{1}{2}\varepsilon\right)\|\eta\|_W+\tfrac{1}{2}\varepsilon\|\Delta^\varepsilon\eta\|_{W^{(1)}},\end{aligned}$$

which implies (7.2), and proves the lemma. □

8. The discrete Schwarzians converge

In this section, we show that for some sequence of $\varepsilon\to 0$, the Schwarzians h_k converge in C^∞. In a later section, the limit will be identified, and it will follow that the limit exists even without restricting to a subsequence.

We shall sometimes write h_k^ε for h_k, to stress the dependence on ε. From Lemma 6.1 we know that the functions h_k^ε are bounded on compact subsets of Ω, with a bound independent of ε. Lemma 5.1 then shows that the functions $\Delta^\varepsilon h_k^\varepsilon$ are also uniformly bounded on compact subsets of Ω. Applying Lemma 7.1, we see that also $\partial_j^\varepsilon h_k^\varepsilon$ have such a bound. It then follows, by Lemma 2.1, that for some sequence of ε tending to 0 the continuous limits

$$\mathcal{H}_k=\lim_{\varepsilon\to 0} h_k^\varepsilon \tag{8.1}$$

exist, and are locally Lipschitz functions on Ω.

Note for future use that

$$\mathcal{H}_{k+3}=\mathcal{H}_k, \tag{8.2}$$

because $h_k^\varepsilon(v)=h_{k+3}^\varepsilon(L_k^\varepsilon v)$. Using this together with (5.4) gives

$$\mathcal{H}_0+\mathcal{H}_1+\mathcal{H}_2=0. \tag{8.3}$$

For each $\delta>0$ let V_δ^ε denote the set of vertices of V^ε whose Euclidean distance to $\{z\in\mathbf{C}:|z|>1/\varepsilon\}-\Omega$ is at least δ.

LEMMA 8.1. *Let $\delta>0$, and let n be an integer. Then there are constants $C=C(n,\delta)$, $\alpha=\alpha(n,\delta)>0$ such that*

$$\|\partial_{k_n}^\varepsilon\partial_{k_{n-1}}^\varepsilon\cdots\partial_{k_1}^\varepsilon h_{k_0}^\varepsilon\|_{V_\delta^\varepsilon}<C \tag{8.4}$$

holds whenever $\varepsilon<\alpha$, *and* $k_0, k_1, \ldots, k_n \in \mathbf{Z}_6$. *In other words, the functions* h_k^ε *are uniformly bounded in* $C^\infty(\Omega)$.

Proof. The proof will be inductive. The case $n=0$ is handled by Lemma 6.1. So assume that $n>0$, and that the lemma holds for $0, 1, 2, \ldots, n-1$. Set

$$g = \partial^\varepsilon_{k_{n-1}} \ldots \partial^\varepsilon_{k_1} h^\varepsilon_{k_0}.$$

Then,

$$\Delta^\varepsilon g = \Delta^\varepsilon \partial^\varepsilon_{k_{n-1}} \ldots \partial^\varepsilon_{k_1} h^\varepsilon_{k_0} = \partial^\varepsilon_{k_{n-1}} \ldots \partial^\varepsilon_{k_1} \Delta^\varepsilon h^\varepsilon_{k_0}.$$

From Lemma 5.1 it now follows that $\Delta^\varepsilon g$ is a linear combination of functions of the form

$$\partial^\varepsilon_{k_{n-1}} \ldots \partial^\varepsilon_{k_1} \Lambda \Psi_{j_1},$$

with $\Lambda = L^\varepsilon_{j_2}$ or $\Lambda = I$, the identity operator. Recall, from (5.2), that Ψ_j is a polynomial in ε and the h_k's. Also note the rule for discrete differentiation of a product,

$$\partial^\varepsilon_k(\psi_1\psi_2) = (\partial^\varepsilon_k \psi_1)\psi_2 + (L^\varepsilon_k \psi_1)\partial^\varepsilon_k \psi_2,$$

which is easy to verify. From this rule, it follows that $\partial^\varepsilon_{j_2} \Psi_{j_1}$ is a polynomial in ε and the Λh_j's, where Λ ranges over the operators I, $\partial^\varepsilon_{j_2}$, L_{j_2}. By induction, it follows that

$$\partial^\varepsilon_{k_{n-1}} \ldots \partial^\varepsilon_{k_1} \Psi_k$$

is a polynomial in ε and expressions of the form

$$L^\varepsilon_{j_m} \ldots L^\varepsilon_{j_{s+1}} \partial^\varepsilon_{j_s} \ldots \partial^\varepsilon_{j_1} h_{j_0}, \tag{8.5}$$

where $m \leqslant n-1$. If $v \in V^\varepsilon_{\delta/2}$, $m \leqslant n$, and $4n\varepsilon < \delta$, then $v' = L^\varepsilon_{j_m} \ldots L^\varepsilon_{j_{s+1}}(v)$ is in $v \in V^\varepsilon_{\delta/4}$. Therefore, the inductive hypothesis with $v' = L^\varepsilon_{j_m} \ldots L^\varepsilon_{j_{s+1}}(v)$, $n' = s$, $\delta' = \frac{1}{4}\delta$ applies, and provides a bound for (8.5) at v. Since $\Delta^\varepsilon g$ is a polynomial in ε and the expressions of the form (8.5) with $m \leqslant n, s \leqslant n-1$, it follows that there is a constant $C_1 = C_1(\delta)$ such that

$$\|\Delta^\varepsilon g(v)\|_{V^\varepsilon_{\delta/2}} \leqslant C_1.$$

Because $|g|$ is also bounded on $V^\varepsilon_{\delta/2}$, the Regularity Lemma 7.1 provides a bound for $|\partial^\varepsilon_{k_n} g|$, on V^ε_δ, which completes the induction step and the proof. □

COROLLARY 8.2. *$h_k^\varepsilon \to \mathcal{H}_k$ in $C^\infty(\Omega)$ as $\varepsilon \to 0$.*

Proof. It follows from Lemma 8.1 and Lemma 2.1 that this is true for some sequence of $\varepsilon \to 0$. The general statement will follow later, when (8.1) is proved in full generality. □

Remark 8.3. It can be shown that the rate of convergence in Corollary 8.2 depends only on n and δ; i.e., there is a function $A(n,\delta,\varepsilon)$ such that

$$\|\partial_{k_n}^\varepsilon \dots \partial_{k_1}^\varepsilon h_k^\varepsilon - \partial_{k_n} \dots \partial_{k_1} \mathcal{H}_k\|_{V_\delta^\varepsilon} < A(n,\delta,\varepsilon) \quad \text{and} \quad A(n,\delta,\varepsilon) \to 0 \ \text{ as } \varepsilon \to 0.$$

The following proposition will not be used below, but is interesting in itself, and would probably prove useful for studying the rate of convergence in Corollary 8.2.

PROPOSITION 8.4. *Let $\delta > 0$, let $k \in \mathbf{Z}_6$, and suppose that $2\varepsilon < \delta$. Then*

$$\|\Delta^\varepsilon h_k\|_{V_\delta^\varepsilon} \leqslant \varepsilon^2 C^*,$$

where $C^* = C^*(\delta)$ depends only on δ.

We will need the following notation. Write

$$a \overset{\varepsilon^n}{=} b$$

when $a-b$ is a polynomial in ε, and in the functions h_k and their discrete derivatives of arbitrary order, which is divisible by ε^n. We also use a similar notation for discrete operators. For example, the relations $L_{k-1}^\varepsilon L_{k+1}^\varepsilon = L_k^\varepsilon$ and $L_j^\varepsilon = I + \varepsilon\partial_j^\varepsilon$ give

$$\partial_{k-1}^\varepsilon + \partial_{k+1}^\varepsilon \overset{\varepsilon}{=} \partial_k^\varepsilon. \tag{8.6}$$

Proof. We will prove that $\Delta^\varepsilon h_k \overset{\varepsilon^2}{=} 0$. This actually proves more than the proposition; it shows that the discrete derivatives of $\Delta^\varepsilon h_k$ are also $O(\varepsilon^2)$.

Use Lemma 5.1 and the identities $L_j^\varepsilon = I + \varepsilon\partial_j^\varepsilon$ to write

$$\begin{aligned}
\tfrac{3}{2}\Delta^\varepsilon h_k &= L_{k+1}^\varepsilon \Psi_{k+3} + 2L_{k+1}^\varepsilon \Psi_{k+5} + L_{k+5}^\varepsilon \Psi_k + 2L_{k+5}^\varepsilon \Psi_{k+2} - 3L_k^\varepsilon \Psi_{k+3} - 3\Psi_k \\
&= \Psi_{k+3} + 2\Psi_{k+5} + \Psi_k + 2\Psi_{k+2} - 3\Psi_{k+3} - 3\Psi_k \\
&\quad + \varepsilon(\partial_{k+1}^\varepsilon \Psi_{k+3} + 2\partial_{k+1}^\varepsilon \Psi_{k+5} + \partial_{k+5}^\varepsilon \Psi_k + 2\partial_{k+5}^\varepsilon \Psi_{k+2} - 3\partial_k^\varepsilon \Psi_{k+3}) \qquad (8.7) \\
&= -2\Psi_k + 2\Psi_{k+2} - 2\Psi_{k+3} + 2\Psi_{k+5} \\
&\quad + \varepsilon(\partial_{k+1}^\varepsilon \Psi_{k+3} + 2\partial_{k+1}^\varepsilon \Psi_{k+5} + \partial_{k+5}^\varepsilon \Psi_k + 2\partial_{k+5}^\varepsilon \Psi_{k+2} - 3\partial_k^\varepsilon \Psi_{k+3}).
\end{aligned}$$

Recall that (5.3) gives $\Psi_j = \frac{1}{3}\varepsilon^{-2}(2h_{j-1} + 3h_j + 2h_{j+1} - h_{j+3})$. This relation implies that $-\Psi_k + \Psi_{k+2} - \Psi_{k+3} + \Psi_{k+5} = 0$. Therefore, (8.7) reduces to

$$\tfrac{3}{2}\Delta^\varepsilon h_k = \varepsilon(\partial_{k+1}^\varepsilon \Psi_{k+3} + 2\partial_{k+1}^\varepsilon \Psi_{k+5} + \partial_{k+5}^\varepsilon \Psi_k + 2\partial_{k+5}^\varepsilon \Psi_{k+2} - 3\partial_k^\varepsilon \Psi_{k+3}). \tag{8.8}$$

The relation $h_{k+2}=L^\varepsilon_{k+2}h_{k-1}=h_{k-1}+\varepsilon\partial^\varepsilon_{k+2}h_{k-1}$, and the definition (5.2) of Ψ give

$$\begin{aligned}\Psi_{k+1} &\overset{\varepsilon^2}{=} -h_kh_{k+1}-h_{k+1}h_{k+2}-h_{k+2}h_k\\ &= -h_kh_{k+1}-h_{k+1}(h_{k-1}+\varepsilon\partial^\varepsilon_{k+2}h_{k-1})-(h_{k-1}+\varepsilon\partial^\varepsilon_{k+2}h_{k-1})h_k \qquad (8.9)\\ &\overset{\varepsilon}{=} -h_kh_{k+1}-h_{k+1}h_{k-1}-h_{k-1}h_k \overset{\varepsilon^2}{=} \Psi_k.\end{aligned}$$

This means that

$$\Psi_j \overset{\varepsilon}{=} \Psi_k, \quad \text{for each } j,k\in\mathbf{Z}_6,$$

which also implies that $\partial^\varepsilon_m\Psi_j\overset{\varepsilon}{=}\partial^\varepsilon_m\Psi_k$, for $m,j,k\in\mathbf{Z}_6$. So (8.8) further reduces to

$$\tfrac{3}{2}\Delta^\varepsilon h_k \overset{\varepsilon^2}{=} \varepsilon(3\partial^\varepsilon_{k+1}+3\partial^\varepsilon_{k+5}-3\partial^\varepsilon_k)\Psi_k. \qquad (8.10)$$

Now the relation (8.6) shows that $\frac{3}{2}\Delta^\varepsilon h_k\overset{\varepsilon^2}{=}0$, and the proof is complete. □

Remark 8.5. Another interesting identity is

$$\tfrac{3}{2}\varepsilon\partial^\varepsilon_j h_{j+3}\overset{\varepsilon^2}{=}h_j+h_{j+2}+h_{j+4}.$$

This follows from (5.3) with $k=j+2$.

9. The contact transformations

Let $v\in V^\varepsilon_\Omega$ be some interior vertex, and let $k\in\mathbf{Z}_6$. Define the *contact transformation* $Z^\varepsilon_k=Z^\varepsilon_k(v)$ to be the Möbius transformation that takes each of the three points $\varepsilon,\varepsilon\omega^2,\varepsilon\omega^4$ to the three points in $P^\varepsilon_v\cap P^\varepsilon_{L^\varepsilon_k v}$, $P^\varepsilon_{L^\varepsilon_k v}\cap P^\varepsilon_{L^\varepsilon_{k+1}}$ and $P^\varepsilon_v\cap P^\varepsilon_{L^\varepsilon_{k+1}}$, respectively.

We now derive expressions for the discrete derivatives of $Z^\varepsilon_k(v)$ with respect to v. These will enable us to show that the Möbius transformation Z^ε_k converge C^∞ as $\varepsilon\to0$.

Let R be the rotation $R(z)=\omega^2 z$. It is clear that

$$Z^\varepsilon_{k+2}(L^\varepsilon_k v)=Z^\varepsilon_k(v)\cdot R. \qquad (9.1)$$

Let $A=A^\varepsilon$ be the Möbius transformation taking $\varepsilon,\varepsilon\omega^2,\varepsilon\omega^4$ to $0,1,\infty$, respectively. Using the Möbius transformations $m_k(v)$ from §4, we may write

$$Z^\varepsilon_k(v)=m_{k+1}(v)^{-1}\cdot A.$$

With the notation $M_k=m_{k+1}\cdot m_k^{-1}$, we then have

$$Z^\varepsilon_{k-1}=m_k^{-1}\cdot A=m_{k+1}^{-1}\cdot A\cdot A^{-1}\cdot M_k\cdot A=Z^\varepsilon_k\cdot A^{-1}\cdot M_k\cdot A.$$

Set

$$Q_k = A^{-1} \cdot M_k \cdot A. \tag{9.2}$$

Then the above relation can be abbreviated

$$Z^\varepsilon_{k-1} = Z^\varepsilon_k \cdot Q_k. \tag{9.3}$$

Equations (9.1) and (9.3) will allow us to write the expression for $Z^\varepsilon_j(L^\varepsilon_k v)$ in terms of $Z^\varepsilon_k(v)$ and the transition matrices Q_k, R:

$$\begin{aligned} L^\varepsilon_k Z^\varepsilon_k &= L^\varepsilon_k (Z^\varepsilon_{k+2} \cdot Q_{k+2} \cdot Q_{k+1}) = L^\varepsilon_k Z^\varepsilon_{k+2} \cdot L^\varepsilon_k Q_{k+2} \cdot L^\varepsilon_k Q_{k+1} \\ &= Z^\varepsilon_k \cdot R \cdot L^\varepsilon_k Q_{k+2} \cdot L^\varepsilon_k Q_{k+1}. \end{aligned} \tag{9.4}$$

Matrix representations of R and A are

$$R = \begin{pmatrix} \omega & 0 \\ 0 & \omega^{-1} \end{pmatrix}, \quad A = \begin{pmatrix} -\omega^2 & \varepsilon\omega^2 \\ 1 & \varepsilon\omega \end{pmatrix},$$

and the expression (4.5) for M_k can be written

$$M_k = \begin{pmatrix} 0 & i \\ i & -\sqrt{3}\, s_k \end{pmatrix} = \begin{pmatrix} 0 & i \\ i & -\sqrt{3}\,(1+\varepsilon^2 h_k) \end{pmatrix}. \tag{9.5}$$

This gives the following expression for Q_k:

$$Q_k = \begin{pmatrix} i\omega(1+\varepsilon^2 h_k) & i\omega^2\varepsilon(2+\varepsilon^2 h_k) \\ \varepsilon i\omega h_k & i\omega^2(1+\varepsilon^2 h_k) \end{pmatrix}.$$

Observe that Q_k is polynomial in ε and h_k. A direct computation gives the leading terms for $R \cdot L^\varepsilon_k Q_{k+2} \cdot L^\varepsilon_k Q_{k+1}$ as

$$R \cdot L^\varepsilon_k Q_{k+2} \cdot L^\varepsilon_k Q_{k+1} = I + \varepsilon \begin{pmatrix} 0 & 2i\sqrt{3} \\ \omega^5 L^\varepsilon_k h_{k+1} + \omega^4 L^\varepsilon_k h_{k+2} & 0 \end{pmatrix} + \varepsilon^2 O(1), \tag{9.6}$$

where I is the identity matrix, and $O(1)$ denotes some matrix that is polynomial in $\varepsilon, L^\varepsilon_k h_{k+1}, L^\varepsilon_k h_{k+2}$. The equations (9.4) and (9.6) give an expression for the discrete derivative $\partial^\varepsilon_k Z^\varepsilon_k$ as

$$\partial^\varepsilon_k Z^\varepsilon_k = Z^\varepsilon_k \cdot \begin{pmatrix} 0 & 2i\sqrt{3} \\ \omega^5 L^\varepsilon_k h_{k+1} + \omega^4 L^\varepsilon_k h_{k+2} & 0 \end{pmatrix} + \varepsilon Z^\varepsilon_k \cdot O(1). \tag{9.7}$$

An entirely similar computation gives

$$\partial^\varepsilon_{k-2} Z^\varepsilon_k = Z^\varepsilon_k \cdot \begin{pmatrix} 0 & 3-\sqrt{3}\,i \\ \omega h_{k-1} + h_k & 0 \end{pmatrix} + \varepsilon Z^\varepsilon_k \cdot O(1). \tag{9.8}$$

(Of course, the $O(1)$-matrix here is not necessarily the same as in (9.7).) Because of the relations $\partial^\varepsilon_{k+3}=-\partial^\varepsilon_k L^\varepsilon_{k+3}$ and $\partial^\varepsilon_{k-1}=\partial^\varepsilon_k+\partial^\varepsilon_{k-2}L^\varepsilon_k$, the expressions for all the derivatives $\partial^\varepsilon_j Z^\varepsilon_k$, $j\in\mathbf{Z}_6$, can be obtained from (9.7) and (9.8).

Recall that z_0 denotes an arbitrary but fixed point in Ω. For each small $\varepsilon>0$ let $v_0=v_0^\varepsilon$ be some vertex in V_Ω^ε which is closest to z_0. Let

$$\widehat{Z}^\varepsilon_k(v)=Z^\varepsilon_k(v_0)^{-1}Z^\varepsilon_k(v).$$

Then we have from (9.7) and (9.8) and the similar expressions for the other $\partial^\varepsilon_j Z^\varepsilon_k$,

$$L^\varepsilon_j\widehat{Z}^\varepsilon_k=\widehat{Z}^\varepsilon_k\cdot(I+\varepsilon O(1)).$$

Since the absolute values of the entries in $(I+\varepsilon O(1))^n$ are bounded by e^{Cn}, for some constant C, and since $\widehat{Z}^\varepsilon_k(v_0)=I$, it follows that the matrices $\widehat{Z}^\varepsilon_k$ are bounded in compact subsets of Ω, independently of ε. From the similar relations for Z^ε_k, we have for each $j\in\mathbf{Z}_6$,

$$\partial^\varepsilon_j\widehat{Z}^\varepsilon_k=\widehat{Z}^\varepsilon_k\cdot O(1). \tag{9.9}$$

Note that the $O(1)$-term is bounded in $C^\infty(\Omega)$. Therefore, repeated differentiation of (9.9) shows that $\widehat{Z}^\varepsilon_k$ is bounded in $C^\infty(\Omega)$ uniformly in ε. By Lemma 2.1, it follows that for some sequence of $\varepsilon\to0$, the limit

$$\widehat{\mathcal{Z}}_k=\lim_{\varepsilon\to0}\widehat{Z}^\varepsilon_k$$

exists, and the convergence is $C^\infty(\Omega)$.

Multiply equation (9.7) on the left by $Z^\varepsilon_k(v_0)^{-1}$, and take a limit as $\varepsilon\to0$, to obtain

$$\partial_k\widehat{\mathcal{Z}}_k=\widehat{\mathcal{Z}}_k\cdot\begin{pmatrix}0 & 2i\sqrt{3}\\ \omega^5\mathcal{H}_{k+1}+\omega^4\mathcal{H}_{k+2} & 0\end{pmatrix}. \tag{9.10}$$

Applying a similar procedure to (9.8) gives

$$\partial_{k-2}\widehat{\mathcal{Z}}_k=\widehat{\mathcal{Z}}_k\cdot\begin{pmatrix}0 & 3-\sqrt{3}\,i\\ \omega\mathcal{H}_{k-1}+\mathcal{H}_k & 0\end{pmatrix}. \tag{9.11}$$

The identities (9.10), (9.11), (8.2) and (8.3) now imply

$$\partial_{k-2}\widehat{\mathcal{Z}}_k=\omega^{-2}\partial_k\widehat{\mathcal{Z}}_k,$$

which shows that $\widehat{\mathcal{Z}}_k(z)$ is a (matrix-valued) analytic function of z.

Observe that the equations (9.10) and (9.11) show that the determinant of $\widehat{\mathcal{Z}}_k(z)$ is constant in Ω. At z_0 this determinant is 1. Therefore, $\widehat{\mathcal{Z}}_k(z)$ is a Möbius transformation for every $z\in\Omega$.

The next step is to show that there is also convergence of a subsequence of Z_k^ε as $\varepsilon \to 0$. For this, all that is needed is to show that the transformations $Z_k^\varepsilon(v_0)$ are bounded independently of ε. Let $\widehat{\mathcal{Z}}_k(z)(w)$ denote the image of w under the Möbius transfromation $\widehat{\mathcal{Z}}_k(z)$. Since $\widehat{\mathcal{Z}}_k(z_0)=I$, we find from (9.10),

$$\left.\frac{\partial}{\partial z}\right|_{z=0} \widehat{\mathcal{Z}}_0(z)(0) = 2i\sqrt{3} \neq 0.$$

It follows that for a sufficiently small $\alpha>0$, the three points $w_0=\widehat{\mathcal{Z}}_0(z_0)(0)$, $w_1=\widehat{\mathcal{Z}}_0(z_0-\alpha)(0)$, $w_2=\widehat{\mathcal{Z}}_0(z_0+\alpha)(0)$ are all distinct. Set $z_1=z_0-\alpha$ and $z_2=z_0+\alpha$.

From the Rodin–Sullivan Theorem and the definition of Z_0^ε, it follows that

$$\lim_{\varepsilon\to 0} Z_0^\varepsilon(z)(\varepsilon) = f(z),$$

where f is the Riemann map $f:\Omega\to U$. For each $\varepsilon>0$, let v_1, v_2 be vertices of V^ε closest to z_1, z_2, respectively. The Möbius transformation $Z_0^\varepsilon(v_0)$ takes each of the three points $\widehat{Z}_0^\varepsilon(v_0)(\varepsilon), \widehat{Z}_0^\varepsilon(v_1)(\varepsilon), \widehat{Z}_0^\varepsilon(v_2)(\varepsilon)$ to a point close to $f(z_0), f(z_1), f(z_2)$, respectively. The former triplet of points are close to w_0, w_1, w_2, respectively. Consequently, $\lim_{\varepsilon\to 0} Z_0^\varepsilon(v_0)$ exists, and is the Möbius transformation that takes w_0, w_1, w_2 to $f(z_0), f(z_1), f(z_2)$, respectively. It is easy to see that the same must hold for the other transformations $Z_k^\varepsilon(v_0)$. Consequently, there is C^∞-convergence,

$$\mathcal{Z}_k = \lim_{\varepsilon\to 0} Z_k^\varepsilon,$$

along some subsequence of $\varepsilon\to 0$.

THEOREM 9.1. *Let f be the Riemann map from Ω to the unit disk U. Then*

$$\mathcal{S}(f) = 4(\mathcal{H}_0+\omega^4\mathcal{H}_1+\omega^2\mathcal{H}_2). \tag{9.12}$$

Using (8.3), this also gives

$$6\mathcal{H}_k = \mathrm{Re}(\omega^{2k}\mathcal{S}(f)),$$

for $k\in\mathbf{Z}_6$. The theorem then implies that Corollary 8.2 and (8.1) are valid for every sequence of $\varepsilon\to 0$, not just for one particular sequence.

Proof. Write

$$\mathcal{Z}_0(z) = \begin{pmatrix} a(z) & b(z) \\ c(z) & d(z) \end{pmatrix}.$$

It follows from $f(z)=\mathcal{Z}_0(z)(0)$ that $f(z)=b(z)/d(z)$. Equation (9.10) is also valid for the matrix $\mathcal{Z}_k$, and therefore $b(z)$ and $d(z)$ both satisfy the differential equation

$$w'' = 2i\sqrt{3}\,(\omega^5\mathcal{H}_1+\omega^4\mathcal{H}_2)w. \tag{9.13}$$

Consequently, we have $(b'd-bd')'=0$, and therefore $b'd-bd'$ is a constant, say α. Hence, $(b/d)'=\alpha/d^2$. Using the fact that d satisfies (9.13), the definition (4.1) of $\mathcal{S}(f)$, and the identity (8.3), it is easy now to verify that $\mathcal{S}(f)=\mathcal{S}(b/d)$ is the same as the right-hand side of (9.12). □

10. C^∞-convergence of disk packings

Let the situation be as described in §3. For each $v\in V_\Omega^\varepsilon$, let $g^\varepsilon(v)$ be the contact point of P_v^ε with $P_{L_0^\varepsilon v}^\varepsilon$. We recall that $f^\varepsilon(v)$ denotes the center of the disk P_v^ε, and $r^\varepsilon(v)$ the radius of P_v^ε. Clearly, Theorem 1.1 follows from the following theorem.

THEOREM 10.1. *In the setting of* §3, f^ε *and* g^ε *converge to the conformal map* $f:\Omega\to U$ *in* $C^\infty(\Omega)$, *and* $2r^\varepsilon/\varepsilon$ *converges in* $C^\infty(\Omega)$ *to* $|f'|$.

Proof. Let us start with g^ε. We have $g^\varepsilon(v)=Z_0^\varepsilon(v)(\varepsilon)$. Write $Z_0^\varepsilon(v)$ as a matrix,

$$Z_0^\varepsilon(v)=\begin{pmatrix} a^\varepsilon(v) & b^\varepsilon(v)\\ c^\varepsilon(v) & d^\varepsilon(v)\end{pmatrix}.$$

Similarly, let the entries of $\mathcal{Z}_0(z)$ be $a(z),b(z),c(z),d(z)$. Then $\varepsilon c^\varepsilon+d^\varepsilon$ converges C^∞ to d and $\varepsilon a^\varepsilon+b^\varepsilon$ converges C^∞ to b. Note that d is nonzero in Ω, because $b(z)/d(z)=\mathcal{Z}_0(z)(0)=f(z)$, and the determinant of $\mathcal{Z}_0(z)$ is nonzero. Consequently, by (2) and (3) of Lemma 2.2, it follows that $(\varepsilon a^\varepsilon+b^\varepsilon)/(\varepsilon c^\varepsilon+d^\varepsilon)\to b/d$ in C^∞. But that is the same as $g^\varepsilon\to f$.

Let c_1 be the circle that contains the three points $\varepsilon,\varepsilon\omega^2,\varepsilon\omega^4$. Because $Z_0^\varepsilon(v)$ maps the three points $\varepsilon,\varepsilon\omega^2,\varepsilon\omega^4$ to $P_v^\varepsilon\cap P_{L_0^\varepsilon v}^\varepsilon$, $P_{L_1^\varepsilon v}^\varepsilon\cap P_{L_0^\varepsilon v}^\varepsilon$, $P_v^\varepsilon\cap P_{L_1^\varepsilon v}^\varepsilon$, respectively, it maps c_1 onto the dual circle of the triangular interstice. Consequently, $Z_0^\varepsilon(v)$ maps the circle c_2 passing through ε and $\varepsilon\omega^4$ which is orthogonal to c_1 onto ∂P_v^ε. Let εp_2 be the center, and $\varepsilon\varrho_2$ be the radius of c_2. Then p_2 and ϱ_2 are constants.

How can we find a formula for $f^\varepsilon(v)$, the center of P_v^ε, in terms of $Z_0^\varepsilon(v)$? The inversion of $f^\varepsilon(v)$ in ∂P_v^ε is ∞, obviously. The preimage of that inversion under $Z_0^\varepsilon(v)$ is the pole of $Z_0^\varepsilon(v)$. Let q be the inversion in c_2 of the pole of $Z_0^\varepsilon(v)$. Note that when a point z_2 is the image of a point z_1 under an inversion in a circle c, then any Möbius transformtion m will take z_2 to the inversion of $m(z_1)$ in the circle $m(c)$. Consequently, $Z_0^\varepsilon(v)(q)=f^\varepsilon(v)$. The pole of $Z_0^\varepsilon(v)$ is just the point $-d^\varepsilon(v)/c^\varepsilon(v)$. This gives

$$q=\varepsilon p_2+\frac{\varepsilon^2\varrho_2^2}{-\varepsilon\overline{p_2}-\overline{d^\varepsilon(v)}/\overline{c^\varepsilon(v)}}=\varepsilon p_2+\frac{\varepsilon^2\varrho_2^2\overline{c^\varepsilon(v)}}{-\varepsilon\overline{c^\varepsilon(v)}p_2-\overline{d^\varepsilon(v)}}.$$

It is therefore clear from Lemma 2.2 that $f^\varepsilon(v)=Z_0^\varepsilon(v)(q)$ converges C^∞ to f.

Note that

$$r^\varepsilon(w)+r^\varepsilon(L_k^\varepsilon w)=|f^\varepsilon(w)-f^\varepsilon(L_k^\varepsilon w)|=\varepsilon|\partial_k^\varepsilon f^\varepsilon(v)|.$$

The three instances of this relation, $(w=v,k=0)$, $(w=v,k=1)$ and $(w=L_0^\varepsilon v,k=2)$, give

$$r^\varepsilon(v)=\tfrac{1}{2}\varepsilon(|\partial_0^\varepsilon f^\varepsilon(v)|+|\partial_1^\varepsilon f^\varepsilon(v)|-|\partial_2^\varepsilon f^\varepsilon(L_0 v)|). \tag{10.1}$$

Because $\partial_0^\varepsilon f^\varepsilon$ converges C^∞ to f' it follows from Lemma 2.2 that $|\partial_0^\varepsilon f|$ converges C^∞ to $|f'|$. Similarly, $|\partial_k^\varepsilon f|\to|f'|$. Consequently, $2r^\varepsilon/\varepsilon$ converges C^∞ to $|f'|$. □

11. Remarks: More general combinatorics and disk patterns

Consider a locally finite disk packing R in the plane, which is invariant under two linearly independent translations, and assume that all the interstices are triangular. It is well known that one may use the rescaled packings $R^\varepsilon=\varepsilon R=\{\varepsilon D: D\in R\}$, $\varepsilon>0$, to construct approximations to conformal mappings, and that the derivatives up to the second order converge (see, e.g., [8] or [9]). We now discuss the definition and convergence of the Schwarzian in this general case.

We will use G^ε to denote the embedded graph of the packing εR, and denote by Ω^ε the domain which approximates Ω; R_Ω^ε the subpacking of εR contained in Ω; G_Ω^ε the graph of R_Ω^ε; and V_Ω^ε the vertex set of G_Ω^ε. Let P^ε be the disk packing in $\overline{U}$ whose graph is equivalent to G_Ω^ε, such that the "boundary" disks are all tangent to the unit circle ∂U. We may then define f^ε to be the map which maps the centers of disks in R_Ω^ε to the centers of the corresponding disks in P^ε. The problem is, does f^ε converge in C^∞? The meaning of the convergence is that the restriction of f^ε to the intersection of V_Ω^ε with any lattice converges.

For each interior edge e in the graph G_Ω^ε, we may define two Möbius invariants $s'(e)$ and $s''(e)$, corresponding to the packings R_Ω^ε and P^ε, respectively. (The invariant $s'(e)$ is obtained by comparing the configuration of the four circles in R_Ω^ε related to e to a configuration of four circles related to an edge in the hexagonal combinatorics, and similarly for $s''(e)$.) Then the Schwarzian derivative on the edge may be defined by $h(e)=\varepsilon^{-2}(s''(e)-s'(e))$. The equations of $s'(e)$ and $s''(e)$ can be derived similarly.

The proof of the uniform boundedness of $h(e)$ (Lemma 6.1) is almost identical in this case. In fact, the proof of [6] works even better in the general abstract setting (see [8] for the necessary modifications). However, except for a few nice cases, it seems difficult to derive the equation for the "Laplacian" of the Schwarzians from the equations of the Möbius invariants.

One may consider disk patterns instead of disk packings. In this case, C^0-convergence follows by a compactness argument using the rigidity theorem of [7] and the topological

lemma of [9]. In the special case of the square grid pattern (or in short SG pattern), the Möbius invariants and their equations have been worked out in [16], as we remarked earlier. One may define the discrete Schwarzians and find a similar formula for their Laplacians as in §5 of the present paper. On the other hand, because the angles of intersection between pairs of disks in a square grid pattern are either 0 or $\frac{1}{2}\pi$, the inversions on the circles generate a Kleinian group. Thus the method of [6] can be extended, although some nontrivial modifications are needed. This implies the boundedness of the Schwarzians. So the technique of this paper does generalize to SG patterns. The general case is still open.

There is an alternative method, which has been successful in proving the rigidity of locally finite disk patterns in the plane (see [7]). Using a refinement of that argument, it is possible to study the logarithm of the ratio of radii of a pair of disk patterns, and show that they converge to the harmonic function $\log|f'(z)|$, where $f(z)$ denotes the conformal mapping from Ω to U. We conjecture that the convergence is C^∞. If true, it would then be an elementary matter to deduce the C^∞-convergence of the discrete function which maps the center of a disk to the center of the corresponding disk.

References

[1] AHARONOV, D., The hexagonal packing lemma and discrete potential theory. *Canad. Math. Bull.*, 33 (1990), 247–252.

[2] — The hexagonal packing lemma and the Rodin Sullivan conjecture. *Trans. Amer. Math. Soc.*, 343 (1994), 157–167.

[3] BOWERS, P. L., The upper Perron method for labeled complexes with applications to circle packings. *Math. Proc. Cambridge Philos. Soc.*, 114 (1993), 321–345.

[4] COLIN DE VERDIÈRE, Y., Un principe variationnel pour les empilements de cercles. *Invent. Math.*, 104 (1991), 655–669.

[5] DOYLE, P. G., HE, Z.-X. & RODIN, B., Second derivatives of circle packings and conformal mappings. *Discrete Comput. Geom.*, 11 (1994), 35–49.

[6] HE, Z.-X., An estimate for hexagonal circle packings. *J. Differential Geom.*, 33 (1991), 395–412.

[7] — Rigidity of infinite disk patterns. Preprint, 1996.

[8] HE, Z.-X. & RODIN, B., Convergence of circle packings of finite valence to Riemann mappings. *Comm. Anal. Geom.*, 1 (1993), 31–41.

[9] HE, Z.-X. & SCHRAMM, O., On the convergence of circle packings to the Riemann map. *Invent. Math.*, 125 (1996), 285–305.

[10] KOEBE, P., Kontaktprobleme der konformen Abbildung. *Ber. Verh. Sächs. Akad. Wiss. Leipzig, Math.-Phys. Kl.*, 88 (1936), 141–164.

[11] LEHTO, O., *Univalent Functions and Teichmüller Spaces.* Graduate Texts in Math., 109. Springer-Verlag, New York, 1987.

[12] MARDEN, A. & RODIN, B., On Thurston's formulation and proof of Andreev's theorem, in *Computational Methods and Function Theory* (Valparaiso, 1989), pp. 103–115. Lecture Notes in Math., 1435. Springer-Verlag, Berlin, 1990.

[13] RODIN, B., Schwarz's lemma for circle packings. *Invent. Math.*, 89 (1987), 271–289.
[14] — Schwarz's lemma for circle packings, II. *J. Differential Geom.*, 30 (1989), 539–554.
[15] RODIN, B. & SULLIVAN, D., The convergence of circle packings to the Riemann mapping. *J. Differential Geom.*, 26 (1987), 349–360.
[16] SCHRAMM, O., Circle patterns with the combinatorics of the square grid. *Duke Math. J.*, 86 (1997), 347–389.
[17] STEPHENSON, K., A probablistic proof of Thurston's conjecture on circle packings. Preprint.
[18] THURSTON, W. P., *The Geometry and Topology of 3-manifolds.* Princeton Univ. Notes, Princeton, NJ, 1982.
[19] — The finite Riemann mapping theorem. Unpublished talk given at the *International Symposium in Celebration of the Proof of the Bieberbach Conjecture* (Purdue University, 1985).

ZHENG-XU HE
Department of Mathematics
University of California, San Diego
La Jolla, CA 92093-0112
U.S.A.
zhe@math.ucsd.edu

ODED SCHRAMM
Department of Mathematics
The Weizmann Institute of Science
Rehovot 76100
Israel
schramm@wisdom.weizmann.ac.il

Received December 13, 1996

GAFA, Geom. funct. anal.
Vol. 10 (2000) 266 – 306
1016-443X/00/020266-41 $ 1.50+0.20/0

© Birkhäuser Verlag, Basel 2000

GAFA Geometric And Functional Analysis

EMBEDDINGS OF GROMOV HYPERBOLIC SPACES

M. Bonk and O. Schramm

Abstract

It is shown that a Gromov hyperbolic geodesic metric space X with bounded growth at some scale is roughly quasi-isometric to a convex subset of hyperbolic space. If one is allowed to rescale the metric of X by some positive constant, then there is an embedding where distances are distorted by at most an additive constant.

Another embedding theorem states that any δ-hyperbolic metric space embeds isometrically into a complete geodesic δ-hyperbolic space.

The relation of a Gromov hyperbolic space to its boundary is further investigated. One of the applications is a characterization of the hyperbolic plane up to rough quasi-isometries.

1 Introduction

The study of Gromov hyperbolic spaces has been largely motivated and dominated by questions about Gromov hyperbolic groups. This paper studies the geometry of Gromov hyperbolic spaces without reference to any group or group action.

One of our main theorems is

Embedding Theorem 1.1. *Let X be a Gromov hyperbolic geodesic metric space with bounded growth at some scale. Then there exists an integer n such that X is roughly similar to a convex subset of hyperbolic n-space $\mathbb{H}^n$.*

The precise definitions appear in the body of the paper, but now we briefly discuss the meaning of the theorem. The condition of bounded growth at some scale is satisfied, for example, if X is a bounded valence graph, or a Riemannian manifold with bounded local geometry. Saying

M. Bonk is partially supported by TMR fellowship ERBFMBICIT 961462. O. Schramm holds the Sam and Ayala Zacks Chair.

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_10, © Springer Science+Business Media, LLC 2011

that X is roughly similar to a metric space Y means that there is a map $f : X \to Y$ and constants $k, \lambda > 0$, such that

$$\left|\lambda d_X(x,y) - d_Y(f(x),f(y))\right| \leqslant k\,, \qquad \forall x, y \in X\,,$$

and $\sup_{y\in Y} d_Y(y, f(X)) \leqslant k$. This is much stronger than saying that X and Y are roughly quasi-isometric. (Roughly quasi-isometric is the term we use for what is elsewhere "quasi-isometric" or "roughly-isometric".)

It follows in particular that complex hyperbolic m-space embeds in $\mathbb{H}^n$ for some $n = n(m)$. It would be interesting to determine or estimate the smallest possible $n(m)$.

The converse of the theorem is straightforward: a geodesic metric space X that is roughly similar to a convex subset W of $\mathbb{H}^n$ has bounded growth at some scale and is Gromov hyperbolic. This is even known to be true if you weaken the assumption that X is roughly similar to W and instead assume that X is roughly quasi-isometric to W.

The outline of the proof of Theorem 1.1 is quite simple. We verify that the Gromov boundary ∂X has finite Assouad dimension, and then Assouad's theorem [A] implies that (if we rescale the metric of X by a constant) ∂X has a bilipschitz embedding in $\mathbb{R}^{n-1}$, for some n. As $\mathbb{R}^{n-1} \cup \{\infty\}$ is the boundary of $\mathbb{H}^n$, further assuming that the set of all geodesics is cobounded in X (that is, there is a finite upper bound for the distance of any point in X to the union of the geodesics in X), the embedding of ∂X in $\partial\mathbb{H}^n$ induces an embedding of X in $\mathbb{H}^n$. The general statement then easily reduces to this case.

Another embedding theorem presented here says that any δ-hyperbolic metric space X embeds isometrically in a complete geodesic δ-hyperbolic metric space. This allows us to generalize statements about geodesic Gromov hyperbolic metric spaces to the non-geodesic setting.

It is known that a rough quasi-isometry $f : X \to Y$ between Gromov hyperbolic metric spaces induces a map $\partial f : \partial X \to \partial Y$, which is bi-hölder and quasiconformal. However, it turns out that ∂f satisfies a stronger condition. It is what is called below a power quasisymmetry. Conversely, under mild conditions on X and Y, any power quasisymmetry from ∂X to ∂Y induces a rough quasi-isometry from X to Y.

Previously, F. Paulin [P] has obtained a characterization of X from ∂X under the assumption that there is a group acting isometrically and co-compactly on X. His methods and results are different.

Given any bounded metric space Z, we define a Gromov hyperbolic space $\mathrm{Con}(Z)$ such that $\partial\,\mathrm{Con}(Z) = Z$. This is quite similar to a construc-

tion by Gromov. It turns out that under mild assumptions $\mathrm{Con}(\partial X)$ is roughly quasi-isometric to X, when X is Gromov hyperbolic.

As an application of the machinery developed, some results about two dimensional Gromov hyperbolic spaces are obtained. We give characterization of $\mathbb{H}^2$ up to rough quasi-isometries.

It is also shown that a planar Gromov hyperbolic graph with bounded vertex degree and finitely many vertices in any compact subset of the plane is roughly quasi-isometric to a convex subset of $\mathbb{H}^2$, provided that there is a finite upper bound for the distance of a vertex in G to the union of all geodesics in G.

The paper is organized as follows. After reviewing some notions of coarse geometry in section 2, and some basics of Gromov hyperbolic spaces in section 3, we show in section 4 that each Gromov hyperbolic space can be embedded in a complete geodesic Gromov hyperbolic space without changing the hyperbolicity constant δ. In section 5 we collect some auxiliary results about Gromov hyperbolic spaces that are needed later on.

In sections 6–8 we discuss the boundary functor $X \mapsto \partial X$, the convex hull functor $Z \mapsto \mathrm{Con}(Z)$, and their relation. This is the technical core of the paper. We give particular emphasis to the correspondence of maps between Gromov hyperbolic spaces and maps between their boundaries.

In section 9 we recall the notion of Assouad dimension, and establish a sufficient condition for a Gromov hyperbolic space to have a boundary with finite Assouad dimension.

The last two sections give applications of the methods developed in earlier sections. In section 10 Theorem 1.1 is proved, and in section 11 the real hyperbolic plane is characterized up to rough quasi-isometry among geodesic metric spaces.

Acknowledgments. The authors would like to express thanks to Yoav Benyamini, Jianguo Cao, Juha Heinonen, Joram Lindenstrauss, Geoff Mess, Frédéric Paulin, Gideon Schechtman, Richard Schwartz, Zlil Sela and Jussi Väisälä for helpful advice and interesting discussions. The present study of Gromov hyperbolic spaces was partly motivated by the work of Phil Bowers and Ken Stephenson on circle packings whose nerve is a planar hyperbolic graph.

Our collaboration started at the Oberwolfach function theory meeting in the spring of 1996. Thanks to the Mathematisches Forschungsinstitut Oberwolfach for its hospitality. Part of this research was completed while the first author was staying at University of Jyväskylä in Spring 1997.

He is pleased to acknowledge the hospitality of the local Department of Mathematics.

2 Coarse Geometry Definitions

Rough quasi-isometries. Let (X, d) be a metric space. If the metric d is fixed, we denote the distance $d(x, y)$ of two points $x, y \in X$ also by $|x - y|$. The diameter of a set $A \subset X$ is denoted by $\operatorname{diam}(A)$, and the distance of two set $A, B \subset X$ by $\operatorname{dist}(A, B)$.

A set $A \subset X$ is *k-cobounded* (in X) for $k \geqslant 0$, if every point $x \in X$ has distance at most k from A. If A is k-cobounded for some $k \geqslant 0$ we say that A is *cobounded*. In the same way, we will drop the parameters of a notion if the values of the parameters do not matter.

Let $f : X \to Y$ be a map (not necessarily continuous) between metric spaces X and Y, and let $\lambda \geqslant 1$ and $k \geqslant 0$ be constants. Suppose that $f(X)$ is k-cobounded in Y. If, in addition, for all $x, y \in X$,

$$\lambda^{-1}|x - y| - k \leqslant \big|f(x) - f(y)\big| \leqslant \lambda|x - y| + k\,, \tag{2.1}$$

then f is called a *(λ, k)-rough quasi-isometry*. If

$$\lambda|x - y| - k \leqslant \big|f(x) - f(y)\big| \leqslant \lambda|x - y| + k\,, \tag{2.2}$$

then f is a *(λ, k)-rough similarity*. If

$$|x - y| - k \leqslant \big|f(x) - f(y)\big| \leqslant |x - y| + k\,, \tag{2.3}$$

then f is a *k-rough isometry*.

If $f : X \to Y$ is a map that satisfies (2.1) for all $x, y \in X$ without $f(X)$ being cobounded in Y, then we call f a (λ, k)-rough quasi-isometry of X *into* Y or a *rough quasi-isometric embedding*. We use similar language for a map that satisfies (2.2) or (2.3), but does not necessarily have cobounded image in its target space. Each rough isometry is also a rough similarity, and each rough similarity is a rough quasi-isometry. If f is bijective and satisfies (2.1) for $k = 0$, then it is called a *λ-quasi-isometry* or *λ-bilipschitz* map.

The classes of maps just introduced have appeared under various names in the literature. Here and in the following we have tried to adopt the following system for our terminology. The word "rough" refers to change of an equality, inequality, etc., defining a notion by an additive constant, and the word "quasi" to a change by a multiplicative constant. Keeping this in mind, the above terms are more or less self-explanatory. We sometimes deviate from this system, if we feel that a certain term is well-established.

Two maps $f, g : X \to Y$ are *roughly equivalent*, written $f \approx g$, if there exists some constant $k \geqslant 0$ such that $|f(x) - g(x)| \leqslant k$ for $x \in X$. This defines an equivalence relation on the set of mappings from X to Y. The equivalence class of a map f is called the *rough mapping class* of f. If $f_1, f_2 : X \to Y$ and $g_1, g_2 : Y \to Z$ are rough quasi-isometric embeddings, then $f_1 \approx f_2$ and $g_1 \approx g_2$ imply $g_1 \circ f_1 \approx g_2 \circ f_2$. Compositions of rough quasi-isometric embeddings are maps of the same type. The same statement is true for embeddings by rough similarities or rough isometries. From this it follows that it is possible to define categories, where the objects are metric spaces and the morphisms are rough mapping classes of our various types of embeddings.

A *rough-inverse* of a rough quasi-isometry $f : X \to Y$ is a rough quasi-isometry $g : Y \to X$ such that $g \circ f \approx \mathrm{id}_X$ and and $f \circ g \approx \mathrm{id}_Y$. Every rough quasi-isometry has a rough inverse. If in addition the rough quasi-isometry is a rough similarity or a rough isometry, then the rough inverse is also a rough similarity or a rough isometry, respectively.

We say that two metric spaces X and Y are *roughly quasi-isometric*, written $X \sim Y$, if there exists a rough quasi-isometry $f : X \to Y$. This is an equivalence for the class of metric spaces. In a similar way we define that X and Y are *roughly similar*, $X \simeq Y$, or *roughly isometric*, $X \cong Y$. These are also equivalences. The equivalence $X \cong Y$ implies $X \simeq Y$, and $X \simeq Y$ implies $X \sim Y$.

Geodesics and rough geodesics in metric spaces. Let X be a metric space. A *path* in X is a map $\gamma : I \to X$, where I is an interval in $\mathbb{R}$. We will also denote by γ the image set of the path γ. A path γ *joins* two points x and y in X, if $I = [a, b]$, $\gamma(a) = x$ and $\gamma(b) = y$. The path γ is called a *ray starting*, or *emanating*, from the *initial point* x, if $I = [0, \infty)$ and $\gamma(0) = x$. If $\gamma : I \to X$ is a (λ, k)-rough quasi-isometry of I into X, then γ is called a (λ, k)-*roughly quasi-isometric path*.

A *geodesic*, a *geodesic ray* or a *geodesic segment* in X is an isometry $\gamma : I \to X$ into X, where I is $\mathbb{R}$, $[0, \infty)$ or a closed segment in $\mathbb{R}$, respectively. Similarly, if $\gamma : I \to X$ is a k-rough isometry of $I = \mathbb{R}$ into X, we call it a k-*rough geodesic*. We also speak of k-*roughly geodesic rays* or k-*roughly geodesic segments*, if $I = [0, \infty)$ or I is a closed segment in $\mathbb{R}$, respectively.

We call subsets of X geodesics, rough geodesics, etc., if they are images of a path in the corresponding class.

A *geodesic metric space* is a metric space X such that for any two points $x, y \in X$ there is a geodesic segment joining x and y. We denote

any geodesic segment with endpoints x, y by $[x, y]$. Note that this notation is ambiguous, since $[x, y]$ may not be unique.

The space X is *k-roughly geodesic*, if for every $x, y \in X$ there exists a k-roughly geodesic segment joining x and y. It is *k-almost geodesic*, if for every $x, y \in X$ and every $t \in [0, |x - y|]$, there is some $z \in X$ with $||x - z| - t| \leqslant k$ and $||y - z| - (|x - y| - t)| \leqslant k$.

The reader, who might be content with studying only geodesic metric spaces, may question the motivation for introducing the concepts of almost geodesic and roughly geodesic. It turns out that the setting of roughly geodesic spaces is more fitting with the philosophy of Gromov hyperbolic spaces, and it is worthwhile to work in that setting. In particular, the spaces $\mathrm{Con}(Z)$, which we will construct and will be important for us, are roughly geodesic, and it will be unnecessarily cumbersome to modify the construction to make them truly geodesic. Remark 5.3 below further discusses this issue.

Constants. In the following, we will encounter various constants whose precise values usually does not matter.

The letter C will denote positive numerical constants. Similarly, $C(a, b, c, \dots)$ will denote universal positive functions of the parameters $a, b, c, \dots$. Sometimes we write $C = C(a, b, c, \dots)$ to emphasize the parameters on which $C(a, b, c, \dots)$ depends and abbreviate $C(a, b, c, \dots)$ to C. We will use the same letter for possibly different constants and functions.

The notation $O(1)$ will be used for real numbers whose absolute value is bounded by a positive numerical constant that can be computed explicitly in principle. Similarly, we use the notation $O_{a,b,c,\dots}(1)$ for real numbers whose absolute value is bounded by a constant which can be chosen in a way only depending on the parameters $a, b, c, \dots$.

3 Some Basics of Gromov Hyperbolic Spaces

We here review some definitions and elementary facts concerning Gromov hyperbolic spaces. For a more detailed account of this material, the reader is referred to [CDP], [GH] or [Gr].

The definition of Gromov hyperbolicity. Let X be a metric space. Given three points $x, y, w \in X$, the *Gromov product* of x and y with respect to the basepoint w is defined as

$$(x|y)_w = \tfrac{1}{2}\big(|x - w| + |y - w| - |x - y|\big)\,.$$

It measures the failure of the triangle inequality to be an equality, and is always nonnegative. Often we will fix a basepoint $w = o \in X$. Then we abbreviate the Gromov product as $(x|y) = (x|y)_o$, and let $|x| = |x - o|$ denote the distance of x from the basepoint. For all $x, y, v, w \in X$ we have

$$\big|(x|y)_v - (x|y)_w\big| \leqslant |v - w| \,. \tag{3.1}$$

This formula is tacitly used in the following when it is convenient to change the basepoint.

Let $a \vee b$ denote the maximum of two numbers $a, b \in \overline{\mathbb{R}}$, and let $a \wedge b$ denote the minimum. The space X is *δ-hyperbolic*, if the following inequality holds for all $x, y, z, w \in X$,

$$(x|z)_w \geqslant (x|y)_w \wedge (y|z)_w - \delta \,. \tag{3.2}$$

This is equivalent to the more symmetric inequality

$$|x - z| + |y - w| \leqslant \big(|x - y| + |z - w|\big) \vee \big(|y - z| + |x - w|\big) + 2\delta \,. \tag{3.3}$$

It turns out that if (3.2) holds for fixed $w \in X$ and arbitrary $x, y, z \in X$, then X is 2δ-hyperbolic. If X is δ-hyperbolic for some $\delta < \infty$, we say that X is *Gromov hyperbolic*.

The Rips thin triangles definition of Gromov hyperbolicity. Let X be a geodesic metric space. We say that X has *δ-thin triangles*, if for every geodesic triangle $[x, y] \cup [y, z] \cup [z, x] \subset X$ and every point $w \in [x, y]$, the distance from w to $[y, z] \cup [z, x]$ is at most δ. There exists some constant $C > 0$ with the following property. If X has δ-thin triangles, then X is $C\delta$-hyperbolic, and if X is δ-hyperbolic, then it has $C\delta$-thin triangles. Hence, in the context of geodesic metric spaces, the condition of having thin triangles provides an alternative definition for Gromov hyperbolicity. This definition was introduced by I. Rips.

The boundary of a Gromov hyperbolic space. Let X be δ-hyperbolic and $w \in X$. A sequence of points $\{x_i\} \subset X$ is said to *converge at infinity*, if

$$\lim_{i,j \to \infty} (x_i|x_j)_w = \infty \,.$$

Two sequences $\{x_i\}, \{y_i\}$ that converge at infinity are *equivalent*, if

$$\lim_{i \to \infty} (x_i|y_i)_w = \infty \,.$$

This defines an equivalence relation for sequences in X converging at infinity. It is easy to see that convergence at infinity of a sequence and equivalence of two sequences does not depend on the choice of the point w. The *boundary* ∂X of X is defined as the set of equivalence classes of sequences converging at infinity.

Let $a, b \in \partial X$. The Gromov product $(a|b)_w$ of a and b with respect to the point $w \in X$ is defined as follows:

$$(a|b)_w = \sup \left\{ \liminf_{i\to\infty} (x_i|y_i)_w \;\middle|\; \{x_i\} \in a\,,\ \{y_i\} \in b \right\}.$$

Similarly, for $a \in \partial X$ and $y \in X$ we define

$$(a|y)_w = (y|a)_w = \sup \left\{ \liminf_{i\to\infty} (x_i|y)_w \;\middle|\; \{x_i\} \in a \right\}.$$

Note that $(a|b)_w \in [0, \infty]$ for $a, b \in X \cup \partial X$, and $(a|b)_w = \infty$ if and only if $a, b \in \partial X$, $a = b$. It can be shown that

$$(a|b)_w - C\delta \leqslant \liminf_{i\to\infty} (x_i|y_i)_w \leqslant \limsup_{i\to\infty} (x_i|y_i)_w \leqslant (a|b)_w + C\delta\,, \qquad \forall\, a, b \in \partial X\,,\ \{x_i\} \in a\,,\ \{y_i\} \in b\,. \quad (3.4)$$

A similar inequality holds if one of the points a, b is in X. Inequality (3.2) generalizes to points in ∂X; more precisely,

$$(x|z)_w \geqslant (x|y)_w \wedge (y|z)_w - C\delta\,, \quad \forall\, w \in X\,,\ x, y, z \in X \cup \partial X\,. \quad (3.5)$$

Geodesic stability. If $A \subset X$ and $B \subset X$ are two subsets of a metric space X, define their *Hausdorff distance* to be

$$d_{\mathrm{Haus}}(A, B) = \Big(\sup_{a\in A} \Big(\inf_{b\in B} |a - b| \Big) \Big) \vee \Big(\sup_{b\in B} \Big(\inf_{a\in A} |a - b| \Big) \Big).$$

Suppose X is a δ-hyperbolic geodesic metric space, and let $\lambda \geqslant 1$, $k \geqslant 0$. Then there exists a constant $C = C(\delta, \lambda, k)$ with the following property. If $\gamma : [a, b] \to X$ is a (λ, k)-roughly quasi-isometric path, $x = \gamma(a)$ and $y = \gamma(b)$ are the endpoints of γ, and $[x, y]$ is any geodesic segment joining x and y, then $d_{\mathrm{Haus}}(\gamma, [x, y]) \leqslant C$.

4 Embedding into a Geodesic Metric Space

Theorem 4.1. *Let X be a δ-hyperbolic metric space. Then there is an isometric embedding $i : X \to Y$ of X into a complete δ-hyperbolic geodesic metric space Y.*

The following lemma provides the basic step in the construction of Y and i.

Lemma 4.2. *Let X be a δ-hyperbolic metric space, let $a, b \in X$, and set $D = |a - b|$. Then there is an isometric embedding i of X into a δ-hyperbolic metric space $(\widehat{X}, d)$ which contains a point m satisfying $d(m, i(a)) = d(m, i(b)) = D/2$.*

The main point is that the hyperbolicity constant δ of $\widehat{X}$ is the same as that of X.

Proof. Assume that there is no point $x \in X$ satisfying $|x-a| = |x-b| = D/2$. Let $\widehat{X} = X \cup \{m\}$, where m is a new point not appearing in X. When $x, y \in X$, define $d(x,y) = |x-y|$, and for any $x \in X$, let

$$d(x,m) = d(m,x) = \tfrac{D}{2} + \sup_{w \in X} \big(|x-w| - |a-w| \vee |b-w| \big) . \quad (4.1)$$

Note that this is always finite. Finally, let $d(m,m) = 0$.

We now verify that d is a metric on $\widehat{X}$. Let $x, y \in X$. When substituting a and b for w in (4.1), we find

$$\begin{aligned} d(x,m) &\geqslant D/2 + \big(|x-a| - |a-a| \vee |b-a| \big) \vee \big(|x-b| - |a-b| \vee |b-b| \big) \\ &= |x-a| \vee |x-b| - D/2 . \end{aligned} \quad (4.2)$$

Since $D \leqslant |x-a| + |x-b|$, and there is no point in X satisfying $|x-a| = |x-b| = D/2$, the right hand side of (4.2) must be positive. Consequently, $d(z,z') > 0$ when $z \neq z'$ are in $\widehat{X}$. Now (4.2) is used to verify one case in the triangle inequality for d.

$$\begin{aligned} d(x,m) + d(m,y) &\geqslant |x-a| \vee |x-b| + \sup_{w \in X} \big(|y-w| - |a-w| \vee |b-w| \big) \\ &\geqslant |x-a| \vee |x-b| + \big(|y-x| - |a-x| \vee |b-x| \big) \\ &= |x-y| = d(x,y) . \end{aligned}$$

Another case of the triangle inequality is proved as follows,

$$\begin{aligned} d(m,x) + d(x,y) &= D/2 + \sup_{w \in X} \big(|x-w| - |a-w| \vee |b-w| \big) + |x-y| \\ &\geqslant D/2 + \sup_{w \in X} \big(|y-w| - |a-w| \vee |b-w| \big) \\ &= d(m,y) . \end{aligned}$$

Since the triangle inequality for d holds for any three points in X, it now follows that it holds for any three points in $\widehat{X}$, and hence $\widehat{X}$ is a metric space.

We now verify that $(\widehat{X}, d)$ is δ-hyperbolic. For this, it is enough to prove the inequality (3.3) for m and points $x, y, z \in X$. For an arbitrary $w \in X$, we have

$$|x-w| + |y-z| \leqslant \big(|y-w| + |x-z| \big) \vee \big(|z-w| + |x-y| \big) + 2\delta .$$

Therefore,

$$d(m,x) + d(y,z) = D/2 + \sup_{w \in X} \big(|x-w| - |a-w| \vee |b-w| \big) + |y-z|$$

$$
\begin{aligned}
&\leqslant D/2 + 2\delta \\
&\qquad + \sup_{w\in X} \big((|y-w| + |x-z|) \vee (|z-w| + |x-y|) - |a-w| \vee |b-w|\big) \\
&= \big(D/2 + |x-z| + \sup_{w\in X}(|y-w| - |a-w| \vee |b-w|)\big) \\
&\qquad \vee \big(D/2 + |x-y| + \sup_{w\in X}(|z-w| - |a-w| \vee |b-w|)\big) + 2\delta \\
&= \big(d(m,y) + d(x,z)\big) \vee \big(d(m,z) + d(x,y)\big) + 2\delta\,,
\end{aligned}
$$

which verifies that $(\widehat{X}, d)$ is δ-hyperbolic.

Observe that

$$
\begin{aligned}
d(m,a) &= D/2 + \sup_{w\in X} \big(|a-w| - |a-w| \vee |b-w|\big) \\
&= D/2 + \sup_{w\in X} 0 \wedge \big(|a-w| - |b-w|\big)\,.
\end{aligned}
$$

As b may be chosen for w, this shows that $d(m,a) = D/2$. Similarly, $d(m,b) = D/2$, and the proof is complete. □

Proof of 4.1. Suppose Z is a metric space and $x, y \in Z$. We say that m is a midpoint of x and y if $|x-m| = |y-m| = |x-y|/2$. The space Z has the midpoint property, if for all $x, y \in Y$ there exists a midpoint $m \in Z$. In the following, when a metric space Z_1 admits an isometric embedding into another metric space Z_2, we can think of Z_1 as being contained in Z_2 and write $Z_1 \subset Z_2$.

The first step in the proof is to construct a metric space $X' \supset X$, which has the midpoint property and is δ-hyperbolic.

For a δ-hyperbolic metric space Z and $a, b \in Z$, denote by $Z[a,b]$ the space constructed according to the previous lemma. Let $\phi : \omega \to X \times X$ be a bijective mapping from some ordinal ω onto the set of pairs in X. By transfinite induction we define a metric space $X(\alpha)$ for each ordinal $\alpha \leqslant \omega + 1$ with the property that $X(\alpha) \supset X(\beta)$ if $\alpha > \beta$.

Set $X(0) = X$. When $\alpha = \beta + 1$ is a successor ordinal, let $X(\alpha)$ be $X(\beta)[a,b]$, where $(a,b) = \phi(\beta)$. For limit ordinals α, set $X(\alpha) = \cup_{\beta<\alpha} X(\beta)$. Note that nested unions of δ-hyperbolic spaces are δ-hyperbolic. It follows that $X^* = X(\omega+1)$ is a δ-hyperbolic metric space which contains X and has the property that for any two points in X there exists a midpoint in X^*.

Now define inductively $X_0 = X$, and $X_{i+1} = X_i^*$ for $i \in \mathbb{N}$. Let $X' = \cup_{i\in\mathbb{N}} X_i$. Then X' has the required properties.

To construct a complete metric space $Y \supset X$ which has the midpoint property and is δ-hyperbolic we use transfinite induction again. For this,

let ω_0 be the first uncountable ordinal.

We define a metric spaces $Z(\alpha)$ for each ordinal $\alpha < \omega_0$ with the property that $Z(\alpha) \supset Z(\beta)$ if $\alpha > \beta$. Let $Z(0)$ be the completion of X'. When $\alpha = \beta + 1$ is a successor ordinal, let $Z(\alpha)$ be the completion of $Z(\beta)'$. For limit ordinals α, let $Z(\alpha)$ be the completion of $\cup_{\beta<\alpha} Z(\beta)'$. Note that completions of δ-hyperbolic spaces are δ-hyperbolic. It follows by transfinite induction that for $\alpha < \omega_0$ all spaces $Z(\alpha)$ are δ-hyperbolic, complete and have the property that if $x, y \in Z(\alpha)$, then $Z(\alpha + 1)$ contains a midpoint of x and y.

Define $Y = \cup_{\alpha<\omega_0} Z(\alpha)$. By the above properties of the spaces $Z(\alpha)$, this space contains X, is δ-hyperbolic and has the midpoint property. Furthermore, Y is complete. For if $\{x_i\}$ is a Cauchy-sequence in Y, then each element x_i is contained in some $Z(\alpha_i)$ where $\alpha_i < \omega_0$. Since ω_0 is the first uncountable ordinal, there exits $\beta < \omega_0$ such that $\alpha_i \leqslant \beta$ for all $i \in \mathbb{N}$. Hence the Cauchy sequence $\{x_i\}$ is contained in $Z(\beta)$. So it has a limit in the complete space $Z(\beta) \subset Y$.

The space Y is geodesic. This follows from the midpoint property and the completeness of Y. For if $x, y \in Y$ are arbitrary and $D = |x - y|$, then by induction it can be shown that there exists an isometry ϕ from the set $S = \{kD/2^n : n, k \in \mathbb{N},\ 0 \leqslant k \leqslant 2^n\}$ into Y with $\phi(0) = x$ and $\phi(D) = y$. By the completeness of Y, ϕ extends to an isometry of $[0, D]$ into Y. The image of this isometry is a geodesic segment $[x, y]$. We leave the details to the reader.

Thus, Y has all the required properties and the proof is complete. □

5 Rough Geodesics in Almost Geodesic Metric Spaces

In this section, we collect various geometric statements that will be used later. Almost all of this material has already appeared in the literature, perhaps in slightly different form. Therefore, we will mostly skip the proofs or only outline the main ideas. An important observation that often serves as a guiding principle in proving statements about a δ-hyperbolic space X is the fact that given any n-point set $M \subset X$ there exists a k-rough isometric embedding of M into a metric tree with $k = k(\delta, n)$ (cf. [GH, pp. 33–38]). In this way geometric statements about finite point sets in X can be reduced to an essentially combinatorial problem about the location of a finite point set in a tree.

We start with a lemma that is very useful to estimate distances of points

in a Gromov hyperbolic space. An intuitive interpretation of this statement is given below.

LEMMA 5.1. *Let X be a δ-hyperbolic metric space. Let o be the basepoint of X, let $a, b \in X \cup \partial X$, and let $x, y \in X$. Suppose that for some $c \geqslant 0$ we have $(a|x) \geqslant |x| - c$ and $(b|y) \geqslant |y| - c$. Then*

$$|x - y| = |x| + |y| - 2\big((a|b) \wedge |x| \wedge |y|\,\big) + O_{\delta,c}(1)\,.$$

Proof. On the one hand, (3.5) gives

$$\begin{aligned}(a|b) &\geqslant (a|x) \wedge (x|b) - C\delta \\ &\geqslant (a|x) \wedge (x|y) \wedge (y|b) - C\delta \geqslant |x| \wedge (x|y) \wedge |y| + O_{\delta,c}(1)\,.\end{aligned}$$

Since $(x|y) \leqslant |x| \wedge |y|$, this implies $|x| \wedge (a|b) \wedge |y| \geqslant (x|y) + O_{\delta,c}(1)$.

On the other hand,

$$\begin{aligned}(x|y) &\geqslant (x|a) \wedge (a|y) - C\delta \\ &\geqslant (x|a) \wedge (a|b) \wedge (b|y) - C\delta \geqslant |x| \wedge (a|b) \wedge |y| + O_{\delta,c}(1)\,.\end{aligned}$$

The combination of these two inequalities is

$$(x|y) = |x| \wedge (a|b) \wedge |y| + O_{\delta,c}(1)\,.$$

The lemma follows, because $|x - y| = |x| + |y| - 2(x|y)$. □

A roughly geodesic ray $\gamma : [0, \infty) \to X$ in a Gromov hyperbolic space *converges at infinity*, in the following sense. For each sequence $\{t_i\} \to \infty$, the sequence $\{\gamma(t_i)\}$ converges at infinity, and the equivalence class of this sequence does not depend on the choice of $\{t_i\}$. The equivalence class of $\{\gamma(t_i)\}$ is called the *endpoint* of γ on ∂X, or the *limit* of γ. We will denote it by $\lim_{t\to\infty} \gamma(t)$ or simply by $\lim \gamma$. The endpoints of a rough geodesic $\gamma : \mathbb{R} \to X$ are defined similarly.

Suppose γ_1 and γ_2 are two roughly geodesic rays in a Gromov hyperbolic space X with initial point o and endpoints a and b, respectively. Suppose $a \neq b$. We can apply Lemma 5.1 for points $x = \gamma_1(t)$ and $y = \gamma_2(t)$, $t \geqslant 0$. In this situation the lemma says that the roughly geodesic rays run close to each other as long as $0 \leqslant t \leqslant (a|b)$, but they start to spread at $t = (a|b)$.

PROPOSITION 5.2. *Let X be a δ-hyperbolic, k-almost geodesic metric space. Then there exists a constant $k' = k'(\delta, k)$ with the following properties:*

(1) *For all points $x, y \in X$, there exists a k'-roughly geodesic segment $\gamma : [a, b] \to X$ with $\gamma(a) = x$ and $\gamma(b) = y$.*
(2) *For all points $x \in X$, $y \in \partial X$, there exists a k'-roughly geodesic ray $\gamma : [0, \infty) \to X$ with $\gamma(0) = x$ and $\lim_{t\to\infty} \gamma(t) = y$.*

(3) *For all points $x, y \in \partial X$, $x \neq y$, there exists a k'-rough geodesic $\gamma : \mathbb{R} \to X$ with $\lim_{t\to-\infty} \gamma(t) = x$ and $\lim_{t\to\infty} \gamma(t) = y$.*

Proof. All three statements have similar proofs, so we prove only (2). For each $t \geqslant 0$, let $z_t \in X$ be some point satisfying $(y|z_t) \geqslant t$ and let $y_t \in X$ satisfy $|y_t| \leqslant t + k$ and $|z_t - y_t| \leqslant |z_t| - t + k$. Similarly, let $x_t \in X$ satisfy $|x_t| \leqslant t + k$ and $|x - x_t| \leqslant |x| - t + k$. Let $a := (x|y)$, $b := |x|$, and set $\gamma(s) := x_{b-s}$ for $s \in [0, b - a]$ and $\gamma(s) := y_{2a-b+s}$ for $s > b - a$. It is left to the reader to verify that γ satisfies the requirements. □

It follows from the definitions that each roughly geodesic metric space is also an almost geodesic metric space. Part (1) of the previous proposition implies the converse of this statement for Gromov hyperbolic spaces.

REMARK 5.3. The reader may wonder why we introduced the concepts of almost geodesic spaces, roughly geodesic rays, etc. It seems that less baggage of terminology would be needed if we had restricted ourselves to geodesic metric spaces. Even without being able to prove statements in the utmost generality, the essential features and ideas of the theory might be preserved.

Parts (2) and (3) of Proposition 5.2 can serve as a justification for introducing the concepts of roughly geodesic rays and rough geodesics. When we assume that X is geodesic, then the existence of geodesic rays and geodesics in parts (2) and (3) will not be true in general, unless we make additional assumptions, for example, that X is proper (closed balls are compact). In addition to that, the concepts of almost and roughly geodesic spaces, and the corresponding notions of roughly geodesic segments, etc., are better adapted to the philosophy of Gromov hyperbolic spaces: Bounded distortions do not matter, since they do not affect the structure of the space on large scales. While the property of being a geodesic metric space is very sensitive to these distortions, the property of being almost and roughly geodesic is more robust, and preserved under rough isometries, for example.

In spite of this, sometimes it does make sense to restrict oneself to geodesic metric spaces. A handy tool for generalizing results about Gromov hyperbolic geodesic metric spaces to the general case is Theorem 4.1. For example, geodesic stability generalizes in the following way.

PROPOSITION 5.4. *Let X be a δ-hyperbolic metric space, and γ_1, γ_2 be (λ, k)-roughly quasi-isometric paths in X with the same endpoints. Then $d_{\mathrm{Haus}}(\gamma_1, \gamma_2) = O_{\delta,\lambda,k}(1)$.*

Proof. Embed X in a δ-hyperbolic geodesic metric space Y. If x and y are the common endpoints of γ_1, γ_2, and $[x, y]$ is any geodesic segment in Y joining x and y, then the geodesic stability of Y implies $d_{\mathrm{Haus}}(\gamma_1, [x, y]) = O_{\delta,\lambda,k}(1)$ and $d_{\mathrm{Haus}}(\gamma_2, [x, y]) = O_{\delta,\lambda,k}(1)$. The claim follows from the triangle inequality. □

PROPOSITION 5.5. *Suppose X and Y are δ-hyperbolic k-almost geodesic metric spaces, and $f : X \to Y$ is a (λ, c)-rough quasi-isometry of X into Y. Then for all $x, y, z, w \in X$*

(1) $\lambda^{-1}(x|y)_w - C \leqslant (x'|y')_{w'} \leqslant \lambda(x|y)_w + C$,
(2) *and if* $(x|z)_w - (x|y)_w \geqslant 0$, *then*

$\lambda^{-1}\big((x|z)_w - (x|y)_w\big) - C \leqslant (x'|z')_{w'} - (x'|y')_{w'} \leqslant \lambda\big((x|z)_w - (x|y)_w\big) + C$.

Here, image points under f are indicated by a prime, and $C = C(\delta, k, \lambda, c)$.

Statement (1) will be used to define the map $\partial f : \partial X \to \partial Y$ for a rough quasi-isometry $f : X \to Y$ (cf. Prop. 6.3). From Corollary 6.4, which is a generalization of (2), it will follow that ∂f is a power quasisymmetry (cf. Thm. 6.5).

Note that if $(x|z)_w - (x|y)_w \leqslant 0$ in (2), we get a similar inequality with the roles of λ and λ^{-1} reversed. To see this exchange y and z. This shows that (2) holds under the slightly weaker assumption $(x|z)_w - (x|y)_w \geqslant -C\delta$. We will need this improved version of (2) in the proof of Corollary 6.4.

Proof. The statement is essentially Prop. 15 on pp. 89–90 in [GH]. We will only outline the main idea of the proof. In order to prove (2) ((1) is just a special case putting $y = w$) we note that the expression $(x|z)_w - (x|y)_w$ is roughly the distance between the rough center of the triangle $[x, z, w]$ and the rough center of the triangle $[x, y, w]$. From Proposition 5.4 it follows that the image of the rough center of $[x, z, w]$ is within bounded distance to the rough center of $[x', z', w']$, and similarly for $[x, y, w]$. The proposition follows. □

A metric space X is called *k-visual with respect to $o \in X$*, if each point in X lies on a k-roughly geodesic ray emanating from o. We call X *k-visual*, if it is k-visual with respect to some point $o \in X$.

It is easy to see that if there exists a point $o \in X$ such that the union of the k_1-roughly geodesic rays from o is k_2-cobounded, then X is k-visual with $k = k(k_1, k_2)$. The following proposition gives an improved version of this statement for Gromov hyperbolic spaces.

PROPOSITION 5.6. *Let X be a δ-hyperbolic metric space with basepoint o. Suppose that the union of all (λ, c)-roughly quasi-isometric rays emanating*

from o is d-cobounded in X. Then X is k-visual with respect to o and k-roughly geodesic, where $k = k(\delta, \lambda, c, d)$.

The proposition implies in particular that every visual Gromov hyperbolic space is roughly geodesic.

Proof. Since the proof only employs standard techniques, we will only outline the main ideas and leave the details to the reader.

It is not hard to show that if $\gamma : [0, \infty) \to X$ is a (λ, c)-rough quasi-isometry into some δ-hyperbolic space, then there exists a map $s : [0, \infty) \to [0, \infty)$ with $s(0) = 0$ such that $\gamma' = \gamma \circ s$ is a $C(\delta, \lambda, c)$-rough isometry into X and $d_{\mathrm{Haus}}(\gamma, \gamma') = O_{\delta,\lambda,c}(1)$. In other words, a roughly quasi-isometric ray admits a "rough reparameterization" to a roughly geodesic ray.

From this statement it follows that under the assumptions of the proposition the union of all k-roughly geodesic rays emanating from o is all of X for appropriate $k = k(\delta, \lambda, c, d)$.

In order to show the second statement let $x, y \in X$ be arbitrary. By the first part of the proof, there exist $C(\delta, \lambda, c, d)$-roughly geodesic rays γ_1, γ_2 emanating from o such that $x \in \gamma_1$ and $y \in \gamma_2$. Now a $C(\delta, \lambda, c, d)$-roughly geodesic segment joining x and y can be found by glueing together appropriate pieces of γ_1 and γ_2. □

6 The Boundary Functor ∂

The relation of a Gromov hyperbolic space to its (Gromov) boundary ∂X has been extensively studied (the definitions are given below). Suppose that $f : X \to Y$ is rough quasi-isometry of Gromov hyperbolic geodesic spaces. It is well known that there is an induced map $\partial f : \partial X \to \partial Y$, and ∂f is Hölder and quasiconformal. Indeed ∂f satisfies a stronger condition, namely, it is a power quasisymmetry. This observation is used to show that under certain mild conditions X is determined by the quasisymmetry class of ∂X up to rough quasi-isometries. F. Paulin [P] obtained a different characterization of X from ∂X, but it seems that for his characterization one must assume that there is a group acting co-compactly on X.

The next goal is to identify the precise relevant structure on ∂X that is induced by the metric on X. The correct structure would depend on how seriously the metric on X is taken. More precisely, if you are interested only in properties of X that are invariant under rough quasi-isometries, the corresponding structure on ∂X would be coarser than if you were interested in X up to rough isometries.

DEFINITION. Let X and Y be two metric spaces, $f : X \to Y$ be a bijection and $\alpha > 0$, $\lambda \geqslant 1$.

(1) The map f is an (α, λ)-*snowflake map* if for all $x, y \in X$

$$\lambda^{-1}|x-y|^{\alpha} \leqslant \big|f(x)-f(y)\big| \leqslant \lambda|x-y|^{\alpha}\,.$$

(2) The map f is an (α, λ)-*quasisymmetry* if for all distinct points $x, y, z \in X$

$$\frac{|f(x)-f(z)|}{|f(x)-f(y)|} \leqslant \eta_{\alpha,\lambda}\left(\frac{|x-z|}{|x-y|}\right).$$

Here

$$\eta_{\alpha,\lambda}(t) = \begin{cases} \lambda t^{1/\alpha} & \text{for} \quad 0 < t < 1\,, \\ \lambda t^{\alpha} & \text{for} \quad 1 \leqslant t\,. \end{cases}$$

It seems that the class of maps in (1) has not been given a name in the literature. We call them snowflake maps, because these maps behave similar as the map giving the parameterization of the well-known von Koch snowflake curve.

In contrast to this, quasisymmetries have appeared in the literature before (cf. [TuV1], [V]). Our usage of this term slightly differs from the common one. In general, a map is called a quasisymmetry if the above condition holds for some increasing homeomorphism $\eta : (0, \infty) \to (0, \infty)$. So our notion of quasisymmetry is stronger. To emphasize this distinction we call a map a *power quasisymmetry*, if it is an (α, λ)-quasisymmetries for some $\alpha > 0$ and $\lambda \geqslant 1$. For some spaces, e.g., connected spaces, quasisymmetries are always power quasisymmetries (cf. [TuV1, Cor. 3.12]).

Every snowflake map is a power quasisymmetry. It is straightforward to check that inverse maps and compositions of snowflake maps are again snowflake maps. The same statement is true for power quasisymmetries. In fact, the inverse of an (α, λ)-quasisymmetry is an $(\alpha, C(\lambda, \alpha))$-quasisymmetry. Quasisymmetries are homeomorphisms.

Two metrics d_1 and d_2 on a space X are called *bilipschitz equivalent*, *snowflake equivalent* or *quasisymmetrically equivalent* if the identity map $\mathrm{id}_X : (X, d_1) \to (X, d_2)$ is bilipschitz, a snowflake map or a quasisymmetry, respectively. This defines equivalence relations for the metrics on a space X.

Following is the standard construction of the metrics on ∂X, where X is a Gromov hyperbolic space. If $x, y \in \partial X, w \in X, \epsilon > 0$, let

$$d_{\partial X,w,\epsilon}(x,y) = d_{w,\epsilon}(x,y) = \inf\left\{\sum_{i=1}^{n} e^{-\epsilon(x_{i-1}|x_i)_w}\right\}, \tag{6.1}$$

where the infimum extends over all finite sequences $x = x_0, x_1, x_2, \ldots, x_n = y$ in ∂X. Here the convention $e^{-\infty} = 0$ is understood.

LEMMA 6.1. *There is some constant $\epsilon_0 > 0$ with the following property. If X is δ-hyperbolic and $\epsilon\delta \leqslant \epsilon_0$, then*

$$\tfrac{1}{2} e^{-\epsilon(x|y)_w} \leqslant d_{w,\epsilon}(x,y) \leqslant e^{-\epsilon(x|y)_w}\,, \quad \forall\, x, y \in \partial X\,. \tag{6.2}$$

For the proof, see, for example, [GH, Ch. 7].

This lemma leads to the following definitions.

DEFINITIONS. The canonical gauge $\mathcal{G}(X)$ on ∂X is the set of all metrics of the form $d = d_{w,\epsilon}$. We say that (ϵ_1, d_1) is *B-equivalent* to (ϵ_2, d_2) if there is a constant $c > 0$ such that

$$c^{-1} d_2^{\epsilon_1} \leqslant d_1^{\epsilon_2} \leqslant c\, d_2^{\epsilon_1}\,.$$

A *B-structure* on Z is an equivalence class of this equivalence relation.

The lemma tells us that $(\epsilon, d_{w,\epsilon})$ and $(\epsilon', d_{w',\epsilon'})$ are B-equivalent when $w, w' \in X$ and $\epsilon, \epsilon' > 0$ satisfy $\epsilon\delta \leqslant \epsilon_0$ and $\epsilon'\delta \leqslant \epsilon_0$. Therefore, when X is Gromov hyperbolic, there is an associated B-structure on ∂X. It is called the *canonical B-structure* on ∂X.

Mostly, it is convenient to work with a fixed metric in the canonical gauge and to suppress the ambiguity in choosing this metric. We often abbreviate such a metric by $d_{\partial X}$.

PROPOSITION 6.2. *The boundary ∂X of any Gromov hyperbolic space X is bounded and complete.*

Note that boundedness and completeness are properties that do not depend on the particular choice of the metric $d_{\partial X}$ in $\mathcal{G}(X)$. There are situations where ∂X is not compact. For example, when X is infinite dimensional hyperbolic space, ∂X is homeomorphic to the unit sphere in Hilbert space.

Proof. The boundedness of ∂X follows from the fact that the Gromov product is nonnegative.

Fix a basepoint $o \in X$. Let $\{y_i\} \subset \partial X$ be a $d_{\partial X}$-Cauchy sequence for some metric in the canonical gauge. Then $\lim_{i,j\to\infty}(y_i|y_j) = \infty$. There exists a sequence $\{x_i\} \subset X$ with $\lim_{i\to\infty}(x_i|y_i) = \infty$. Then $\{x_i\}$ converges at infinity. If $y \in \partial X$ is the equivalence class of $\{x_i\}$, then $\lim_{i\to\infty} y_i = y$. □

We will now discuss the functorial properties of $X \mapsto \partial X$.

PROPOSITION 6.3. *Suppose X, Y, Z are Gromov hyperbolic almost geodesic metric spaces, and $f : X \to Y$, $g : Y \to Z$ are roughly quasi-isometric embeddings.*

(1) *If $\{x_i\} \subset X$ converges at infinity, then $\{f(x_i)\} \subset Y$ converges at infinity. If $\{x_i\}$ and $\{y_i\}$ are equivalent sequences in X converging at infinity, then $\{f(x_i)\}$ and $\{f(y_i)\}$ are also equivalent.*

(2) *If $a \in \partial X$ and $\{x_i\} \in a$, let $b \in \partial Y$ be the equivalence class of $\{f(x_i)\}$ and define $\partial f(a) = b$. Then $\partial f : \partial X \to \partial Y$ is well-defined. Moreover, $\partial(g \circ f) = \partial g \circ \partial f$.*

(3) *If $f_1, f_2 : X \to Y$ are rough quasi-isometric embeddings and $f_1 \approx f_2$, then $\partial f_1 = \partial f_2$.*

(4) *The map ∂f is injective. If the image of f is cobounded in Y, i.e., if f is a rough quasi-isometry, then ∂f is a bijection.*

Proof. The statements in (1)–(3) immediately follow from the definitions and (1) of Proposition 5.5. For the first part of (4) note that if $\{x_i\}$ and $\{y_i\}$ are sequences in X converging at infinity such that $\{f(x_i)\}$ and $\{f(y_i)\}$ are equivalent, then $\lim_{i\to\infty}(f(x_i)|f(y_i)) = \infty$. This implies $\lim_{i\to\infty}(x_i|y_i) = \infty$ by 5.5.(2). Thus $\{x_i\}$ and $\{y_i\}$ are equivalent.

If f is a rough quasi-isometry, then it has a rough inverse $h : Y \to X$ which is a rough quasi-isometry. Since $\partial(\mathrm{id}_X) = \mathrm{id}_{\partial X}$, the statements (2), (3) and the relations $h \circ f \approx \mathrm{id}_X$, $f \circ h \approx \mathrm{id}_Y$ imply $\partial h \circ \partial f = \mathrm{id}_{\partial X}$ and $\partial f \circ \partial h = \mathrm{id}_{\partial Y}$. From this it follows that ∂f is bijective. □

COROLLARY 6.4. *The inequalities in Proposition 5.5 are also valid if $x, y, z \in \partial X$, if x', y', z' are the images of these points under ∂f, and if we assume in part (2) that x, y, z are distinct points.*

Proof. This follows from the definition of ∂f, Proposition 5.5, and (3.4). Note that for the proof of the second part we need inequality 5.5.(2) under the slightly stronger assumption $(x|z)_w - (x|y)_w \geqslant -C\delta$, $x, y, z, w \in X$ (cf. the remark following Prop. 5.5). □

Theorem 6.5. *Suppose X and Y are Gromov hyperbolic almost geodesic metric spaces, and let $f : X \to Y$ be a map.*

(1) *If f is a rough similarity, then ∂f is a snowflake map.*

(2) *If f is a rough quasi-isometry, then ∂f is a power quasisymmetry.*

Proof. Let x' denote $\partial f(x)$ for $x \in \partial X$. Note that ∂f is bijective by Proposition 6.3 (4).

The properties of ∂f in question do not depend on which metrics $d_{\partial X} \in \mathcal{G}(X)$ and $d_{\partial Y} \in \mathcal{G}(Y)$ in the canonical gauges on ∂X and ∂Y we choose. We may assume that the image w' of the basepoint w of X is the basepoint in Y. Then there exists constant $\epsilon, \epsilon' > 0$ such that for all $x, y \in \partial X$

$$d_{\partial X}(x, y) \asymp e^{-\epsilon(x|y)} \quad \text{and} \quad d_{\partial Y}(x', y') \asymp e^{-\epsilon'(x'|y')}\,. \tag{6.3}$$

Here $\asymp$ means equality up to a multiplicative constant independent of x and y.

From this we see that statement (1) is equivalent to the existence of constants $\lambda > 0$ and $K \geqslant 0$ such that

$$\lambda(x|y) - K \leqslant (x'|y') \leqslant \lambda(x|y) + K \tag{6.4}$$

for all $x, y \in \partial X$. If f is a rough similarity, an inequality like this holds for $x, y \in X$ (the prime indicating the image under f in this case). The definition of ∂f, and (3.4) imply that (6.4) also holds for $x, y \in \partial X$.

Statement (2) is a straightforward consequence of (6.3), 5.5.(2), and Corollary 6.4. □

We can summarize Proposition 6.2, Proposition 6.3, and Theorem 6.5 as follows. Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be categories, where the objects are Gromov hyperbolic almost geodesic metric spaces. The morphisms of $\mathcal{C}_1$ are rough similarities, while the morphisms of $\mathcal{C}_2$ are rough quasi-isometries. Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be categories, where the objects are bounded and complete B-structures. The morphisms of $\mathcal{D}_1$ are snowflake maps, and the morphisms of $\mathcal{D}_2$ are power quasisymmetries. Then $X \mapsto \partial X$ is a functor from $\mathcal{C}_i$ to $\mathcal{D}_i$ for $i \in \{1, 2\}$. Actually, as the results in section 7 will indicate, it is more appropriate to consider rough mapping classes of rough similarities and rough quasi-isometries as the morphisms of $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively.

7 The Metric Space Con(Z)

Given a bounded metric space (Z, d), we now construct a Gromov hyperbolic space $\mathrm{Con}(Z)$. The space $\mathrm{Con}(Z)$ has properties analogous to the hyperbolic convex hull of a set in the boundary of a real hyperbolic space. We refer to section 10 for some further discussion. Our construction is similar to one given by Gromov [Gr, 1.8.A.(b)] and to a construction of Trotsenko and Väisälä [TV].

Set

$$\mathrm{Con}(Z) = Z \times \big(0, D(Z)\big] .$$

Here and in the following we let $D(Z) = \mathrm{diam}(Z)$ if $\mathrm{diam}(Z) > 0$, and $D(Z) = 1$ if $\mathrm{diam}(Z) = 0$, i.e., if Z consists of a single point. It is convenient to include this trivial case in the definition, for otherwise we would have to exclude it explicitly in some of the following statements.

Define $\rho : \mathrm{Con}(Z) \times \mathrm{Con}(Z) \to [0, \infty)$ by

$$\rho\big((z, h), (z', h')\big) = 2 \log \left(\frac{d(z, z') + h \vee h'}{\sqrt{hh'}} \right) . \tag{7.1}$$

The motivation of the factor 2 comes from the fact that it corresponds to curvature -1 for real hyperbolic spaces.

LEMMA 7.1. *ρ is a metric on* $\mathrm{Con}(Z)$.

Proof. The triangle inequality,

$$\rho\big((z,h),(z'',h'')\big) \leqslant \rho\big((z,h),(z',h')\big) + \rho\big((z',h'),(z'',h'')\big)\,, \quad (7.2)$$

reduces to

$$\frac{d(z,z'')+h\vee h''}{\sqrt{hh''}} \leqslant \frac{d(z,z')+h\vee h'}{\sqrt{hh'}} \cdot \frac{d(z',z'')+h'\vee h''}{\sqrt{h'h''}}\,,$$

which is equivalent to

$$h'\big(d(z,z'')+h\vee h''\big) \leqslant \big(d(z,z')+h\vee h'\big)\big(d(z',z'')+h'\vee h''\big)\,. \quad (7.3)$$

Since $d(z,z'') \leqslant d(z,z')+d(z',z'')$, inequality (7.3) holds, which verifies the triangle inequality. The other properties of a metric are immediate. □

Henceforth, $\mathrm{Con}(Z)$ will always be equipped with this metric ρ. We will use the same letter ρ for metrics on different spaces $\mathrm{Con}(Z)$.

Theorem 7.2. *There are constants $\delta \geqslant 0$, $k \geqslant 0$ with the following property. If (Z,d) is a bounded metric space, then $\mathrm{Con}(Z)$ is δ-hyperbolic, k-visual, and k-roughly geodesic.*

Proof. First we prove the Gromov hyperbolicity of $\mathrm{Con}(Z)$.

Suppose we are given numbers $r_{ij} \geqslant 0$ such that $r_{ij} = r_{ji}$ and $r_{ij} \leqslant r_{ik}+r_{kj}$ for $i,j,k \in \{1,2,3,4\}$. Then $r_{12}r_{34} \leqslant 4((r_{13}r_{24})\vee(r_{14}r_{23}))$. To see this, we may assume that r_{13} is the smallest of the quantities r_{ij} appearing on the right hand side of this inequality. Then $r_{12} \leqslant r_{13}+r_{32} \leqslant 2r_{23}$ and $r_{34} \leqslant r_{31}+r_{14} \leqslant 2r_{14}$. The inequality follows.

Now let $x_i = (z_i,h_i)$, $i \in \{1,2,3,4\}$, be four arbitrary points in $\mathrm{Con}(Z)$. Set $d_{ij} = d(z_i,z_j)$ and $r_{ij} = d_{ij}+h_i\vee h_j$. The numbers r_{ij} satisfy the above requirements. Hence

$$\begin{aligned}&(d_{1,2}+h_1\vee h_2)(d_{3,4}+h_3\vee h_4)\\ &\leqslant 4\big((d_{1,3}+h_1\vee h_3)(d_{2,4}+h_2\vee h_4)\big)\vee\big((d_{1,4}+h_1\vee h_4)(d_{2,3}+h_2\vee h_3)\big)\,.\end{aligned}$$

This translates to

$$\rho(x_1,x_2)+\rho(x_3,x_4) \leqslant \big(\rho(x_1,x_3)+\rho(x_2,x_4)\big)\vee\big(\rho(x_1,x_4)+\rho(x_2,x_3)\big)+C\,,$$

which shows that $\mathrm{Con}(Z)$ is C-hyperbolic (cf. (3.3)).

Choose a point $z_0 \in Z$, and let $o = (z_0, D(Z)) \in \mathrm{Con}(Z)$. If $p = (z,h) \in \mathrm{Con}(Z)$ is an arbitrary point, a C-roughly geodesic ray emanating from o and passing through p can be obtained from an appropriate parameterization of the set $\{o\}\cup\{z\}\times(0,D(Z)]$. More precisely, define the ray $\gamma : [0,\infty) \to \mathrm{Con}(Z)$ by $\gamma(0) = o$ and $\gamma(t) = (z, D(Z)e^{-t})$ for $t > 0$. It is easy to check that γ has the required properties.

Finally, $\mathrm{Con}(Z)$ is C-roughly geodesic as follows from Proposition 5.6 and the first two parts of the proof. □

We want to show that the assignment $Z \mapsto \mathrm{Con}(Z)$ is a functor of appropriately defined categories. The basic problem is to define an induced map $\widehat{f} : \mathrm{Con}(X) \to \mathrm{Con}(Y)$ for a map $f : X \to Y$, $x \mapsto x'$. This is possible if f is a power quasisymmetry. As we mentioned in the introduction, the idea is similar to the fuzzy extension of a quasiconformal map on $\mathbb{R}^n$ to upper half-space in $\mathbb{R}^{n+1}$ by Tukia and Väisälä [TuV2]. One would like to define $\widehat{f}(x,h) = (x', |x'-y'|)$, where y is a point in X with $h = |x-y|$. The fact that $\widehat{f}$ may not be well-defined is not serious problem. For if z is any other point with $h = |x-z|$, then $|x'-y'|$ and $|x'-z'|$ are comparable if f is a power quasisymmetry. In this case, the ρ-distance of $(x', |x'-y'|)$ and $(x', |x'-z'|)$ in $\mathrm{Con}(Y)$ is bounded by a constant only depending on the parameters of f. So $\widehat{f}$ is "roughly" well-defined.

More serious is the problem that there might exist no y with $h = |x-y|$, or even worse, no y such that $|x-y|$ is comparable to h. In this case, one has to "interpolate" between those heights h which are attained as distances of points.

The following lemma gives the basic step in the construction of $\widehat{f}$. For $z \in Z$, we denote by R_z the ray in $\mathrm{Con}(Z)$ that ends at $z \in Z$: $R_z = \{z\} \times (0, D(Z)] \subset \mathrm{Con}(Z)$. The maps f_x of the lemma will be used in the definition of $\widehat{f}$, and will be the restrictions of $\widehat{f}$ to the rays R_x (cf. the definition following Lemma 7.3).

LEMMA 7.3. *Suppose that $f : X \to Y$, $x \mapsto x' = f(x)$, is an (α, c)-quasisymmetry of bounded metric spaces X, Y. For all $x \in X$ we can define a map $f_x : R_x \to R_{x'}$ with the following properties:*

(1) *There exist $\lambda = \lambda(\alpha, c) \geqslant 1$, $k = k(\alpha, c) \geqslant 0$ such that all maps f_x, $x \in X$, are (λ, k)-rough quasi-isometries.*

(2) *If $x, y \in X$, $x \neq y$, $h = |x-y|$, and $f_x(x,h) = (x', h')$, then*
$$C(\alpha, c)^{-1} |x'-y'| \leqslant h' \leqslant |x'-y'| C(\alpha, c)\,.$$

(3) *Let $x \in X$, $t_1, t_2 \in (0, D(X)]$, $f_x(x, t_1) = (x', t_1')$, $f_x(x, t_2) = (x', t_2')$. Then*
$$t_1 \leqslant t_2 \Rightarrow t_1' \leqslant t_2'\,,$$
i.e., f_x is non-decreasing on R_x.

(4) *For all $x, y \in X$, $x \neq y$, and $h \in [|x-y|, D(X)]$*
$$\rho\big(f_x(x,h), f_y(y,h)\big) = O_{\alpha,c}(1)\,.$$

(5) *If Z is a metric space with finite diameter and $g : Y \to Z$ is a (β, c')-quasisymmetry, then for all $x \in X$ and $p \in R_x$*

$$\rho\big((g \circ f)_x(p), g_{x'}(f_x(p))\big) = O_{\alpha,\beta,c,c'}(1)\,.$$

In particular, $(g \circ f)_x \approx g_{f(x)} \circ f_x$.

(6) *If f is a bilipschitz or snowflake map, then the maps f_x, $x \in X$, can be defined such that they are rough isometries or rough similarities with the same parameters, respectively. The parameters only depend on the parameters of f. These maps satisfy statements similar to* (2)–(5) (*cf. proof*).

Proof. We may assume $\operatorname{diam}(X) > 0$. For $x \in X$ let $S_x \subset \mathbb{N}$ be the set of all $l \in \mathbb{N}$ for which the annulus $A_X(x,l) = \{z \in X : 2^{-l-1}\operatorname{diam}(X) < |z-x| \leqslant 2^{-l}\operatorname{diam}(X)\}$ is nonempty. The set S_x can be regarded as the "scale spectrum" of X at x. Note that $m = \min S_x \in \{0,1\}$. Similarly, let $S_{x'}$ be the scale spectrum of Y at the point x'.

We define a map $\phi_x : S_x \to S_{x'}$ by

$$\phi_x(l) = \sup\big\{l' \;\big|\; \exists y{\in}X : 2^{-l-1}\operatorname{diam}(X) < |y{-}x|\,,\; y'{\in}A_Y(x',l')\big\}\,, \quad \forall l{\in}S_x\,.$$

Obviously, ϕ_x is non-decreasing on S_x. If $l \in S_x$ there exists a point $y \in A_X(x,l)$. If we take any such point y, then for a unique $l' \in S_{x'}$ we have $y' \in A_Y(x',l')$. From the fact that f is an (α,c)-quasisymmetry it follows that $|l' - \phi_x(l)| = O_{\alpha,c}(1)$.

For $l_1, l_2 \in S_x$ choose points $y_1 \in A_X(x,l_1)$, and $y_2 \in A_X(x,l_2)$. Since f is an (α,c)-quasisymmetry,

$$C(\alpha,c)^{-1}\eta_{\alpha,c}(2^{l_1-l_2})^{-1} \leqslant \frac{|y_1' - x'|}{|y_2' - x'|} \leqslant C(\alpha,c)\eta_{\alpha,c}(2^{l_2-l_1})\,. \quad (7.4)$$

This and the remark following the definition of ϕ_x show

$$\lambda^{-1}|l_2 - l_1| - C(\alpha,c) \leqslant |\phi_x(l_2) - \phi_x(l_1)| \leqslant \lambda|l_2 - l_1| + C(\alpha,c)\,, \quad (7.5)$$

where $\lambda = \lambda(\alpha) \geqslant 1$.

The set $\phi_x(S_x)$ is $C(\alpha,c)$-cobounded in $S_{x'}$. For if $l' \in S_{x'}$ is arbitrary, then, since f is bijective, there exists a point $y \in X$, $y \neq x$, such that $y' \in A_Y(x',l')$. For some $l \in S_x$ we have $y \in A_X(x,l)$. By the above remark $|\phi_x(l) - l'| = O_{\alpha,c}(1)$.

Since $\{0,1\} \cap S_{x'} \neq 0$ and ϕ_x is non-decreasing, this coboundedness implies

$$\phi_x(m) \leqslant C(\alpha,c)\,. \quad (7.6)$$

We now extend the map $\phi_x : S_x \to S_{x'}$ to a map $\phi_x : [0,\infty) \to [0,\infty)$. We use the same notation for the new and the old map. The extension is essentially obtained by linear interpolation.

More precisely, let $M = \sup S_x \in \mathbb{N} \cup \{\infty\}$. Define $\phi_x(t) = \phi_x(m)$ for $t \in [0, m)$. If $t \in (m, M) \backslash S_x$, there a smallest interval (l_1, l_2) with endpoints $l_1, l_2 \in S_x$ containing t. Then $t = \mu l_1 + (1-\mu) l_2$ for a unique $\mu \in (0, 1)$. Define $\phi_x(t) = \mu \phi_x(l_1) + (1-\mu)\phi_x(l_2)$. Finally, if $M < \infty$ and $t > M$ put $\phi_x(t) = \phi_x(M) + (t - M)$.

Obviously, ϕ_x is continuous and non-decreasing. Furthermore, $\lim_{t\to\infty} \phi_x(t) = \infty$. This follows from the definition of ϕ_x if $M < \infty$ and from (7.5) if $M = \infty$. Now (7.6) shows that $\phi_x([0, \infty))$ is $C(\alpha, c)$-cobounded in $[0, \infty)$.

Since $\phi_x : S_x \to S_{x'}$ is non-decreasing and satisfies (7.5), it follows that $\phi_x : [0, \infty) \to [0, \infty)$ satisfies (7.5) for all $l_1, l_2 \in [0, \infty)$. This statement is straightforward to prove. We leave the details to the reader. (Note that (7.5) is true, when l_1, l_2 are in some minimal non-degenerate interval with endpoints in $S_x \cup \{0, \infty\}$. Introducing appropriately chosen points in $S_x \cap [l_1, l_2]$, the general case of (7.5) can be reduced to this special case.)

These considerations show that $\phi_x : [0, \infty) \to [0, \infty)$ is a non-decreasing rough quasi-isometry with parameters only depending on α and c.

Now let $D = \operatorname{diam}(X)$, $D' = \operatorname{diam}(Y)$, and define for $(x, h) \in \operatorname{Con}(X)$

$$f_x(x, h) = \left(x', D' 2^{-\phi_x(\log_2(D/h))}\right). \tag{7.7}$$

The statements about the maps f_x follow directly from the properties of the maps ϕ_x. This is clear for (1) and (3). For (2) note that $\phi_x(\log_2(D/|x-y|)) = \log_2(D'/|x'-y'|) + O_{\alpha,c}(1)$ for $x, y \in X$, $x \neq y$. Statement (4) follows from (2) and (3).

To prove (5), define $\psi_y : S_y \to S_{g(y)}$, $y \in Y$, and $\theta_x : S_x \to S_{g(f(x))}$, $x \in X$, in the same way for g and $g \circ f$, respectively, as the maps ϕ_x, $x \in X$, were defined for f. Let $x \in X$. Making the parameters of the maps f and g larger if necessary, we may assume that both are power quasisymmetries with the same parameters, α and c, say. This implies that for $t \in S_x$

$$\theta_x(t) = \psi_{x'}(\phi_x(t)) + O_{\alpha,c}(1). \tag{7.8}$$

To establish this equation for general $t \in [0, \infty)$ assume $t \in (m, M) \backslash S_x$. The cases $t \in [0, m]$ or $t \in [M, \infty)$ are easier and can be treated in a similar way as below.

Let (l_1, l_2) be the smallest interval with endpoints in S_x containing x and write $t = \mu l_1 + (1-\mu) l_2$ with $\mu \in (0, 1)$. Then $\phi_x(t) = \mu l_1' + (1-\mu) l_2'$, where $l_1' = \phi_x(l_1)$ and $l_2' = \phi_x(l_2)$. In particular, $\phi_x(t) \in [\min S_{x'}, \sup S_{x'}]$, and so there exists some minimal interval $[\tilde{l}_1, \tilde{l}_2]$ with endpoints in $S_{x'}$ containing $\phi_x(t)$. The interval $[\tilde{l}_1, \tilde{l}_2]$ is degenerate if $\phi_x(t) \in S_{x'}$.

Since ϕ_x is non-decreasing and $\phi(S_x)$ is $C(\alpha, c)$-cobounded in $S_{x'}$, we have $\tilde{l}_1 = l'_1 + O_{\alpha,c}(1)$ and $\tilde{l}_2 = l'_2 + O_{\alpha,c}(1)$. From the definition of θ_x and ϕ_x, equation (7.8) for $t \in S_x$, and the fact that ψ_x is a rough quasi-isometry with parameters only depending on α and c, we obtain

$$\begin{aligned}\theta_x(t) &= \mu\psi_{x'}(l'_1) + (1-\mu)\psi_{x'}(l'_2) + O_{\alpha,c}(1)\\ &= \mu\psi_{x'}(\tilde{l}_1) + (1-\mu)\psi_{x'}(\tilde{l}_2) + O_{\alpha,c}(1)\\ &= \psi_{x'}(\phi_x(t)) + O_{\alpha,c}(1)\,.\end{aligned}$$

This shows that (7.8) holds for all $t \in [0, \infty)$. The maps $(g \circ f)_x$, $x \in X$, and g_y, $y \in Y$, are defined similarly as in (7.7) using the maps θ_x and ψ_y, respectively. Statement (5) then follows from (7.8).

To see that (6) is true, define $f_x(x, h) = (x', (D'/D)h)$ for $(x, h) \in \mathrm{Con}(X)$ if f is a bilipschitz map. Statements (1)–(5) are immediate in this case with all constants only depending on the bilipschitz constant of f (and the bilipschitz constant of g in (5)).

Suppose that f is an (α, c)-snowflake map. If $S(X) := \bigcup_{x\in X} S_x$ is an infinite set, then α is uniquely determined. In this case let $\alpha(f) = \alpha$. If $S(X)$ is a finite set, let $\alpha(f) = 1$.

Note that the property of a space X that $S(X)$ is finite or infinite is invariant under snowflake maps. This implies that if $g : Y \to Z$ is a snowflake map, then $\alpha(g \circ f) = \alpha(g)\alpha(f)$.

The set $S(X)$ is finite, if and only if $d(X) := \inf_{x,y\in X, x\neq y} |x - y| > 0$. In this case, f is a λ-bilipschitz map with $\lambda = \lambda(\alpha, c, \mathrm{diam}(X), d(X))$.

Now define $f_x(x, h) = (x', D'(h/D)^{\alpha(f)})$.

The stated properties of the maps f_x are immediate to check. The constants in (1), (2), (4) will only depend on α and c. If $S(X)$ is finite, the constant in (2) will also depend on $d(X)$ and $\mathrm{diam}(X)$. The snowflake version of statement (5) is true with constant 0. □

If X and Y are bounded metric spaces, and $f : X \to Y$ is a power quasisymmetry (which includes snowflake or bilipschitz maps), we define a map $\widehat{f} : \mathrm{Con}(X) \to \mathrm{Con}(Y)$ by

$$\widehat{f}(x, h) = f_x(x, h)\,, \quad \forall (x, h) \in \mathrm{Con}(X)\,.$$

Here f_x, $x \in X$, are the maps defined in the last lemma.

Theorem 7.4. *Suppose that $f : X \to Y$ is a power quasisymmetry of bounded metric spaces X, Y. Then $\widehat{f} : \mathrm{Con}(X) \to \mathrm{Con}(Y)$ is a rough quasi-isometry. If f is a snowflake or bilipschitz map, then $\widehat{f}$ is a rough similarity or a rough isometry, respectively.*

If Z is a bounded metric space, and $g : Y \to Z$ is a map of the same type as f, i.e., a power quasisymmetry, a snowflake map, or a bilipschitz map, then $\widehat{g \circ f} \approx \widehat{g} \circ \widehat{f}$.

Proof. Assume $f : X \to Y$, $x \mapsto x' = f(x)$ is an (α, c)-quasisymmetry.

We have $\widehat{f}(\mathrm{Con}(X)) = \bigcup_{x \in X} f_x(R_x)$, and $\bigcup_{x \in X} R_{x'} = \mathrm{Con}(Y)$. Since $f_x(R_x)$ is $C(\alpha, c)$-cobounded in $R_{x'}$ for all $x \in X$ by Lemma 7.3.(1), the set $\widehat{f}(\mathrm{Con}(X))$ is $C(\alpha, c)$-cobounded in $\mathrm{Con}(Y)$.

To show that $\widehat{f} : \mathrm{Con}(X) \to \mathrm{Con}(Y)$ is a rough quasi-isometry, we have to establish inequality (2.1) for $\widehat{f}$. For two arbitrary points $q_1 = (x_1, h_1), q_2 = (x_2, h_2) \in \mathrm{Con}(X)$ we express the distance of q_1 and q_2 and the distance of their image points $q_1' = (x_1', h_1')$ and $q_2' = (x_2', h_2')$ under $\widehat{f}$ by distances only involving pairs of points lying on the same ray R_z. Then (2.1) will follow from the definition of $\widehat{f}$ and the fact that the maps f_x, $x \in X$, are rough quasi-isometries with the same parameters.

To that purpose define $h = |x_1 - x_2| \vee h_1 \vee h_2$, $p_1 = (x_1, h)$, $p_2 = (x_2, h)$, $p_1' = \widehat{f}(p_1) = (x_1', h')$, $p_2' = \widehat{f}(p_2) = (x_2', \tilde{h}')$. The definition of ρ shows

$$\rho(q_1, q_2) = \rho(q_1, p_1) + \rho(p_2, q_2) + O(1)\,. \tag{7.9}$$

If $h = |x_1 - x_2|$, then from Lemma 7.3.(2) and (3) it follows that

$$C(\alpha, c)^{-1}|x_1' - x_2'| \leqslant h', \tilde{h}' \leqslant C(\alpha, c)|x_1' - x_2'|\,, \quad \text{and} \quad h_1' \vee h_2' \leqslant h', \tilde{h}'\,.$$

This implies

$$\rho(q_1', q_2') = \rho(q_1', p_1') + \rho(p_2', q_2') + O_{\alpha,c}(1)\,. \tag{7.10}$$

If $h \neq |x_1 - x_2|$, i.e., if $h = h_1 \vee h_2$, we may without loss of generality assume that $h = h_1$. Then $q_1' = p_1'$ and Lemma 7.3.(4) shows that $\rho(p_1', p_2') = O_{\alpha,c}(1)$. Thus, (7.10) also holds in this case.

Consequently, Lemma 7.3.(1) now gives

$$\begin{aligned} \rho(q_1', q_2') &= \rho(q_1', p_1') + \rho(p_2', q_2') + O_{\alpha,c}(1) \\ &\leqslant C(\alpha, c)\big(\rho(q_1, p_1) + \rho(p_2, q_2)\big) + O_{\alpha,c}(1) \\ &= C(\alpha, c)\rho(q_1, q_2) + O_{\alpha,c}(1)\,. \end{aligned}$$

An inequality in the opposite direction can be obtained in the same way. This shows that $\widehat{f}$ is a rough quasi-isometry.

If f is a snowflake or a bilipschitz map, the same argument based on (7.9), (7.10) and Lemma 7.3.(6) shows that $\widehat{f}$ is a rough similarity or a rough isometry, respectively. Finally, the last statement follows from Lemma 7.3.(5). □

Again we may summarize the results of this section in the language of categories. Let $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$ be categories, where the objects are bounded

metric spaces, and the morphisms are power quasisymmetries, snowflake maps, and bilipschitz maps, respectively.

Let $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ be categories, where the objects are visual Gromov hyperbolic spaces, and the morphisms are rough mapping classes of rough quasi-isometries, rough similarities, and rough isometries, respectively. Then $Z \mapsto \mathrm{Con}(Z)$ is a functor from $\mathcal{C}_i$ to $\mathcal{D}_i$ for $i \in \{1, 2, 3\}$.

In particular, we get the following correspondence for the types of maps

$$\begin{aligned} \text{power quasisymmetry} &\;\widehat{=}\; \text{rough quasi-isometry}, \\ \text{snowflake map} &\;\widehat{=}\; \text{rough similarity}, \\ \text{bilipschitz map} &\;\widehat{=}\; \text{rough isometry}. \end{aligned}$$

Note that the first two correspondences (going from right to left) are exactly what we found in the last section for the functor $X \mapsto \partial X$.

8 The Relation of ∂X and $\mathrm{Con}(Z)$

In this section we will investigate the relation of the functors $X \mapsto \partial X$ and $Z \mapsto \mathrm{Con}(Z)$. We content ourselves with proving two statements about the objects of the categories involved. Similar questions can also be studied for the morphisms.

Theorem 8.1. *Suppose (Z, d) is a complete bounded metric space, and let $X = \mathrm{Con}(Z)$. Then the spaces ∂X and Z can be identified as sets, and d is bilipschitz to a metric in $\mathcal{G}(X)$, the canonical gauge on ∂X.*

Proof. Take as the basepoint $o \in X$ some point of the form (z_0, D), where $D = D(Z)$. Let $x = (z, h)$, $x' = (z', h') \in X$. Then

$$\begin{aligned} (x|x') &= \log\left(\frac{d(z_0, z)+D}{\sqrt{Dh}}\right) + \log\left(\frac{d(z_0, z')+D}{\sqrt{Dh'}}\right) - \log\left(\frac{d(z, z')+h \vee h'}{\sqrt{hh'}}\right) \\ &= -\log\left(d(z, z') + h \vee h'\right) + O_D(1)\,. \end{aligned}$$

From this follows that a sequence $\{(z_i, h_i)\}$ in $\mathrm{Con}(Z)$ converges at infinity if and only if $\{z_i\}$ is a Cauchy sequence in Z and $\lim_{i\to\infty} h_i = 0$. Since Z is complete, $\{z_i\}$ has a limit y in Z. It is immediate to check that if a sequence $\{(z_i', h_i')\}$ is equivalent to $\{(z_i, h_i)\}$, then $\lim_{i\to\infty} z_i = \lim_{i\to\infty} z_i' = y$. So by assigning y to the equivalence class of $\{(z_i, h_i)\}$ we get a well-defined map from ∂X to Z. It is straightforward to verify that this map is bijective. Thus, ∂X and Z can be identified as sets.

Moreover, from the above expression and (3.4) we conclude

$$(z|z') = -\log(d(z, z')) + O_D(1)\,, \qquad \forall z, z' \in \partial X = Z\,.$$

This shows that d is bilipschitz to a metric in the canonical gauge on ∂X. □

Theorem 8.2. *Suppose X is a visual Gromov hyperbolic metric space. Then X and $\mathrm{Con}(\partial X)$ are roughly similar.*

The proof will actually show more. If d is a metric in the canonical gauge on ∂X for which there exists a point $w \in X$, and a constant $a \geqslant 1$ such that

$$a^{-1} e^{-(z|z')_w} \leqslant d(z,z') \leqslant a e^{-(z|z')_w}\,, \quad z, z' \in \partial X\,,$$

then X and $\mathrm{Con}((\partial X, d))$ are roughly isometric.

Proof. Assume X is δ-hyperbolic and k-visual with respect to $o \in X$. If we take different metrics d_1, d_2 in the canonical gauge on ∂X, then the spaces $(\partial X, d_1)$ and $(\partial X, d_2)$ are snowflake equivalent. Theorem 7.4 implies that the spaces $\mathrm{Con}((\partial X, d_1))$ and $\mathrm{Con}((\partial X, d_2))$ are roughly similar. Therefore, the assertion does not depend on which metric in the canonical gauge on ∂X we choose. Fix such a metric d on ∂X. Then we may assume that

$$a^{-1} \exp\big(-\epsilon(z|z')\big) \leqslant d(z,z') \leqslant a \exp\big(-\epsilon(z|z')\big)\,, \quad \forall z, z' \in \partial X\,. \tag{8.1}$$

Making k larger if necessary, by Proposition 5.6 and Proposition 5.2.(2) for each $z \in \partial Z$ we can choose a k-roughly geodesic ray $\gamma_z : [0,\infty) \to X$ with $\gamma(o) = 0$ and $\lim \gamma_z = z$.

Let $D = D((\partial X, d))$. For $(z,h) \in \mathrm{Con}(\partial X)$, we define $f(z,h) = \gamma_z(\epsilon^{-1} \log(D/h))$. We show that $f : \mathrm{Con}(\partial X) \to X$ is a rough similarity.

First, note that the image of f is cobounded in X. For each $x \in X$ lies on some k-roughly geodesic ray emanating from o. If $z = \lim \gamma$, then from Lemma 5.1 it follows that $|\gamma(t) - \gamma_z(t)| = O_{\delta,k}(1)$ for $t \geqslant 0$. So x has distance at most $O_{\delta,k}(1)$ to γ_z which is contained in the image of f.

Let (z,h) and (z',h') be two arbitrary points in $\mathrm{Con}(\partial X)$. Observe that $(z|f(z,h)) = \epsilon^{-1} \log(D/h) + O_{\delta,k}(1)$, and $(z'|f(z',h')) = \epsilon^{-1} \log(D/h') + O_{\delta,k}(1)$. Therefore, from Lemma 5.1 we get

$$\begin{aligned}
&\epsilon\big|f(z,h) - f(z',h')\big| \\
&\quad = \log(D/h) + \log(D/h') - 2\big(\epsilon(z|z') \wedge \log(D/h) \wedge \log(D/h')\big) + O_{\delta,k,\epsilon}(1) \\
&\quad = -\log h - \log h' + 2\big(-\epsilon(z|z') \vee \log h \vee \log h'\big) + O_{\delta,k,\epsilon,D}(1) \\
&\quad = 2 \log \left(\frac{e^{-\epsilon(z|z')} \vee h \vee h'}{\sqrt{hh'}} \right) + O_{\delta,k,\epsilon,D}(1)\,.
\end{aligned}$$

Now apply (8.1) and the inequalities $(u+v)/2 \leqslant u \vee v \leqslant u+v$ to obtain

$$\begin{aligned}\epsilon\big|f(z,h)-f(z',h')\big| &= 2\log\left(\frac{d(z,z')+h\vee h'}{\sqrt{hh'}}\right)+O_{\delta,k,\epsilon,D,a}(1)\\ &= \rho\big((z,h),(z',h')\big)+O_{\delta,k,\epsilon,D,a}(1). \qquad (8.2)\end{aligned}$$

Since the image of f is cobounded in X, it follows from (8.2) that f is a rough quasi-similarity. Note that if $\epsilon = 1$, then f is a rough isometry. □

The last two statements may be summarized as follows. If Z is a complete bounded metric space, then

$$Z \overset{\text{sf}}{=} \partial(\operatorname{Con} Z)\,,$$

where $\overset{\text{sf}}{=}$ means snowflake equivalence of the spaces. Note that for a space to be snowflake equivalent to the boundary of a Gromov hyperbolic space it is necessary that the space is complete and bounded by Proposition 6.2.

On the other hand, if X is a visual Gromov hyperbolic space, then

$$X \simeq \operatorname{Con}(\partial X)\,,$$

where $\simeq$ means rough similarity of the spaces. It is not hard to see that the property of being visual and Gromov hyperbolicity are preserved under rough similarities. So the assumptions on X are necessary by Theorem 7.2.

These two statements show that in some sense $X \mapsto \partial X$ and $Z \mapsto \operatorname{Con}(Z)$ are inverse to each other.

9 Growth and Assouad Dimension

In this section we prove Theorem 9.2 which will be used in sections 10 and 11.

Definition (cf. [A]). Let X be a metric space. For $\alpha, \beta > 0$, let $S(\alpha,\beta)$ be the maximal cardinality of a set $V \subset X$ such that $\alpha \leqslant |x-y| \leqslant \beta$ for all $x, y \in V$, $x \neq y$. Define t to be the infimum of all numbers $s \geqslant 0$ such that for some constant $K \geqslant 0$ the inequality

$$S(\alpha,\beta) \leqslant K(\beta/\alpha)^s$$

holds for all $0 < \alpha \leqslant \beta$. (It is understood that $t = \infty$, if no such s exists.) Then $\dim_A(X) = t$ is the *Assouad dimension* of X.

See [L] for a further discussion of the Assouad dimension. We will need the following theorem (cf. [A]).

Theorem 9.1 (Assouad's Theorem). *Let (Z,d) be a metric space with finite Assouad dimension, and let $p \in (0,1)$. Then there is some integer n such that the metric space (Z,d^p) admits a bilipschitz embedding into $\mathbb{R}^n$.*

Here, d^p is the metric $d^p(x,y) = d(x,y)^p$, $x, y \in Z$.

A metric space X is called *doubling*, if for all r, R with $R > r > 0$ there exists $N \in \mathbb{N}$ only depending on R/r such that every open ball of radius R in X can be covered by N open balls of radius r. It can be shown that a metric space has finite Assouad dimension if and only if it is doubling.

The next theorem gives a sufficient condition for a Gromov hyperbolic geodesic metric space X to have a boundary with finite Assouad dimension. To state the theorem we make the following definition.

DEFINITION. A metric space X has *bounded growth at some scale*, if there are constants r, R with $R > r > 0$, and $N \in \mathbb{N}$ such that every open ball of radius R in X can be covered by N open balls of radius r.

Theorem 9.2. *Let X be a Gromov hyperbolic geodesic metric space with bounded growth at some scale. Then the Assouad dimension of ∂X is finite.*

As we remarked above, the assertion is equivalent to ∂X being doubling. So a weak doubling condition on X implies that ∂X is doubling. Note that the properties of a metric space to have finite Assouad dimension and to be doubling are preserved by quasisymmetric maps. In particular, the statement about ∂X in the theorem is independent of which metric in the canonical gauge we choose.

Examples for Gromov hyperbolic geodesic metric spaces having bounded growth at some scale are all complete simply-connected Riemannian n-manifolds X with sectional curvature κ satisfying $-b^2 \leqslant \kappa \leqslant -a^2 < 0$. In this case open balls of radius 1 in X are bilipschitzly equivalent to the unit ball in $\mathbb{R}^n$ with bilipschitz constant only depending on a and b. This statement follows from Topogonov's comparison theorem, and implies that X has bounded growth at some scale. In particular, complex hyperbolic spaces are Gromov hyperbolic geodesic metric spaces with bounded growth at some scale.

Note that any bounded degree graph has bounded growth at some scale. Hence, every finitely generated hyperbolic group has bounded growth at some scale.

The proof of 9.2 is similar to the proof of Proposition 11 from [GH, Ch. 7].

Proof. Assume X is δ-hyperbolic, and fix some metric $d_{\partial X} = d_{o,\epsilon}$ in the canonical gauge on ∂X. Rescaling the metric on X by the factor ϵ if necessary, we may assume $\epsilon = 1$. Then

$$(z|z') = -\log\big(d_{\partial X}(z,z')\big) + O_\delta(1), \qquad \forall z, z' \in \partial X\,. \tag{9.1}$$

Denote by $B(q,s)$ the open ball in X with center $q \in X$ and radius $s > 0$. By assumption, there exist constants $R > r > 0$, $N \in \mathbb{N}$ such that every open ball of radius R in X can be covered by N open balls of radius r. Suppose that $B(p, 2R-r)$ is some ball of radius $2R-r$ and center $p \in X$. Then $B(p,R)$ can be covered by N open balls of radius r with centers $p_1, \ldots, p_N \in X$, say. Since X is a geodesic metric space, it follows that $B(p,2R-r)$ is covered by the balls $B(p_1,R), \ldots, B(p_N,R)$. Therefore, $B(p,2R-r)$ can be covered by N^2 open balls of radius r. By induction it follows that any open ball of radius $mR-(m-1)r$, $m \in \mathbb{N}$, can be covered by N^m open balls of radius r.

Given any $z \in \partial X$, for every $t > 0$ there is an $x \in X$ such that $|x| = t$ and $(x|z) - t = O_\delta(1)$. To verify this, take $w \in X$ with $(w|z) \geqslant t$ and let x be the point satisfying $|x| = t$ on the geodesic segment $[o,w]$.

Fix some α, β with $0 < \alpha \leqslant \beta \leqslant 1$. Suppose $z_1, z_2, \ldots, z_n$ are points in ∂X such that

$$\alpha \leqslant d_{\partial X}(z_i, z_j) \leqslant \beta\,, \quad i \neq j\,. \tag{9.2}$$

Then (9.1) implies

$$-\log\beta - C(\delta) \leqslant (z_i|z_j) \leqslant -\log\alpha + C(\delta)\,, \qquad i \neq j\,. \tag{9.3}$$

Let $x_1, \ldots, x_n$ be points satisfying $|x_j| = -\log\beta$, $\big|-\log\beta - (x_j|z_j)\big| \leqslant C(\delta)$, and let $y_1, \ldots, y_n$ be points satisfying $|y_j| = -\log\alpha$, $\big|-\log\beta - (y_j|z_j)\big| \leqslant C(\delta)$. Then it is easy to see from Lemma 5.1 and (9.3) that

$$\begin{aligned} |x_i - x_j| &= O_\delta(1)\,, \\ |y_i - y_j| &= 2(-\log\alpha + \log\beta) + O_\delta(1)\,, \\ |x_i - y_i| &= \log\beta - \log\alpha + O_\delta(1)\,, \end{aligned} \tag{9.4}$$

for all $i \neq j$. It follows that the open ball of radius

$$R^* = \log(\beta/\alpha) + C(\delta)$$

centered at x_1 contains all the points y_i.

Let m be the least integer such that $m(R-r) \geqslant R^*$. We have seen that an open ball of radius $m(R-r)$ can be covered by N^m balls of radius r. So the set $\{y_1, \ldots, y_n\}$ can be covered by N^m balls of radius r. We now assume that $\log(\beta/\alpha)$ is much larger than r. It then follows from the middle equation in (9.4) that $|y_i - y_j| > 2r$ for $i \neq j$. Consequently, any ball of radius r can contain at most one of the points y_j. This gives

$$n \leqslant N^m \leqslant N^{1+R^*/(R-r)} \leqslant C(R,r,N,\delta)(\beta/\alpha)^{C(R,r,N,\delta)}\,.$$

By changing the constants if necessary, the same is true even without the assumption that $\log(\beta/\alpha)$ is much larger than r. Since the diameter of X

is finite, the assumption $\beta \leqslant 1$ is also inconsequential. We conclude that the Assouad dimension of ∂X is finite. □

10 Embeddings into Real Hyperbolic Spaces

The definition of the metric ρ on $\mathrm{Con}(X)$ was motivated by the upper half-space model of real hyperbolic n-space $\mathbb{H}^n$. In this model

$$\mathbb{H}^n = \mathbb{R}^n_+ = \big\{(x_1, \dots, x_n) \in \mathbb{R}^n : x_n > 0\big\}$$

is equipped with the Riemannian metric given by the length element

$$ds^2 = \tfrac{1}{x_n^2}(dx_1^2 + \cdots + dx_n^2)\,.$$

The space $\mathbb{H}^n$ is a Riemannian n-manifold with constant sectional curvature -1.

We identify $\mathbb{R}^{n-1}$ with the hyperplane in $\mathbb{R}^n$ given by the equation $x_n = 0$. For $x \in \mathbb{R}^n$ we write $x = (z, h)$, where $z \in \mathbb{R}^{n-1}$, $h \in \mathbb{R}$. Then $\mathbb{H}^n = \{(z, h) : z \in \mathbb{R}^{n-1}, h > 0\}$.

It is not hard to find explicit expressions for the hyperbolic distance $d_h((z, h), (z', h'))$ of two points $(z, h), (z', h') \in \mathbb{H}^n$. For our purpose it is sufficient to know that for some universal constant $C > 0$

$$\left| d_h\big((z, h), (z', h')\big) - 2 \log \left(\frac{|z - z'| + h \vee h'}{\sqrt{hh'}} \right) \right| \leqslant C\,. \quad (10.1)$$

Here $|z - z'|$ is the euclidean distance of z and z'.

The spaces $\mathbb{H}^n$, $n \in \mathbb{N}$, are geodesic, and from (10.1) and the first part of the proof of Theorem 7.2 it follows that they are C-hyperbolic.

The Gromov boundary $\partial\mathbb{H}^n$ can be identified with $\mathbb{R}^{n-1} \cup \{\infty\}$. If $M \subset \mathbb{R}^n_+$ is bounded with respect to the euclidean metric, then (M, d_h) is C-hyperbolic, and ∂M can be identified with the intersection of the euclidean closure of M and $\mathbb{R}^{n-1}$. Moreover, if $w \in \mathbb{H}^n$, then

$$(z|z')_w = -\log|z - z'| + O_{M,w}(1)\,, \quad \forall z, z' \in \partial M \subset \mathbb{R}^{n-1}\,. \quad (10.2)$$

In particular, the euclidean metric is bilipschitz to a metric in the canonical gauge of ∂M. These statements can be verified using (10.1), and considerations as in the proof of Theorem 8.1.

If $A \subset \mathbb{H}^n \cup \partial\mathbb{H}^n$, let $\mathrm{hull}(A) \subset \mathbb{H}^n$ be the intersection of all closed half-spaces $H \subset \mathbb{H}^n$ such that $A \subset H \cup \partial H$. Recall that in the upper half-space model of $\mathbb{H}^n$, half-spaces in $\mathbb{H}^n$ are determined by hyperplanes in $\mathbb{R}^n$ perpendicular to $\mathbb{R}^{n-1}$ or spheres with center in $\mathbb{R}^{n-1}$. The set $\mathrm{hull}(A)$ is the smallest set $M \subset \mathbb{H}^n$ that is hyperbolically convex and closed, and

satisfies $A \subset M \cup \partial M$. Here we have to assume that A consists of more than one point if $A \subset \partial\mathbb{H}^n$, for otherwise $\mathrm{hull}(A) = \emptyset$.

We need the following facts about $\mathrm{hull}(A)$.

PROPOSITION 10.1. (1) *Let $A \subset \mathbb{H}^n$. For each $p \in \mathrm{hull}(A)$, there exists a point $q \in \mathbb{H}^n$ that lies on a geodesic segment with endpoints in A and satisfies $d_h(p,q) \leqslant O(1)$.*

(2) *Let $A \subset \partial\mathbb{H}^n$ be a set with more than one point, and fix $o \in \mathrm{hull}(A)$. For each $p \in \mathrm{hull}(A)$, there exists a point $q \in \mathbb{H}^n$ that lies on a geodesic ray from o to some point in A and satisfies $d_h(p,q) \leqslant O(1)$.*

(3) *Let $Z \subset \mathbb{R}^{n-1} \subset \partial\mathbb{H}^n$ be a compact set containing more than one point. Then $\mathrm{hull}(Z) \cong \mathrm{Con}(Z)$.*

Remember the notation $X \sim Y$, if two metric spaces X, Y are roughly quasi-isometric, $X \simeq Y$ if they are roughly similar, and $X \cong Y$, if there are roughly isometric.

It is worthwhile to note that the constant $O(1)$ in the statement is absolute. In particular, it does not depend on the dimension n.

Proof. Statement (1) and (2) can easily be proved, for example, by using the Euclidean unit ball in $\mathbb{R}^n$ as the Klein model for $\mathbb{H}^n$ (where the hyperbolic geodesics are line segments) with p corresponding to 0. We leave the details to the reader.

(3) Since Z is compact, $\partial(\mathrm{hull}(Z)) = Z$. To see this note that $Z \subset \partial(\mathrm{hull}(Z))$. On the other hand, assume $z \in \mathbb{R}^{n-1} \backslash Z \subset \partial\mathbb{H}^n$. Since Z is compact, there exists a small closed euclidean ball B centered at z which is disjoint from Z. Then $H = B \cap \mathbb{R}^n_+$ is a closed half-space in $\mathbb{H}^n$ disjoint from $\mathrm{hull}(Z)$. Hence, $z \notin \partial(\mathrm{hull}(Z))$.

By the second part of (1), the union of the geodesic rays in $\mathrm{hull}(Z)$ emanating from some fixed basepoint $o \in \mathrm{hull}(Z)$ is cobounded. Therefore, $\mathrm{hull}(Z)$ is visual.

Since $\mathrm{hull}(Z)$ is also Gromov hyperbolic, Theorem 8.2 shows that $\mathrm{hull}(Z) \simeq \mathrm{Con}(Z)$. To get the stronger statement $\mathrm{hull}(Z) \cong \mathrm{Con}(Z)$ note that the euclidean metric on $Z = \partial(\mathrm{hull}(Z))$ is bilipschitz to a metric in the canonical gauge on $\partial(\mathrm{hull}(Z))$ by formula (10.2). The assertion now follows from the remark after Theorem 8.2. □

Theorem 10.2. *Let X be a Gromov hyperbolic geodesic metric space with bounded growth at some scale. Then there exists an integer n such that X is roughly similar to a convex subset of hyperbolic n-space $\mathbb{H}^n$.*

Note that the theorem has an easy converse. If X is geodesic and

roughly similar to a subset of $\mathbb{H}^n$, then X is Gromov hyperbolic and has bounded growth at some scale.

Let d be the metric on X. The theorem says that for some $c > 0$ the metric space (X, cd) is roughly isometric to a convex subset of $\mathbb{H}^n$. The proof shows that there is a $C = C(\delta(X)) > 0$ such that c may be chosen as any number in $(0, C)$. However, one cannot always take $c = 1$, as demonstrated by Proposition 10.3 below.

Proof. First suppose that the union of the geodesic rays starting from some basepoint $o \in X$ is cobounded in X, and that ∂X contains more than one point.

Then X is visual, and so $X \simeq \mathrm{Con}((\partial X, d))$ by Theorem 8.2. Here, d is some fixed metric in the canonical gauge on ∂X. By Theorem 7.4 applied to snowflake maps, $\mathrm{Con}((\partial X, d)) \simeq \mathrm{Con}((\partial X, d^{1/2}))$. Theorem 9.2 shows that ∂X has finite Assouad dimension. By Theorem 9.1, $(\partial X, d^{1/2})$ admits a bilipschitz embedding into $\mathbb{R}^{n-1}$ for sufficiently large n. Let Z be the bilipschitz image of $(\partial X, d^{1/2})$ in $\mathbb{R}^{n-1}$. By Theorem 7.4 for bilipschitz maps, $\mathrm{Con}((\partial X, d^{1/2})) \cong \mathrm{Con}(Z)$. Since ∂X is complete, bounded, and contains more than one point, Z also has these properties. In particular, Z is compact. Proposition 10.1 shows that $\mathrm{Con}(Z) \cong \mathrm{hull}(Z)$. Therefore, $X \simeq \mathrm{hull}(Z) \subset \mathbb{H}^n$, proving the assertion in this case.

Now drop the additional assumptions on X. Since X has bounded growth at some scale, there exist constants $R > r > 0$, $N \in \mathbb{N}$ such that every open ball of radius R in X can be covered by N open balls of radius r. Let X_0 be a maximal set of points in X with the property that the distance between any two points in X_0 is at least $5R$.

At each point $x_0 \in X$, we glue an isometric copy of the ray $[0, \infty)$ to X by identifying x_0 with 0, the initial point of the ray. The new space $\widehat{X}$ carries a unique metric which agrees with the metric on X, the metrics on the rays glued to X, and makes $\widehat{X}$ a geodesic space. If a geodesic segment connects two points lying on different rays glued to X, then this segment contains the initial points of the rays. Using this and the thin triangle definition of hyperbolicity, it is not hard to check that $\widehat{X}$ is Gromov hyperbolic.

If we assign to each point $x_0 \in X_0$ the limit point in $\partial\widehat{X}$ of the ray glued to x_0, then we get an injective embedding of X_0 into $\partial\widehat{X}$. From this we see that we may assume that $\partial\widehat{X}$ contains more than one point. For otherwise, X_0 is a one point set. Then X has bounded diameter. The theorem is true in this case, X being roughly similar to a point.

Fix some basepoint $o \in X$. For each $x_0 \in X_0$, the union of a geodesic segment $[o, x_0]$, and the ray glued to X at x_0 is a geodesic ray emanating from o. By definition of X_0 this implies that the union of the geodesic rays from o is $5R$-cobounded.

To verify that $\widehat{X}$ has bounded growth at some scale, observe that a ball of radius R in $\widehat{X}$ can intersect at most one component of $\widehat{X} - X$. Consequently, any open ball of radius R in $\widehat{X}$ can be covered by $N + \lceil R/r \rceil$ open balls of radius r, where $\lceil R/r \rceil$ denotes the least integer greater than R/r.

It follows that $\widehat{X}$ satisfies the additional assumptions of the special case treated above. So we know that $\widehat{X} \simeq \widehat{W}$ for some convex set $\widehat{W} \subset \mathbb{H}^n$, where n is sufficiently large. Let W denote the image of X under a rough similarity $f : \widehat{X} \to \widehat{W}$. Then $X \simeq W$.

We claim that $W \cong \mathrm{hull}(W)$. To see this let $w, w' \in W$. There exist points $x, x' \in X$ such that $w = f(x)$ and $w' = f(x')$. Since X is geodesic, there exists a geodesic segment $[x, x']$ joining x and x'. The image $f([x, x'])$ is a roughly quasi-isometric path joining w and w'. By geodesic stability in $\mathbb{H}^n$, $d_{\mathrm{Haus}}(f([x, x']), [w, w']) \leqslant c$, where c depends on the parameters of f only.

This shows that $W = f(X)$ is cobounded in the set consisting of the union of all geodesic segments with endpoints in W. Since the latter set is cobounded in $\mathrm{hull}(W)$ by Proposition 10.1, the set W is also cobounded in $\mathrm{hull}(W)$. Therefore, the inclusion map $i : W \to \mathrm{hull}(W)$ is a rough isometry. We conclude $X \simeq \mathrm{hull}(W) \subset \mathbb{H}^n$, completing the proof of the theorem. □

The following theorem shows that in 10.2 "roughly similar" cannot be replaced by "roughly isometric".

PROPOSITION 10.3. *Let d_h denote the standard metric on $\mathbb{H}^2$, let $c > 1$ and $n \in \mathbb{N}$. Then the metric space $X = (\mathbb{H}^2, c\, d_h)$ is not roughly isometric to a subset of $\mathbb{H}^n$.*

Proof. Recall that $\partial\mathbb{H}^n$ the $(n-1)$-dimensional unit sphere is S^{n-1}, in the following sense. For every $o \in \mathbb{H}^n$ the metric $d_{\partial\mathbb{H}^n, o, 1}$ is bilipschitz to $d_{S^{n-1}}$, the spherical metric on S^{n-1}. We identify $\partial\mathbb{H}^n$ with S^{n-1}.

Suppose that X is roughly isometric to a subset W of $\mathbb{H}^n$. Then it must be roughly isometric to $\mathrm{hull}(W)$. Fix points $o \in X$, $o' \in W$, and fix a small $\epsilon > 0$. It follows that the metric $d_{\partial\mathbb{H}^2, o, c\epsilon}$ is bilipschitz to the metric $d_{\partial\,\mathrm{hull}(W), o', \epsilon}$, which is just the restriction of $d_{\partial\mathbb{H}^n, o', \epsilon}$ to ∂W. It therefore follows that $(d_{S^1})^c$ is bilipschitz to the restriction of $d_{S^{n-1}}$ to ∂W, in the

sense that there is a homeomorphism $g : S^1 \to \partial W \subset S^{n-1}$ and a constant C such that

$$C^{-1} d_{S^{n-1}}\big(g(x), g(y)\big) \leqslant d_{S^1}(x,y)^c \leqslant C d_{S^{n-1}}\big(g(x), g(y)\big)\,,$$

for every $x, y \in S^1$. However, for every $x, y \in S^1$ and every $s > 0$ there is a sequence $x = x_0, x_1, \ldots, x_k = y$ such that $\sum_{j=1}^{k} d_{S^1}(x_{j-1}, x_j)^c < s$. It follows that $d_{S^1}(x,y) = 0$, a contradiction. □

It is interesting to ask what can remain of the conclusion of Theorem 10.2 if the assumption that X has finite growth at some scale is removed.

For any metric space Z let

$$\mathrm{Con}_h(Z) = Z \times (0, \infty)\,,$$

and define a metric d_h on $\mathrm{Con}_h(Z)$ as follows. Given two points $p = (z,h), p' = (z',h') \in \mathrm{Con}_h(Z)$, let $d_h(p,p')$ be the distance between the two points in the hyperbolic plane $\mathbb{H}^2$ whose coordinates in the upper half-plane model are $(0,h)$ and $(|z-z'|, h')$. Note that $\mathrm{Con}_h(\mathbb{R}^{n-1}) = \mathbb{H}^n$. The spaces $\mathrm{Con}_h(Z)$ are δ-hyperbolic for some univeral $\delta \geqslant 0$. This is seen by checking the four point condition (3.3) using (10.1), and the argument of the first part of the proof of Theorem 7.2.

If Y is any bounded subset of Z, then (10.1) implies that the set inclusion $i : \mathrm{Con}(Y) \to \mathrm{Con}_h(Z)$ is a roughly isometric embedding of $\mathrm{Con}(Y)$ into $\mathrm{Con}_h(Z)$.

On first thought it seems tempting to expect that perhaps any Gromov hyperbolic space admits a rough similarity into $\mathrm{Con}_h(H)$, where H is a sufficiently large Hilbert space. This seems to be natural, since $\mathrm{Con}_h(H)$ of a Hilbert space H is the infinite dimensional analogue of $\mathbb{H}^n$. However, it follows from results of Enflo [E] (see also [Ra]) that there is a bounded metric spaces (Z,d) such that none of the spaces (Z, d^p), $p \in (0,1]$, has a bilipschitz embedding into any Hilbert space. This can be used to show that $\mathrm{Con}(Z)$ does not admit a rough similarity into $\mathrm{Con}_h(H)$ for any Hilbert space H.

On the other hand, the following is easy.

Theorem 10.4. *For every Gromov hyperbolic space X there is an l_∞-space $l_\infty(A)$ and a rough similarity of X into $\mathrm{Con}_h(l_\infty(A))$.*

Recall that if A is any set, then $l_\infty(A)$ is the space of all bounded real-valued functions on A equipped with the metric coming from the supremum norm.

Proof. Using Theorem 4.1, we assume with no loss of generality that X is a geodesic metric space. Choose a basepoint $o \in X$. To every point $x \in X$ glue a new copy of the ray $[0, \infty)$, with the initial point of the ray identified with x, and call the resulting space $\widehat{X}$. Endow $\widehat{X}$ with the obvious metric, and note that $\widehat{X}$ is Gromov hyperbolic. Then $\widehat{X}$ has the property that for every point $x \in \widehat{X}$ there is a geodesic ray from o passing through x (cf. the proof of Theorem 10.2). Let $Z = \partial\widehat{X}$. Then $\widehat{X} \simeq \mathrm{Con}(Z)$ by Theorem 8.2. Fix a metric d in the canonical gauge on Z. Then the map $f : (Z, d) \to l_\infty(Z)$, $z \mapsto d(z, \cdot)$, is an isometric embedding of (Z, d) into $l_\infty(Z)$. Thus $\mathrm{Con}(Z) \cong \mathrm{Con}(f(Z))$. Since the inclusion $\mathrm{Con}(f(Z)) \subset \mathrm{Con}_h(l_\infty(Z))$ is a rough isometric embedding by the remark following the definition of Con_h, $\widehat{X}$ admits an embedding into $\mathrm{Con}_h(l_\infty(Z))$ by a rough similarity. The same is then true for X. □

11 Some Two-dimensional Applications

In this section a characterization of the hyperbolic plane among geodesic metric spaces up to rough quasi-isometry is given. It would be interesting to have similar statements for higher dimensional hyperbolic spaces as well. We also give a sufficient condition for a Gromov hyperbolic planar graph to be roughly quasi-isometric to a convex subset of $\mathbb{H}^2$.

DEFINITION. A metric space Z is called a λ-*quasi-circle* for $\lambda \geqslant 1$, if it is homeomorphic to the circle $S^1 = \{x \in \mathbb{C} : |x| = 1\}$, and has the property that for any pair of distinct points $x, y \in Z$, the diameter of at least one of the two components of $Z - \{x, y\}$ is at most $\lambda|x - y|$.

We need the following result (cf. [TuV1, Thm. 4.9]).

Theorem 11.1. *A metric space Z is power quasisymmetric to S^1 if and only if it is a quasi-circle and doubling.*

The theorem was not exactly stated like this in [TuV1], but our statement is equivalent. To see this note that a metric space is doubling if and only if it has finite Assouad dimension, and this is equivalent to the condition that the space is homogeneously totally bounded as defined in [TuV1] (cf. Def. 2.7 and Rem. 3.20). Moreover, every quasisymmetry between connected metric spaces is a power quasisymmetry (cf. [TuV1, Cor. 3.12]).

There are quasi-circles which are not doubling. The unit circle S^1 with the metric $d(x, y) = 1/\log(100/|x - y|)$, $x \neq y$, is an example.

Theorem 11.2. *A geodesic metric space X is roughly quasi-isometric to*

the hyperbolic plane $\mathbb{H}^2$ if and only if X is Gromov hyperbolic, visual, has bounded growth at some scale, and ∂X is a quasi-circle.

The property of a metric space being a quasi-circle is preserved by quasisymmetric maps. In particular, for the condition that ∂X is a quasi-circle it does not matter which metric in the canonical gauge on ∂X we choose.

There are Gromov hyperbolic geodesic metric spaces which are visual, have bounded growth at some scale, and for which ∂X is a topological circle, but which are not roughly quasi-isometric to $\mathbb{H}^2$. For example, one can take X to be the hyperbolic convex hull of a topological circle Z in $\partial\mathbb{H}^3$ which is not a quasi-circle. If X and $\mathbb{H}^2$ were roughly quasi-isometric, then $Z = \partial X$ and $\partial\mathbb{H}^2$ would be power quasisymmetric. This is impossible, since $\partial\mathbb{H}^2$ is a quasi-circle (cf. the proof below).

Proof. First note that $\mathbb{H}^2$ is a Gromov hyperbolic geodesic metric space which is visual and has bounded growth at some scale. If we use the unit disc model of the hyperbolic plane, then $\partial\mathbb{H}^2$ can be identified with the unit circle S^1, and it can be shown that the euclidean metric on S^1 is bilipschitz to a metric in the canonical gauge. Therefore, $\partial\mathbb{H}^2$ is a quasi-circle.

Conversely, if a metric space has all the properties stated in the theorem, then $X \simeq \mathrm{Con}(\partial X)$ by Theorem 8.2. By Theorem 9.2 the Assouad dimension of ∂X is finite, and so ∂X is doubling. Hence ∂X is power quasisymmetric to S^1 by Theorem 11.1. By the first part of the proof, we can apply these considerations to $\mathbb{H}^2$. In particular, $\mathbb{H}^2 \simeq \mathrm{Con}(\partial\mathbb{H}^2)$. Moreover, both $\partial\mathbb{H}^2$ and ∂X are power quasisymmetric to S^1, and so there are power quasisymmetric to each other. This implies $\mathrm{Con}(\partial X) \sim \mathrm{Con}(\partial\mathbb{H}^2)$ by Theorem 7.4. We conclude that $X \sim \mathbb{H}^2$. □

Following is an application of Theorem 11.2 to planar graphs. We consider only connected graphs G and think of G as being equipped with a path metric in the usual way, where the edges are intervals of length 1.

Theorem 11.3. *Let G be a Gromov hyperbolic planar graph with bounded vertex degree. Suppose that the union of all geodesics is cobounded in G, and that every compact subset of the plane contains only finitely many vertices of G. Then G is roughly quasi-isometric to a convex subset of $\mathbb{H}^2$.*

The following lemma is probably known, but we have not found a reference.

LEMMA 11.4. *Suppose that a geodesic metric space Z is the union, $Z = X \cup Y$, of two closed subsets $X, Y \subset Z$. Assume that $X \cap Y$ is*

geodesic, and both X and Y are δ-hyperbolic. Then Z is δ'-hyperbolic, where δ' only depends on δ.

A proof can be based on Rips' thin triangles definition. The details are left to the reader.

We shall also need a recent theorem of P. Bowers.

Theorem 11.5 [B]. *Let G be the 1-skeleton of a triangulation of the plane, which is Gromov hyperbolic. Then the Gromov boundary ∂G is either a single point or a topological circle.*

In the latter case, suppose that γ is a geodesic in G, and let G_1, G_2 be the two halves of G "cut" by γ. Then G_1 and G_2 are Gromov hyperbolic and the closed arcs on the circle ∂G determined by the endpoints of γ are the Gromov boundaries of G_1 and G_2.

Proof of 11.3. Assume that G is δ-hyperbolic.

Our first goal is to obtain a better understanding of the planar embedding of G. A *face* is a component of $\mathbb{R}^2 - G$. Let F be a bounded face. Then there are finitely many vertices of G in ∂F. We shall now show that the number of vertices in ∂F is bounded independently of F.

Given any simple closed path α in G, let $D(\alpha)$ denote the closure of the bounded component of $\mathbb{R}^2 - \alpha$. Let n be the least number of vertices in a simple closed path α in G satisfying $D(\alpha) \supset F$. Let α be a simple closed path in G with n vertices such that $D(\alpha) \supset F$ and with $D(\alpha)$ as small as possible. That is, there is no $\alpha' \neq \alpha$ with n vertices and $F \subset D(\alpha') \subset D(\alpha)$. If we choose three almost equally spaced vertices in α, then α becomes a geodesic triangle. Therefore, from the Rips thin triangles condition it follows that there is an upper bound of $O_\delta(1)$ on n, the number of vertices in α.

No geodesic γ in G intersects the interior of $D(\alpha)$, because otherwise we could contradict the minimality of the length of α or of $D(\alpha)$ by replacing an arc of α by an arc of $\gamma \cap D(\alpha)$. Since the union of all geodesics is cobounded G, it follows that there is an upper bound on the distance from any vertex on ∂F to α. Since the degrees of the vertices in G are bounded, and the length n, of α is bounded, it follows that the number of vertices in ∂F is bounded. Consequently, we assume, with no loss of generality, that each bounded face is a triangle (that is, has three vertices on its boundary), because triangulating the bounded faces gives a planar graph roughly isometric to G.

We now deal with the unbounded faces, and prove that with no loss of generality it may be assumed that the boundary of an unbounded face

is a geodesic. Let F be an unbounded face. Then there are infinitely many vertices in ∂F. Fix two vertices a, b in ∂F, and let β be a simple curve in the interior of F with endpoints a, b. Then there is a geodesic arc $\alpha_{a,b}$ joining a and b such that $D(\alpha_{a,b} \cup \beta)$ is minimal with respect to inclusion. It follows that no geodesic enters $D(\alpha_{a,b} \cup \beta)$. Now let a and b tend to infinity along ∂F, "in opposite directions". Since the union of the geodesics is cobounded in G, there is an upper bound on the lower limit of the combinatorial distance of an arbitrary vertex $v \in \partial F$ from $\alpha_{a,b}$. One consequence is that we may extract a limiting geodesic α_F. It also follows that there are no geodesics that meet the component of $\mathbb{R}^2 - \alpha_F$ which contains F. Let G_F be the set of vertices and edges of G that are contained in the component of $\mathbb{R}^2 - \alpha_F$ that contains F, and let G' be the graph obtained from G by deleting G_F for every unbounded face F. Then G' is roughly isometric to G. Consequently, assume with no loss of generality that $G = G'$ and for every unbounded F, its boundary ∂F is a geodesic in G.

Now let Q be a bounded degree triangulation of the upper half-plane that is roughly quasi-isometric to a hyperbolic half-plane and such that ∂Q, the part of Q on the real axis, is a geodesic. For each unbounded face F of G, glue a graph Q_F combinatorially isomorphic to Q, with ∂Q_F identified with the geodesic ∂F, and with orientations agreeing. Let $\widehat{G}$ be the resulting graph.

Suppose that x_1, x_2, x_3, x_4 are four vertices in $\widehat{G}$. Then there are at most four unbounded faces F such that $\{x_1, x_2, x_3, x_4\}$ is contained in G union with the corresponding Q_F's. Consequently, it follows by applying Lemma 11.4 three times that the distances determined by these four points satisfy (3.3) with some constant δ' independent of the choice of the points x_1, x_2, x_3, x_4. Therefore, $\widehat{G}$ is Gromov hyperbolic. Moreover, it is (combinatorially isomorphic to) a triangulation of the plane where the union of the geodesics is cobounded, and G is isometric with a subset of $\widehat{G}$.

From Theorem 11.5 it follows that the Gromov boundary $\partial\widehat{G}$ is a topological circle. We will show that it is a quasi-circle.

Fix a basepoint $o \in \partial\widehat{G}$. If $x, y \in \partial\widehat{G}$, $x \neq y$, there exists a geodesic γ_{xy} in $\widehat{G}$ joining x and y, i.e., one end of γ_{xy} converges to x and one to y. It can be shown that

$$(x|y) = \operatorname{dist}(o, \gamma_{xy}) + O_\delta(1).$$

It follows that if d is some fixed metric in the canonical gauge on $\partial\widehat{G}$, then

there exists $\epsilon > 0$ such that

$$d(x, y) \asymp e^{-\epsilon \operatorname{dist}(o, \gamma_{xy})} . \tag{11.1}$$

Here $\asymp$ means equality up to multiplicative constants independent of x and y.

Let z_1, z_2 be two distinct points in $\partial\widehat{G}$, let $\alpha_1, \alpha_2 \subset \partial\widehat{G}$ be the two closed arcs on $\partial\widehat{G}$ determined by z_1, z_2, and let $w_1 \in \alpha_1, w_2 \in \alpha_2$ be points chosen to maximize $d(z_1, w_1)$ and $d(z_2, w_2)$. Note that $z_1 \neq w_1$, $z_2 \neq w_2$. Let $\gamma \subset \widehat{G}$ be a geodesic joining z_1 and z_2. Then $\widehat{G} - \gamma$ has two components, one, A_1 say, that contains a geodesic ray with limit point w_1, and one, A_2 say, that contains a geodesic ray with limit point w_2.

Changing notation if necessary, we may assume that $o \in \gamma \cup A_1$. The geodesic $\beta \subset X$ joining w_2 and z_2 lies entirely in $\gamma \cup A_2$. In order to reach β from o, one must cross γ. Therefore, $\operatorname{dist}(o, \gamma) \leqslant \operatorname{dist}(o, \beta)$. By (11.1) this implies

$$\begin{aligned} d\text{-}\operatorname{diam}(\alpha_2) &\leqslant 2d(z_2, w_2) \leqslant c_1 e^{-\epsilon \operatorname{dist}(o,\beta)} \\ &\leqslant c_1 e^{-\epsilon \operatorname{dist}(o,\gamma)} \leqslant c_2 d(z_1, z_2) . \end{aligned}$$

Here the constants $c_1, c_2 > 0$ can be chosen independently of the points z_1, z_2, w_1, w_2. This shows that $\partial\widehat{G}$ is a quasi-circle.

The graph $\widehat{G}$ is visual, as easily follows from the fact that the union of the geodesics is cobounded in $\widehat{G}$. Since $\widehat{G}$ has bounded degree, it has bounded growth at some scale, and $\partial\widehat{G}$ has finite Assouad dimension. Theorem 11.2 then tells us that $\widehat{G}$ is roughly quasi-isometric to $\mathbb{H}^2$. Consequently, G is quasi-isometric to a subset, say W of $\mathbb{H}^2$. As in the proof of Theorem 10.2, it can be shown that W is roughly isometric to hull(W). The theorem follows. □

COROLLARY 11.6. *Assume additionally in Theorem 11.3 that ∂G is connected. Then G is roughly quasi-isometric to the hyperbolic plane, or to a hyperbolic half-plane.*

The proof easily follows from Theorem 11.3, and is left to the reader.

References

[A] P. ASSOUAD, Plongements lipschitziens dans $\mathbb{R}^n$, Bull. Soc. Math. France 111 (1983), 429–448.

[B] P.L. BOWERS, Negatively curved graph and planar metrics with applications to type, Michigan Math. J. 45 (1998), 31–53.

[CDP] M. Coornaert, T. Delzant, A. Papadopoulos, Geométrié et théorie des groupes. Les groupes hyperboliques de Gromov, Springer Lecture Notes in Mathematics 1441, Berlin, 1990.

[E] P. Enflo, On a problem of Smirnov, Ark. Mat. 8 (1969), 107–109.

[GH] E. Ghys, P. de la Harpe, eds., Sur les Groupes Hyperboliques d'après Mikhael Gromov, Progress in Mathematics 83, Birkhäuser, Boston, 1990.

[Gr] M. Gromov, Hyperbolic groups, in "Essays in Group Theory" (G. Gersten, ed.), Math. Sci. Res. Inst. Publ. Springer (1987), 75–263.

[L] J. Luukkainen, Assouad dimension, antifractal metrization, porous sets, and homogeneous measures, J. Korean Math. Soc. 35 (1998), 23–76.

[P] F. Paulin, Un groupe hyperbolique est déterminé par son bord, J. London Math. Soc. 54 (1996), 50–74.

[R] J.G. Ratcliffe, Foundations of Hyperbolic Manifolds, Graduate Texts in Mathematics 149, Springer, New York, 1994.

[Ra] Y. Raynaud, Espaces de Banach superstables, distances stables et homéomorphismes uniformes, Israel J. Math. 44 (1983), 33–52.

[TV] D.A. Trotsenko, J. Väisälä, Upper sets and quasisymmetric maps, preprint.

[TuV1] P. Tukia, J. Väisälä, Quasisymmetric embeddings of metric spaces, Ann. Acad. Sci. Fenn. Ser. A I Math. 5 (1980), 97–114.

[TuV2] P. Tukia, J. Väisälä, Quasiconformal extension from dimension n to $n+1$, Ann. Math. 115 (1982), 331–348.

[V] J. Väisälä, Quasisymmetric embeddings in Euclidean spaces, Trans. Amer. Math. Soc. 264 (1981), 191–204.

Mario Bonk, Institut für Analysis, Technische Universität Braunschweig, D-38106 Braunschweig, Germany
M.Bonk@tu-bs.de

Oded Schramm, Mathematics Department, The Weizmann Institute, Rehovot 76100, Israel
schramm@wisdom.weizmann.ac.il

Submitted: October 1998
Revised version: January 1999

Correction to "Embeddings of Gromov hyperbolic spaces". The parts of the proof of Theorem 9.2 related to the relations (9.4) are somewhat garbled, but it is easy to correct the argument. The idea is to find for each of the points z_j suitable associated points x_j and y_j in X. The points $x_1, \ldots, x_n$ should be close together, the distance $|x_j - y_j|$ should be controlled, but we want the definite separation $|y_i - y_j| > 2r$ for $i \neq j$. As in the paper, one chooses x_j for $j = 1, \ldots, n$ such that $|x_j| = -\log \beta$ and $|-\log \beta - (x_j|z_j)| \leq C(\delta)$. Then $|x_i - x_j| = O_\delta(1)$ as follows from Lemma 5.1.

The point y_j is chosen such that $|y_j| = t$ and $|(y_j|z_j) - t| \leq C(\delta)$ for appropriate $t \geq 0$. If $t \geq -\log \alpha$, then again Lemma 5.1 shows that

$$|x_j - y_j| = t + \log \beta + O_\delta(1)$$

and

$$|y_i - y_j| \geq 2t + 2\log \alpha + O_\delta(1)$$

for $i \neq j$. If we pick $t = -2\log \alpha + \log \beta + C(\delta)$, then

$$|x_j - y_j| = 2(-\log \alpha + \log \beta) + O_\delta(1)$$

and

$$|y_j - y_i| > 2\log(\beta/\alpha).$$

Here we may assume that $\log(\beta/\alpha) \geq r$, because if this inequality is not true, then we replace α by a smaller quantity that is comparable to α up to a factor only depending on r. The rest of the argument is as in the paper.

Mario Bonk

Part II
Noise Sensitivity

Oded Schramm's contributions to Noise Sensitivity

Christophe Garban *

Abstract

We survey in this paper the main contributions of Oded Schramm related to *Noise Sensitivity*. We will describe in particular his various works which focused on the "Spectral analysis" of critical percolation (and more generally of Boolean functions), his work on the shape-fluctuations of first passage percolation and finally his contributions to the model of dynamical percolation.

This paper is dedicated to the memory of Oded Schramm. I feel very fortunate to have known him.

A sentence which summarizes well in what consisted of Oded's work on Noise Sensitivity is the following quote from Jean Bourgain.

There is a general philosophy which claims that if f defines a property of 'High Complexity', then Sup$\hat{f}$*, the support of the Fourier Transform, has to be 'spread out'.*

Through his work on models coming from Statistical physics (in particular percolation), Oded Schramm was often confronted with such functions of *'High complexity'*. For example, in percolation, any large scale connectivity property can be encoded by a Boolean function of the 'inputs' (edges or sites). At criticality, these large scale connectivity functions turn out to be of 'High Frequency' which gives deep information on the underlying model. As we will see along this survey, Oded Schramm developed over the last decade highly original and deep ideas to understand the 'complexity' of Boolean functions.

We will essentially follow the chronology of his contributions in the field; it is quite striking that three distinct periods emerge from Oded's work and they will constitute sections 3, 4 and 5 of this survey, corresponding respectively to the papers [BKS99, SS10b, GPS10].

We have chosen to present numerous sketches of proof, since we believe that the elegance of Oded's mathematics is best illustrated through his proofs, which usually combined imagination, simplicity and powerfulness. This choice is the reason for the length of the present survey.

*UMPA, CNRS UMR 5669, ENS Lyon. Research supported in part by ANR-06-BLAN-00058 and the Ostrowski foundation

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_11, © Springer Science+Business Media, LLC 2011

Contents

1 Introduction **2**
1.1 Historical context . . . 3
1.2 Concept of Noise Sensitivity . . . 4
1.3 Motivations from statistical physics . . . 6
1.4 Precise definition of Noise Sensitivity . . . 8
1.5 Structure of the paper . . . 9

2 Background **9**
2.1 Fourier analysis of Boolean functions . . . 10
2.2 Notion of Influence . . . 12
2.3 Fast review on Percolation . . . 14

3 The "hypercontractivity" approach **18**
3.1 About Hypercontractivity . . . 19
3.2 Applications to Noise sensitivity . . . 21
3.3 Anomalous fluctuations, or Chaos . . . 23

4 The Randomized Algorithm approach **28**
4.1 First appearance of randomized algorithm ideas . . . 29
4.2 The Schramm/Steif approach . . . 29
4.3 Is there any better algorithm ? . . . 35
4.4 Related works of Oded an randomized algorithms . . . 36

5 The "geometric" approach **37**
5.1 Rough description of the approach and "spectral measures" . . . 37
5.2 First and second moments on the size of the spectral sample . . . 40
5.3 "Scale invariance" properties of $\mathscr{S}_{f_n}$ . . . 41
5.4 Strategy assuming a good control on the dependency structure . . . 43
5.5 The weak dependency control we achieved and how to modify the strategy . . . 47

6 Applications to Dynamical Percolation **49**
6.1 The model of dynamical percolation . . . 49
6.2 Second moment analysis and the "exceptional planes of percolation" . . . 51
6.3 Dynamical percolation, later contributions . . . 53

1 Introduction

In this introduction, we will start by describing the scientific context in the late nineties which lead Oded Schramm to work on the sensitivity of Boolean functions. We will then motivate and define the notion of *Noise Sensitivity* and finally we will review the main contributions of Oded that will concern us throughout this survey.

1.1 Historical context

When Oded started working on Boolean functions (with the idea to use them in statistical mechanics), there was already important literature in computer science dedicated to the properties of Boolean functions.

Here is an example of a related problem which was solved before Oded came into the field: in [BOL87] it was conjectured that if f is any Boolean function on n variables (i.e. $f : \{0,1\}^n \to \{0,1\}$), taking the value 1 for half of the configurations of the hypercube $\{0,1\}^n$; then there exists some deterministic set $S = S_f \subset \{1,\dots,n\}$ of less than $c\frac{n}{\log n}$ variables, such that f remains undetermined as long as the variables in S are not assigned (the constant c being a universal constant). This means that for any such Boolean function f, there should always exist a set of small size which is "pivotal" for the outcome.

This conjecture was proved in [KKL88]. Beside the proof of the conjecture, what is most striking about this paper (and which will concern us throughout the rest of this survey) is that, for the first time, techniques brought from Harmonic Analysis were used in [KKL88] to study properties of Boolean functions. The authors wrote at the time:

These new connections with harmonic analysis are very promising.

Indeed, as they anticipated, the main technique they borrowed from harmonic analysis, namely **hypercontractivity**, was later used in many subsequent works. In particular, as we will see later (section 3), hypercontractivity was one of the main ingredients in the landmark paper on *noise sensitivity* written by Benjamini, Kalai and Schramm: [BKS99].

Before going into [BKS99] (which introduced the concept of *Noise sensitivity*), let us mention some of the related works in this field which appeared in the period from [KKL88] to [BKS99] and which made good use of hypercontractivity. We distinguish several directions of research.

- First of all, the results of [KKL88] have been extended to more general cases: non-balanced Boolean functions, other measures than the uniform measure on the hypercube $\{0,1\}^n$ and finally, generalizations to "Boolean" functions of continuous dependence $f : [0,1]^n \to \{0,1\}$. See [BKK+92] as well as [Tal94]. Note that both of these papers rely on hypercontractive estimates.

- Based on these generalizations, Friedgut and Kalai studied in [FK96] the phenomenon of "Sharp Thresholds". Roughly speaking, they proved that any monotone "graph property" for the so-called random graphs $G(n,p)$, $0 \le p \le 1$ satisfies a *sharp threshold* as the number of vertices n goes to infinity; see [FK96] for a more precise statement. In other words, they show that any monotone graph event $\mathcal{A}$ appears "suddenly" while increasing the edge intensity p (whence a "cut-off" like shape for the function $p \mapsto \mathbb{P}_{n,p}(\mathcal{A})$). In some sense their work is thus intimately related to the subject of this survey, since many examples of such "graph properties" concern connectedness, size of clusters and so on..

  The study of *Sharp Thresholds* began in Percolation theory with the seminal

work of Russo [Rus78, Rus82], where he introduced the idea of sharp thresholds in order to give an alternate proof of Kesten's result $p_c(\mathbb{Z}^2) = \frac{1}{2}$. See also the paper by Margulis [Mar74].

- Finally, Talagrand made several remarkable contributions over this period which highly influenced Oded and his coauthors (as we will see in particular through section 3). To name a few of these: an important result of [BKS99] (theorem 3.3 in this survey) was inspired by [Tal96]; the study of fluctuations in First Passage Percolation in [BKS03] (see section 3) was influenced by a result from [Tal94] (this paper by Talagrand was already mentioned above since it somewhat overlapped with [BKK+92, FK96]). More generally the questions addressed along this survey are related to the *concentration of measure phenomenon* which has been deeply understood by Talagrand, see [Tal95].

1.2 Concept of Noise Sensitivity

It is now time to motivate and then define what is *Noise Sensitivity*. This concept was introduced in the seminal work [BKS99] whose content will be described in section 3. As one can see from its title: *Noise sensitivity of Boolean functions and applications to Percolation*, the authors introduced this notion having in mind applications in percolation. Before explaining these motivations in the next subsection, let us consider the following simple situation.

Assume we have some initial data $\omega = (x_1, \ldots, x_n) \in \{0,1\}^n$ and we are interested in some functional of this data that we may represent by a (real-valued) function $f : \{0,1\}^n \to \mathbb{R}$. Often, the functional f will be the indicator function of an event $A \subset \{0,1\}^n$; in other words f will be **Boolean**. Now imagine that there are some errors in the transmission of this data ω and that one receives only the sightly perturbed data ω^ϵ; one would like that the quantity we are interested in, i.e. $f(\omega)$, is not "too far" from what we actually receive, i.e. $f(\omega^\epsilon)$. (A similar situation would be: we are interested in some *real* data ω but there are some inevitable errors in collecting this data and one ends up with ω^ϵ). To take a typical example, if $f : \{0,1\}^n \to \{0,1\}$ represents a *voting scheme*, for example Majority, then it is natural to wonder how robust f is with respect to errors ?

At that point, one would like to know how to properly model the perturbed data ω^ϵ ? The correct answer depends on the probability law which governs our "random" data ω. In the rest of this survey, our data will always be sampled according to the uniform measure, hence it will be sufficient for us to assume that $\omega \sim \mathbb{P} = \mathbb{P}^n = (\frac{1}{2}\delta_0 + \frac{1}{2}\delta_1)^{\otimes n}$, the uniform measure on $\{0,1\}^n$. Other natural measures may be considered instead, but we will stick to this simpler case. Therefore a natural way to model the perturbed data ω^ϵ is to assume that each variable in $\omega = (x_1, \ldots, x_n)$ is **resampled** independently with small probability ϵ. Equivalently, if $\omega = (x_1, \ldots, x_n)$, then ω^ϵ will correspond to the random configuration $(y_1, \ldots, y_n)$, where independently for each $i \in \{1, \ldots, n\}$, with probability $1 - \epsilon$, $y_i := x_i$ and with probability ϵ, y_i is sampled according to a Bernoulli(1/2). It is clear that such a "noising" procedure preserves the uniform measure on $\{0,1\}^n$.

In computer science, one is naturally interested in the *Noise stability* of f which, if f is Boolean (i.e. with values in $\{0,1\}$), can be measured by the quantity $\mathbb{P}\big[f(\omega) \neq f(\omega^\epsilon)\big]$, where as above $\mathbb{P} = \mathbb{P}^n$ denotes the uniform measure on $\{0,1\}^n$ (in fact there is a slight abuse of notation here, since $\mathbb{P}$ samples the coupling $(\omega, \omega^\epsilon)$ hence there is also some extra randomness needed for the randomization procedure). In [MOO05], it is shown using Fourier analysis that in some sense, the most stable Boolean function on $\{0,1\}^n$ is the Majority function (under reasonable assumptions which exclude Dictatorship and so on).

If a functional f happens to be "noise stable", this means that very little information is lost on the outcome $f(\omega)$ knowing the "biased" information ω^ϵ. Throughout this survey, we will be mainly interested in the opposite property, namely *Noise Sensitivity*. We will give precise definitions later, but roughly speaking, f will be called "noise sensitive" if almost ALL information is lost on the outcome $f(\omega)$ knowing the biased information ω^ϵ. This complete loss of information can be measured as follows: f will be "noise sensitive" if $\mathrm{Var}\big[\mathbb{E}\big[f(\omega) \mid \omega^\epsilon\big]\big] = o(1)$. It turns out that if $\mathrm{Var}(f) = O(1)$, it is equivalent to consider the correlation between $f(\omega)$ and $f(\omega^\epsilon)$ and equivalently, f will be called "noise sensitive" if $\mathrm{Cov}(f(\omega), f(\omega^\epsilon)) = \mathbb{E}\big[f(\omega)f(\omega^\epsilon)\big] - \mathbb{E}\big[f\big]^2 = o(1)$.

Remark 1.1. Let us point out that *noise stability* and *noise sensitivity* are two extreme situations. Imagine our functional f can be decomposed as $f = f_{\text{stable}} + f_{\text{sens}}$, then after transmission of the data, we will keep the information on f_{stable} but lose the information on the other component. As such f will be neither stable nor sensitive.

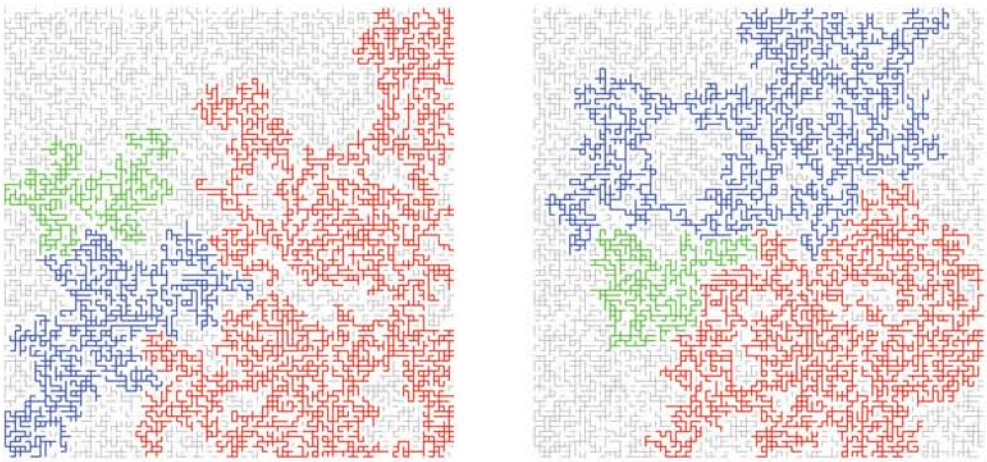


Figure 1.1:

We will end by an example of a noise sensitive functional in the context of percolation. Let us consider a percolation configuration ω_n on the rescaled lattice $\frac{1}{n}\mathbb{Z}^2$ in the window $[0,1]^2$ (see subsection 2.3 for background and references on percolation)

at criticality $p_c = \frac{1}{2}$ (hence ω_n is sampled according to the uniform measure on $\{0,1\}^{E_n}$, where E_n is the set of edges of $\frac{1}{n}\mathbb{Z}^2 \cap [0,1]^2$). In figure 1.1, we represented a percolation configuration ω_{90} and its noised configuration ω_{90}^{ϵ} with $\epsilon = \frac{1}{10}$. In each picture, the three largest clusters are colored in Red, Blue, Green. As is suggested from this small scale picture, the functional giving the size of the largest cluster turns out to be asymptotically "noise sensitive"; i.e. for fixed $\epsilon > 0$ when $n \to \infty$ the entire information about the largest cluster is lost from ω_n to ω_n^{ϵ} (even though most of the "microscopic" information is preserved).

For nice applications of the concept of Noise sensitivity in the context of computer science, see [KS06] and [O'D03].

1.3 Motivations from statistical physics

Beside the obvious motivations in computer science, there were several motivations coming from Statistical physics (mostly from percolation) which lead Benjamini, Kalai and Schramm to introduce and study *noise sensitivity* of Boolean functions. We wish to list some of these in this subsection.

• *Dynamical percolation*
In "real life", models issued from statistical physics undergo *thermal agitation*; consequently, their state evolve in time. For example, in the case of the Ising model, the natural dynamics associated to thermal agitation is the so-called *Glauber Dynamics*.

In the context of percolation, a natural dynamics modeling this *thermal agitation* has been introduced in [HPS97] under the name of **dynamical percolation** (it was also invented independently by Itai Benjamini). This model is defined in a very simple way and we will describe it in detail in section 6; see [Ste09] for a very nice survey.

For percolation in dimension two, it is known that at *criticality*, there is almost surely no infinite cluster. Nevertheless, the following open question was asked back in [HPS97]: if one lets the critical percolation configuration evolve according to this dynamical percolation process, is it the case that there will exist **exceptional times** where an infinite cluster will appear ? As we will later see, such exceptional times indeed exist. This reflects the fact that the *dynamical* behavior of percolation is very different from its *static* behavior.

Dynamical percolation is intimately related to our concept of Noise sensitivity since if (ω_t) denotes the trajectory of a dynamical percolation process, then the configuration at time t is exactly a "noised" configuration of the initial configuration ω_0 ($\omega_t \equiv \omega_0^{\epsilon}$ with an explicit correspondence between t and ϵ, see section 6).

As is suggested by the picture 1.1, large clusters of $(\omega_t)_{t\geq 0}$ "move" (or change) very quickly as the time t goes on. This **rapid mixing** of the large scale properties of ω_t is the reason for the appearance of "exceptional" infinite clusters along the dynamics. Hence, as we will see in section 6, the above question addressed in [HPS97] needs an analysis of the *noise sensitivity* of large clusters. In cite [BKS99], the first results in this direction are shown (they prove that in some sense any large scale connectivity property is "noise sensitive"). However, their control on the sensitivity of large scale properties (later we will rather say: their control on the "frequency"

of large scale properties) was not sharp enough to imply that exceptional times do exist. The question was finally answered in the case of the triangular lattice in [SS10b] thanks to a better understanding of the sensitivity of percolation. The case of $\mathbb{Z}^2$ was carried out in [GPS10]. This conjecture from [HPS97] on the existence of **exceptional times** will be one of our driving motivations throughout this survey.

• *Conformal Invariance of percolation*
In the nineties, an important conjecture popularized by Langlands et al in [LPSA94] stated that critical percolation should satisfy some (asymptotic) *conformal invariance* (in the same way as Brownian motion does). Using Conformal Field Theory, John Cardy made several important predictions based on this conformal invariance principle, see [Car92].

Conformal invariance of percolation was probably the main conjecture on planar percolation at that time and people working in this field, including Oded, were working actively on this question until Stanislav Smirnov solved it on the triangular lattice in [Smi01] (note that a major motivation of the introduction of the SLE_κ processes by Oded in [Sch00] was this conjecture).

At the time of [BKS99], conformal invariance of percolation was not yet proved (it still remains open on $\mathbb{Z}^2$), and the SLEs from [Sch00] were "in construction", so the route towards conformal invariance was still vague. One nice idea from [BKS99] in this direction was to randomly perturb the lattice itself and then claim that the crossing probabilities are almost not affected by the random pertubation of the lattice. This setup is easily seen to be equivalent to proving Noise sensitivity. What was difficult and remained to be done was to show that one could use such random perturbations to go from one "quad" to another conformally equivalent "quad"; see [BKS99] for more details. Note that at the same period, a different direction to attack conformal invariance was developed by Benjamini and Schramm in [BS98a] where they proved a certain kind of conformal invariance for Voronoï percolation.

• *Tsirelson's "Black noise"*
In [TV98], Tsirelson and Vershik constructed sigma-fields which are not "produced" by White-noise (they called these sigma-fields "nonlinearizable" at that time). Tsirelson then realized that a good way to characterize a "Brownian filtration" was to test its stability versus perturbations. Systems intrinsically unstable led him to the notion of *Black noise*; see [Tsi04] for more details. This instability corresponds exactly to our notion of being "noise sensitive", and after [BKS99] appeared, Tsirelson used the concept of *noise sensitivity* to describe his theory of black noises. Finally, black noises are related to percolation, since according to Tsirelson himself, percolation would provide the most important example of a (two-dimensional) Black noise. In some sense [BKS99] shows that if percolation can be seen asymptotically as a *noise* (i.e. a continuous sigma-field which "factorizes", see [Tsi04]), then this noise has to be *black* i.e. all its observables or functionals are noise sensitive. The remaining step (proving that percolation asymptotically factorizes as a *noise*) was proved recently by Schramm and Smirnov ([SS10a]).

• *Anomalous fluctuations*
In a different direction, we will see that Noise Sensitivity is related to the study of

random shapes whose fluctuations are much smaller than "gaussian" fluctuations (this is what we mean by "anomalous fluctuations"). Very roughly speaking, if a random metric space is highly sensitive to noise (in the sense that its geodesics are "noise sensitive"), then it induces a lot of independence within the system itself and the metric properties of the system decorrelate fast (in space or under perturbations). This instability implies in general very small fluctuations for the macroscopic metric properties (like the shape of large balls and so on). In section 3, we will give an example of a random metric space, First Passage Percolation, whose fluctuations can be analyzed with the same techniques as the ones used in [BKS99].

1.4 Precise definition of Noise Sensitivity

Let us now fix some notations and definitions that will be used throughout the rest of the article, especially the definition of *noise sensitivity* which was only informal so far.

First of all, from now on, it will be more convenient to work with the hypercube $\{-1,1\}^n$ rather than with $\{0,1\}^n$. Let us then call $\Omega_n := \{-1,1\}^n$. The advantage of this choice is that the *characters* of $\{-1,1\}^n$ have a more simple form.

In the remainder of the text a **Boolean function** will denote a function from Ω_n into $\{0,1\}$ (except in section 5, where when made precise it could also be into $\{-1,1\}$) and as argued above, Ω_n will be endowed with the uniform probability measure $\mathbb{P} = \mathbb{P}^n$ on Ω_n. Through this survey, we will sometimes extend the study to the larger class of real-valued functions from Ω_n into $\mathbb{R}$. Some of the results will hold for this larger class, but the Boolean hypothesis will be crucial at several different steps.

In the above informal discussion, "noise sensitivity" of a function $f : \Omega_n \to \{0,1\}$ corresponded to $\mathrm{Var}\big[\mathbb{E}\big[f(\omega) \;\big|\; \omega^\epsilon\big]\big]$ or $\mathrm{Cov}(f(\omega), f(\omega^\epsilon))$ being "small". To be more specific, *noise sensitivity* is defined in [BKS99] as an asymptotic property.

Definition 1.2 ([BKS99])**.** *Let $(m_n)_{n\geq 0}$ be an increasing sequence in $\mathbb{N}$. A sequence of Boolean functions $f_n : \{-1,1\}^{m_n} \to \{0,1\}$ is said to be (asymptotically)* **noise sensitive** *if for any level of noise $\epsilon > 0$,*

$$\mathrm{Cov}(f_n(\omega), f_n(\omega^\epsilon)) = \mathbb{E}\big[f_n(\omega) f_n(\omega^\epsilon)\big] - \mathbb{E}\big[f_n\big]^2 \underset{n\to\infty}{\longrightarrow} 0\,. \tag{1.1}$$

In [BKS99], the asymptotic condition was rather that

$$\mathrm{Var}\big[\mathbb{E}\big[f_n(\omega) \;\big|\; \omega^\epsilon\big]\big] \underset{n\to\infty}{\longrightarrow} 0$$

but as we mentioned above, the definitions are easily seen to be equivalent (using the Fourier expansions of f_n).

Remark 1.3. One can extend in the same fashion this definition to the class of (real-valued) functions $f_n : \Omega_{m_n} \to \mathbb{R}$ of bounded variance (bounded variance is needed to guarantee the equivalence of the two above criteria).

Remark 1.4. Note that if $\mathrm{Var}(f_n)$ goes to zero as $n \to \infty$, then (f_n) will automatically satisfy the condition (1.1), hence our definition of noise sensitivity is meaningful only for non-degenerate asymptotic events.

The opposite notion of *noise stability* is defined in [BKS99] as follows

Definition 1.5 ([BKS99])**.** *Let $(m_n)_{n\geq 0}$ be an increasing sequence in $\mathbb{N}$. A sequence of Boolean functions $f_n : \{-1,1\}^n \to \{0,1\}$ is said to be (asymptotically)* **Noise stable** *if*

$$\sup_{n\geq 0} \mathbb{P}\big[f_n(\omega) \neq f_n(\omega^\epsilon)\big] \underset{\epsilon\to 0}{\longrightarrow} 0.$$

1.5 Structure of the paper

In section 2, we will review some necessary background: Fourier analysis of Boolean functions, the notion of *influence* and some facts about percolation. Then the three sections 3, 4 and 5 form the core of this survey. They present three different approaches, each of them enabling to localize with more or less accuracy the "Frequency domain" of percolation.

The approach presented in section 3 is based on a technique, *hypercontractivity*, brought from harmonic analysis. Following [BKS99] and [BKS03] we apply this technique to the sensitivity of percolation as well as to the study of shape fluctuations in First Passage Percolation. In section 4, we describe an approach developed by Schramm and Steif in [SS10b] based on the analysis of randomized algorithms. Section 5, following [GPS10], presents an approach which considers the "Frequencies" of percolation as random sets in the plane; the purpose is then to study the Law of these "Frequencies" and to prove that they behave in some ways like random Cantor sets.

Finally, in section 6 we present applications of the detailed information provided by the last two approaches ([SS10b] and [GPS10]) to the model of dynamical percolation.

The contributions that we have chosen to present reveal personal tastes of the author. Also, the focus here is mainly on the applications in statistical mechanics and particularly percolation. Nevertheless we will try as much as possible, along this survey, to point towards other contributions Oded made close to this theme (such as [BS98b, PSSW07, BSW05, OSSS05]). See also the very nice survey by Oded [Sch07].

Acknowledgments

I wish to warmly thank Jeff Steif who helped tremendously in improving the readability of this survey. The week he invited me to spend in Göteborg was a great source of inspiration for writing this paper. I would also like to thank Gábor Pete: since we started our collaboration three years ago with Oded, it has always been a pleasure to work with him. Finally, I would like to thank Itai Benjamini who kindly invited me to write the present survey.

2 Background

In this section, we will give some preliminaries on Boolean functions which will be used throughout the rest of the present survey. We will start with an analog of

Fourier Series for Boolean functions; then we will define the **Influence** of a variable, a notion which will be particularly relevant in the remainder of the text; we will end the preliminaries section with a brief review on percolation since most of Oded's work in Noise Sensitivity was motivated by applications in percolation theory.

2.1 Fourier analysis of Boolean functions

In this survey, recall that we consider our Boolean functions as functions from the hypercube $\Omega_n := \{-1,1\}^n$ into $\{0,1\}$, and Ω_n will be endowed with the uniform measure $\mathbb{P} = \mathbb{P}^n = (\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1)^{\otimes n}$.

Remark 2.1. In greater generality, one could consider other natural measures like $\mathbb{P}_p = \mathbb{P}_p^n = ((1-p)\delta_{-1} + p\delta_1)^{\otimes n}$; these measures are relevant for example in the study of sharp thresholds (where one increases the "level" p). In the remainder of the text, it will be sufficient for us to restrict to the case of the uniform measure $\mathbb{P} = \mathbb{P}_{1/2}$ on Ω_n.

In order to apply Fourier analysis, the natural setup is to enlarge our discrete space of Boolean functions and to consider instead the larger space $L^2(\{-1,1\}^n)$ of real-valued functions on Ω_n endowed with the inner product:

$$\begin{aligned} \langle f, g \rangle &:= \sum_{x_1,\ldots,x_n} 2^{-n} f(x_1,\ldots,x_n) g(x_1,\ldots,x_n) \\ &= \mathbb{E}\big[fg\big] \text{ for all } f, g \in L^2(\Omega_n)\,, \end{aligned}$$

where $\mathbb{E}$ denotes the expectation with respect to the uniform measure $\mathbb{P}$ on Ω_n.

For any subset $S \subset \{1, 2 \ldots, n\}$, let χ_S be the function on $\{-1,1\}^n$ defined for any $x = (x_1, \ldots, x_n)$ by

$$\chi_S(x) := \prod_{i \in S} x_i\,. \tag{2.1}$$

It is straightforward to see that this family of 2^n functions forms an orthonormal basis of $L^2(\{-1,1\}^n)$. Thus, any function f on Ω_n (and a fortiori any Boolean function f) can be decomposed as

$$f = \sum_{S \subset \{1,\ldots,n\}} \hat{f}(S)\, \chi_S,$$

where $\hat{f}(S)$ are the Fourier coefficients of f. They are sometimes called the **Fourier-Walsh** coefficients of f and they satisfy

$$\hat{f}(S) := \langle f, \chi_S \rangle = \mathbb{E}\big[f \chi_S\big].$$

Note that $\hat{f}(\emptyset)$ corresponds to the average $\mathbb{E}\big[f\big]$. As in classical Fourier analysis, if f is some Boolean function, its Fourier(-Walsh) coefficients provide information on the "regularity" of f.

Of course one may find many other orthonormal bases for $L^2(\{-1,1\}^n)$, but there are many situations for which this particular set of functions $(\chi_S)_{S \subset \{1,\ldots,n\}}$ arises

naturally. First of all there is a well-known theory of Fourier analysis on groups, a theory which is particularly simple and elegant on Abelian groups (thus including our special case of $\{-1,1\}^n$, but also $\mathbb{R}/\mathbb{Z}$, $\mathbb{R}$ and so on). For Abelian groups, what turns out to be relevant for doing harmonic analysis is the set $\hat{G}$ of **characters** of G (i.e. the group homomorphisms from G to $\mathbb{C}^*$). In our case of $G = \{-1,1\}^n$, the characters are precisely our functions χ_S indexed by $S \subset \{1,\dots,n\}$ since they satisfy $\chi_S(x \cdot y) = \chi_S(x)\chi_S(y)$.

These functions also arise naturally if one performs simple random walk on the hypercube (equipped with the Hamming graph structure), since they are the eigenfunctions of the heat kernel on $\{-1,1\}^n$.

Last but not least, the basis (χ_S) turns out to be particularly adapted to our study of **noise sensitivity**. Indeed if $f : \Omega_n \to \mathbb{R}$ is any real-valued function, then the correlation between $f(\omega)$ and $f(\omega^\epsilon)$ is nicely expressed in terms of the Fourier coefficients of f as follows:

$$\begin{aligned}\mathbb{E}\big[f(\omega)f(\omega^\epsilon)\big] &= \mathbb{E}\Big[\Big(\sum_{S_1}\hat{f}(S_1)\chi_{S_1}(\omega)\Big)\Big(\sum_{S_2}\hat{f}(S_2)\chi_{S_2}(\omega^\epsilon)\Big)\Big] \\ &= \sum_S \hat{f}(S)^2\mathbb{E}\big[\chi_S(\omega)\,\chi_S(\omega^\epsilon)\big] \\ &= \sum_S \hat{f}(S)^2(1-\epsilon)^{|S|}\,. \end{aligned} \tag{2.2}$$

Therefore, the "level of sensitivity" of a Boolean function is naturally encoded in its Fourier coefficients. More precisely, for any real-valued function $f : \Omega_n \to \mathbb{R}$, one can consider its **Energy Spectrum** E_f defined by

$$E_f(m) := \sum_{|S|=m} \hat{f}(S)^2,\ \forall\, m \in \{1,\dots,n\}\,.$$

Since $\mathrm{Cov}(f(\omega), f(\omega^\epsilon)) = \sum_{m=1}^n E_f(m)(1-\epsilon)^m$, all the information we need is contained in the Energy Spectrum of f. As argued in the introduction, a function of "high frequency" will be sensitive to noise while a function of "low frequency" will be stable. This allows us to give an equivalent definition of *noise sensitivity* (recall definition 1.2):

Proposition 2.2. *A sequence of Boolean functions $f_n : \{-1,1\}^{m_n} \to \{0,1\}$ is (asymptotically)* **noise sensitive** *if and only if, for any $k \geq 1$*

$$\sum_{m=1}^{k}\sum_{|S|=m}\hat{f}_n(S)^2 = \sum_{m=1}^{k} E_{f_n}(m) \underset{n\to\infty}{\longrightarrow} 0\,.$$

Before introducing the notion of influence, let us give a simple example:

Example: let Φ_n be the Majority function on n variables (a function which is of obvious interest in computer science). $\Phi_n(x_1,\dots,x_n) := \mathrm{sign}(\sum x_i)$, where n is

an odd integer here. It is possible in this case to explicitly compute its Fourier coefficients and when n goes to infinity, one ends up with the following asymptotic formula (see [O'D03] for a nice overview and references therein):

$$E_{\Phi_n}(m) = \sum_{|S|=m} \widehat{\Phi_n}(S)^2 = \begin{cases} \frac{4}{\pi\, m 2^m}\binom{m-1}{\frac{m-1}{2}} + O(m/n) & \text{if } m \text{ is odd}\,, \\ 0 & \text{if } m \text{ is even}\,. \end{cases}$$

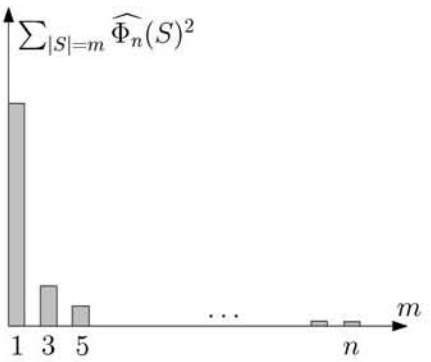


Figure 2.1:

Picture 2.1 represents the shape of the Energy Spectrum of Φ_n: its Spectrum is concentrated on low frequencies which is typical of stable functions.

From now on, most of this paper will be concerned with the description of the Fourier expansion of Boolean functions (and more specifically of their Energy Spectrum).

2.2 Notion of Influence

If $f : \Omega_n \to \mathbb{R}$ is a (real-valued) function, the influence of a variable $k \in [n]$ is a quantity which measures by how much (on average) the function f depends on the fixed variable k. For this purpose, we introduce the functions

$$\nabla_k f : \begin{cases} \Omega_n & \to & \mathbb{R} \\ \omega & \mapsto & f(\omega) - f(\sigma_k \cdot \omega) \end{cases} \quad \text{for all } k \in [n]\,,$$

where σ_k acts on Ω_n by flipping the kth bit (thus $\nabla_k f$ corresponds to a discrete derivative along the kth bit).

The **Influence** $\mathbf{I}_k(f)$ of the kth variable is defined in terms of $\nabla_k f$ as follows

$$\mathbf{I}_k(f) := \|\nabla_k f\|_1 = \|f(\omega) - f(\sigma_k \cdot \omega)\|_1\,.$$

Remark 2.3. If $f : \Omega_n \to \{0,1\}$ is a Boolean function corresponding to an event $A \subset \Omega_n$ (i.e. $f = 1_A$), then $\mathbf{I}_k(f) = \mathbf{I}_k(A)$ is the probability that the kth bit is pivotal for A (i.e. $\mathbf{I}_k(f) = \mathbb{P}\big[f(\omega) \neq f(\sigma_k \cdot \omega)\big]$).

Remark 2.4. If $f : \Omega_n \to \mathbb{R}$ is a **monotone** function (i.e. $f(\omega_1) \leq f(\omega_2)$ when $\omega_1 \leq \omega_2$), then notice that

$$\hat{f}(\{k\}) := \mathbb{E}\big[f\chi_{\{k\}}\big] = \mathbb{E}\big[f(x_1, \ldots, x_n)x_k\big] = \frac{1}{2}\,\mathbf{I}_k(f)\,,$$

by monotonicity of f. This gives a first hint that influences are closely related to the Fourier expansion of f.

We define the **Total influence** of a (real-valued) function f to be $\mathbf{I}(f) := \sum \mathbf{I}_k(f)$. This notion is very relevant in the study of sharp thresholds (see [Rus82, FK96]). Indeed if f is a monotone function, then by Margulis-Russo's formula (see for example [Gri08])

$$\frac{d}{dt}\Big|_{1/2} \mathbb{P}_p(f) = \mathbf{I}(f) = \sum_k \mathbf{I}_k(f)\,.$$

This formula easily extends to all $p \in [0,1]$. In particular, for a monotone event A, a "large" total influence implies a "sharp" threshold for $p \mapsto \mathbb{P}_p(A)$.

As it has been recognized for quite some time already (since Ben Or/Linial), the set of all influences $\mathbf{I}_k(f)$, $k \in [n]$ carries important information about the function f. Let us then call $\mathbf{Inf}(f) := (\mathbf{I}_k(f))_{k\in[n]}$ the **Influence Vector** of f. We have already seen that the L^1 norm of the influence vector encodes properties of the type "Sharp Threshold" for f since by definition $\mathbf{I}(f) = \|\mathbf{Inf}(f)\|_1$. The L^2 norm of this influence vector will turn out to be a key quantity in the study of noise sensitivity of Boolean functions $f : \Omega_n \to \{0,1\}$. We thus define (following the notations of [BKS99])

$$\mathbf{H}(f) := \sum_k \mathbf{I}_k(f)^2 = \|\mathbf{Inf}(f)\|_2^2\,.$$

For Boolean functions f (i.e. with values in $\{0,1\}$), these notions ($\mathbf{I}(f)$ and $\mathbf{H}(f)$) are intimately related with the above Fourier expansion of f. Indeed we will see in the next section that if $f : \Omega_n \to \{0,1\}$, then

$$\mathbf{I}(f) = 4\sum_S |S|\hat{f}(S)^2\,.$$

If one assumes furthermore that f is **monotone**, then from remark 2.4, one has

$$\frac{1}{4}\,\mathbf{H}(f) = \sum_k \hat{f}(\{k\})^2 = \sum_{|S|=1} \hat{f}(S)^2 = E_f(1), \tag{2.3}$$

which corresponds to the "weight" of the level-one Fourier coefficients (this property also holds for real-valued functions, but we will use the quantity $\mathbf{H}(f)$ only in the Boolean case).

We will conclude by the following general *philosophy* that we will encounter throughout the rest of the survey (especially in section 3): if a function $f : \{0,1\}^n \to \mathbb{R}$ is such that each of its variables has a "very small" influence (i.e. $\ll \frac{1}{\sqrt{n}}$), then f should have a behavior very different from a "Gaussian" one. We will see an illustration

of this *rule* in the context of anomalous fluctuations (lemma 3.9). In the Boolean case, these functions (such that all their variables have very small influence) will be noise sensitive (theorem 3.3), which is not characteristic of Gaussian nor White noise behavior.

2.3 Fast review on Percolation

We will only briefly recall what the model is, as well as some of its properties that will be used throughout the text. For a complete account on percolation see [Gri99] and more specifically in our context the lecture notes [Wer07].

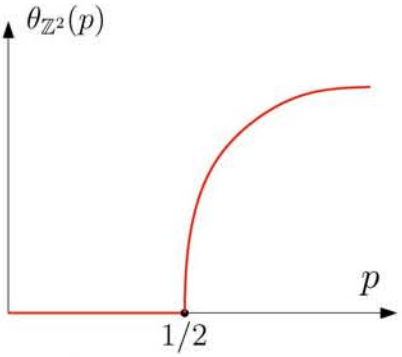
We will be concerned mainly in two-dimensional percolation and we will focus on two lattices: $\mathbb{Z}^2$ and the triangular lattice $\mathbb{T}$. All the results stated for $\mathbb{Z}^2$ in this text are also valid for percolations on "reasonable" 2-d translation invariant graphs for which RSW is known to hold.

Let us describe the model on $\mathbb{Z}^2$. Let $\mathbb{E}^2$ denote the set of edges of the graph $\mathbb{Z}^2$. For any $p \in [0,1]$ we define a random subgraph of $\mathbb{Z}^2$ as follows: independently for each edge $e \in \mathbb{E}^2$, we keep this edge with probability p and remove it with probability $1-p$. Equivalently, this corresponds to defining a random configuration $\omega \in \{-1,1\}^{\mathbb{E}^2}$ where, independently for each edge $e \in \mathbb{E}^2$, we declare the edge to be open ($\omega(e)=1$) with probability p or closed ($\omega(e)=-1$) with probability $1-p$. The law of the so-defined random subgraph (or configuration) is denoted by $\mathbb{P}_p$. In Percolation theory, one is interested in large scale connectivity properties of the random configuration ω.

In particular as one raises the level p, above a certain critical parameter $p_c(\mathbb{Z}^2)$, an infinite cluster (almost surely) appears. This corresponds to the well-known *phase transition* of percolation. By a famous theorem of Kesten this transition takes place at $p_c(\mathbb{Z}^2) = \frac{1}{2}$.

Percolation is defined similarly on the triangular grid $\mathbb{T}$, except that on this lattice we will rather consider **site**-percolation (i.e. here we keep each **site** with probability p). Also for this model, the transition happens at the critical point $p_c(\mathbb{T}) = \frac{1}{2}$.

The phase transition can be measured with the *density function* $\theta_{\mathbb{Z}^2}(p) := \mathbb{P}_p(0 \overset{\omega}{\longleftrightarrow} \infty)$ which encodes important properties of the large scale connectivities of the random configuration ω: it corresponds to the density averaged over the space $\mathbb{Z}^2$ of the (almost surely unique) infinite cluster. The shape of the function $\theta_{\mathbb{Z}^2}$ is pictured on the right (notice the infinite derivative at p_c).



Over the last decade, the understanding of the critical regime has undergone remarkable progress and Oded himself obviously had an enormous impact on these developments. The main ingredients of this productive period were the introduction of the SLE processes by Oded (see the survey on SLEs by Steffen Rohde in the

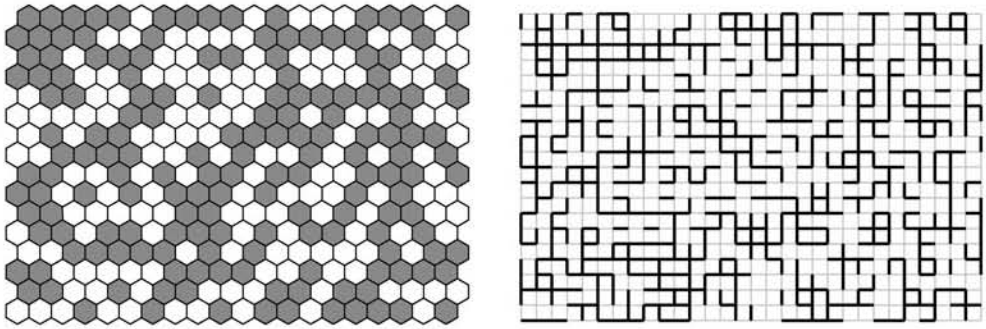
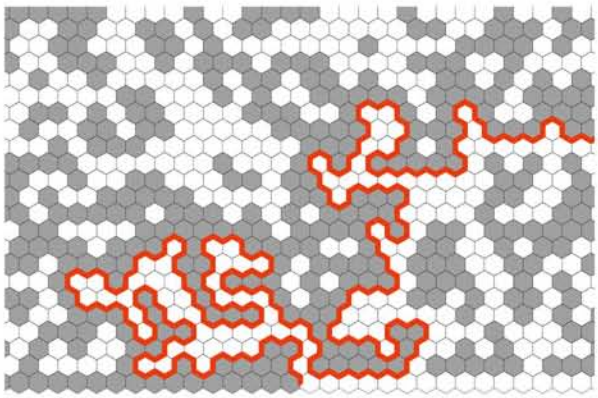


Figure 2.2: Pictures by Oded representing two critical percolation configurations respectively on $\mathbb{T}$ and on $\mathbb{Z}^2$. The sites of the triangular grid are represented by hexagons.

present volume) and the proof of conformal invariance on $\mathbb{T}$ by Stanislav Smirnov ([Smi01]).

At this point one cannot resist to show another famous (and everywhere used) picture by Oded representing an exploration path on the triangular lattice; this red curve which turns right on black hexagons and left on the white ones, asymptotically converges towards SLE_6 (as the mesh size goes to 0).



The proof of conformal invariance combined with the detailed information given by the SLE_6 process enabled one to obtain very precise information on the critical and *near-critical* behavior of $\mathbb{T}$-percolation. For instance it is known that on the triangular lattice, the density function $\theta_{\mathbb{T}}(p)$ has the following behavior near $p_c = 1/2$

$$\theta(p) = (p - 1/2)^{5/36+o(1)},$$

when $p \to 1/2+$ (see [Wer07]).

In the rest of the text, we will often rely on two types of percolation events: namely the *one-arm* and *four-arm* events. They are defined as follows: for any radius $R > 1$, let A_R^1 be the event that the site 0 is connected to distance R by some open path. Also, let A_R^4 be the event that there are four "arms" of alternating color from the site 0 (which can be of either color) to distance R (i.e. there are four connected paths, two open, two closed from 0 to radius R and the closed paths lie between the open paths). See figure 2.3 for a realization of each event.

It was proved in [LSW02] that the probability of the one-arm event decays like

$$\mathbb{P}\big[A_R^1\big] := \alpha_1(R) = R^{-\frac{5}{48}+o(1)}.$$

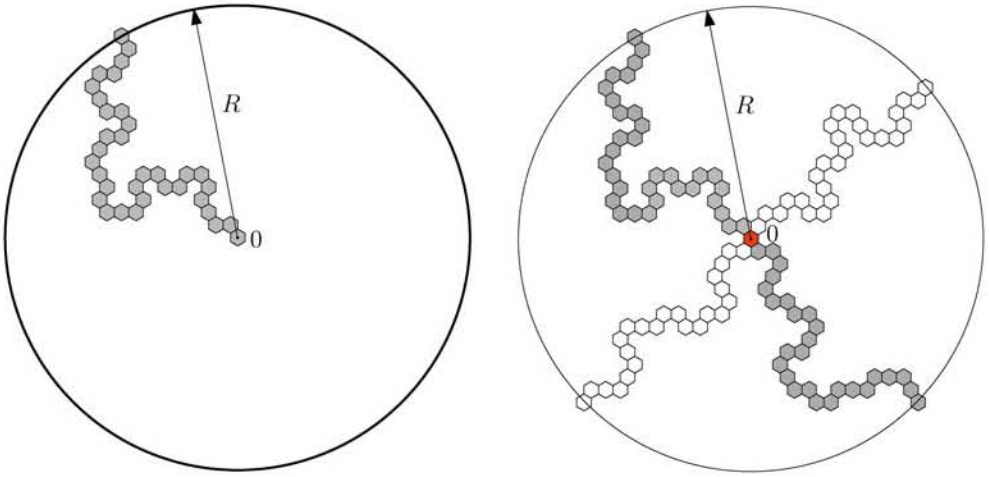


Figure 2.3: A realization of the *one-arm* event is pictured on the left; the *four-arm* event is pictured on the right.

For the four-arms event, it was proved by Smirnov and Werner in [SW01] that its probability decays like

$$\mathbb{P}\big[A_R^4\big] := \alpha_4(R) = R^{-\frac{5}{4}+o(1)}\,.$$

The three exponents we encountered concerning $\theta_{\mathbb{T}}$, α_1 and α_4 (i.e. $\frac{5}{36}$, $\frac{5}{48}$ and $\frac{5}{4}$) are known as *critical exponents*.

The *four-arm* event will be of particular importance throughout the rest of this survey. Indeed suppose the four arms event holds at some site $x \in \mathbb{T}$ up to some large distance R. This means that the site x carries important information about the large scale connectivities within the euclidean ball $B(x,R)$. Changing the status of x will drastically change the "picture" in $B(x,R)$. We call such a point a **pivotal point** up to distance R.

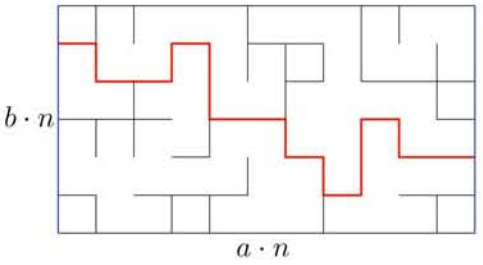
Finally it is often convenient to "divide" these arm-events into different scales. For this purpose, we introduce $\alpha_4(r,R)$ (with $r \le R$) to be the probability that the four-arm event is realized from radius r to radius R ($\alpha_1(r,R)$ is defined similarly for the one-arm event). By independence on disjoint sets, it is clear that if $r_1 \le r_2 \le r_3$ then one has $\alpha_4(r_1,r_3) \le \alpha_4(r_1,r_2)\,\alpha_4(r_2,r_3)$. A very-useful property known as **quasi-multiplicativity** claims that up to constants, these two expressions are the same (this makes the division into several scales practical). This property can be stated as follows.

Proposition 2.5 (**quasi-multiplicativity**, [Kes87]). *For any $r_1 \le r_2 \le r_3$, one has (both for $\mathbb{Z}^2$ and $\mathbb{T}$ percolations)*

$$\alpha_4(r_1,r_3) \asymp \alpha_4(r_1,r_2)\,\alpha_4(r_2,r_3)\,,$$

where the constant involved in $\asymp$ are uniform constants.

See [Wer07, Nol09] for more details. Note also that the same property holds for the one-arm event (but is much easier to prove: it is an easy consequence of the RSW theorem which is stated in figure 2.4) .



The picture on the left represents a configuration for which the left-right crossing of $[a \cdot n, b \cdot n]$ is realized; in this case we define $f_n(\omega) := 1$; otherwise we define $f_n(\omega) := 0$. There is an extremely useful result known as **RSW theorem** which states that for any $a, b > 0$, asymptotically the probability of crossing the rectangle $\mathbb{P}\big[f_n = 1\big]$ remains bounded away from 0 and 1.

Figure 2.4: Left to Right crossing event in a rectangle and RSW Theorem.

To end this preliminary section, let us sketch what will be one of the main goals throughout this survey: if f_n denotes the Left-Right crossing event of a large rectangle $[a \cdot n, b \cdot n]$ (see figure 2.4), then one can consider these observables as Boolean functions. As such they admit a Fourier expansion. Understanding the sensitivity of percolation will correspond to understanding the Energy Spectrum of such observables. We will see by the end of this survey that, as $n \to \infty$, the Energy Spectrum of f_n should roughly look as in figure 2.5.

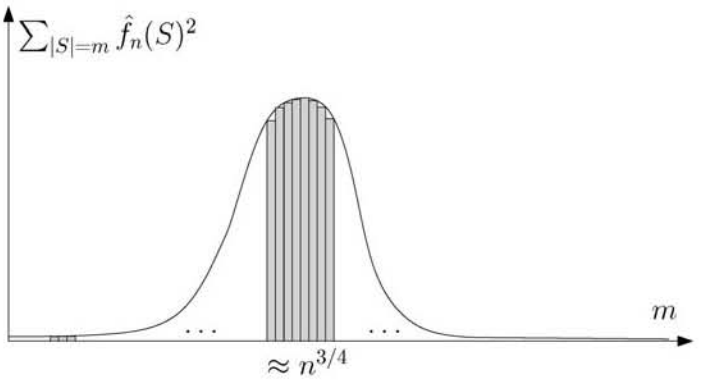


Figure 2.5: Expected shape of the Energy Spectrum of percolation crossing events. We will see in section 5 that most of the Spectral mass of f_n is indeed localized around $n^{3/4}$. Compare this shape of the spectrum with the above spectrum of the Majority function Φ_n.

3 The "hypercontractivity" approach

In this section, we will describe the first noise sensitivity estimates as well as anomalous fluctuations for First Passage Percolation (FPP).

The notion of influence will be essential throughout this section; recall that we defined $\nabla_k f = f - f \circ \sigma_k$ and $\mathbf{I}_k(f) = \|\nabla_k f\|_1$. It is fruitful to consider $\nabla_k f$ as a function from Ω_n into $\mathbb{R}$. Indeed, its Fourier decomposition is directly related to the Fourier decomposition of f itself in the following way: it is straightforward to check that for any $S \subset [n]$

$$\widehat{\nabla_k f}(S) = \begin{cases} 2\hat{f}(S) & \text{if } k \in S \\ 0 & \text{else} \end{cases}$$

If one assumes furthermore that f is Boolean (i.e. $f : \Omega_n \to \{0,1\}$), then our discrete derivatives $\nabla_k f$ take their values in $\{-1,0,1\}$; this implies in particular that for any $k \in [n]$

$$\mathbf{I}_k(f) = \|\nabla_k f\|_1 = \|\nabla_k f\|_2^2 .$$

Using Parseval, this enables us to write the Total influence of f in terms of its Fourier coefficients as follows:

$$\begin{aligned} \mathbf{I}(f) = \sum_k \mathbf{I}_k(f) &= \sum_k \|\nabla_k f\|_2^2 \\ &= \sum_k \|\widehat{\nabla_k f}\|_2^2 \\ &= 4 \sum_S |S| \hat{f}(S)^2, \end{aligned} \tag{3.1}$$

since each "frequency" $S \subset [n]$ appears $|S|$ times. Before the appearance of the notion of "noise sensitivity" there have been several works whose focus were to get lower bounds on influences. For example (see [FK96]), if one wants to provide general criteria for sharp thresholds, one needs to obtain lower bounds on the total influence $\mathbf{I}(f)$. The strategy to do this in great generality was developed in [KKL88], where they proved that for any Balanced Boolean function f (balanced meaning here $\mathbb{P}\big[f = 1\big] = 1/2$), there always exists a variable whose influence is greater than $O(\frac{\log n}{n})$.

Before treating in details the case of noise sensitivity, let us describe in an informal way what was the ingenious approach from [KKL88] to obtain lower bounds on influences. Let us consider some arbitrary balanced Boolean function f. We want to show that at least one of its variables has large influence $\geq c\frac{\log n}{n}$. Suppose all its influences $\mathbf{I}_k(f)$ are 'small' (this would need to be made quantitative), this means that all the functions $\nabla_k f$ have small L^1 norm. Now if f is Boolean (into $\{0,1\}$), then as we have noticed above, $\nabla_k f$ is almost Boolean (its values are in $\{-1,0,1\}$); hence $\|\nabla_k f\|_1$ being small implies that $\nabla_k f$ has a small support. Using the intuition coming from Weyl-Heisenberg uncertainty, $\widehat{\nabla_k f}$ should then be quite spread, in particular, most of its spectral mass should be concentrated on high frequencies.

This intuition (which is still vague at this point) somehow says that having small influences pushes the spectrum of $\nabla_k f$ towards high frequencies. Now summing up as we did in (3.1), but only restricting ourselves to fequencies S of size smaller than some large (well-chosen) $1 \ll M \ll n$, one obtains

$$\begin{aligned} \sum_{0<|S|<M} \hat{f}(S)^2 &\leq 4 \sum_{0<|S|<M} |S| \hat{f}(S)^2 \\ &= \sum_k \sum_{0<|S|<M} \widehat{\nabla_k f}(S)^2 \\ &\text{“} \ll \text{”} \sum_k \|\widehat{\nabla_k f}\|_2^2 \\ &= \mathbf{I}(f)\,, \end{aligned} \tag{3.2}$$

where in the third line, we used the informal statement that $\widehat{\nabla_k f}$ should be supported on high frequencies if f has small influences. Now recall that we assumed f to be balanced, hence

$$\sum_{|S|>0} \hat{f}(S)^2 = 1/4\,.$$

Therefore, in the above equation (3.2), if we are in the case where a positive fraction of the Fourier mass of f is concentrated below M, then (3.2) says that $\mathbf{I}(f)$ is much larger than one. In particular, at least one of the influences has to be 'large'. If, on the other hand, we are in the case where most of the Spectral mass of f is supported on frequencies of size higher than M, then we also obtain that $\mathbf{I}(f)$ is large by the previous formula:

$$\mathbf{I}(f) = 4 \sum_S |S| \hat{f}(S)^2\,.$$

In [KKL88], this intuition is converted into a proof. The main difficulty here is to formalize or rather to implement the above argument, i.e. to obtain spectral information on functions with values in $\{-1, 0, 1\}$ knowing that they have small support. This is done ([KKL88]) using techniques brought from harmonic analysis, namely *Hypercontractivity.*

3.1 About Hypercontractivity

First, let us state what it corresponds to. Let $(K_t)_{t\geq 0}$ be the heat kernel on $\mathbb{R}^n$. Hypercontractivity is a statement which quantifies how functions are regularized under the heat flow. The statement, which goes back to Nelson/Gross can be simply stated as follows :

Lemma 3.1 (Hypercontractivity)**.** *If $1 < q < 2$, there is some $t = t(q) > 0$ (which does not depend on the dimension n) such that for any $f \in L^q(\mathbb{R}^n)$,*

$$\|K_t * f\|_2 \leq \|f\|_q\,.$$

The dependence $t = t(q)$ is explicit but will not concern us in the Gaussian case. Hypercontractivity is thus a regularization statement: if one starts with some initial "rough" L^q function f outside of L^2 and waits long enough ($t(q)$) under the heat flow, we end up being in L^2 with a good control on its L^2 norm.

This concept has an interesting history as is nicely explained in O'Donnell lectures notes (see [O'D]). It was originally invented by Nelson in [Nel66] where he needed regularization estimates on Free Fields (which are the building blocks of quantum field theory) in order to apply these in "constructive field theories". It was then generalized by Gross in his elaboration of Logarithmic Sobolev Inequalities ([Gro75]), which are an important tool in analysis. Hypercontractivity is intimately related to these Log-Sobolev Inequalities (they are somewhat equivalent concepts) and thus has many applications in the theory of Semi-Groups, mixing of Markov chains and so on.

We now state the result in the case which concerns us, the hypercube. For any $\rho \in [0,1]$, let T_ρ be the following "Noise Operator" on the functions of the hypercube: recall from the Preliminary section that if $\omega \in \Omega_n$, we denote by ω^ϵ an ϵ-noised configuration of ω. For any $f : \Omega_n \to \mathbb{R}$, we define $T_\rho f : \omega \mapsto \mathbb{E}\big[f(\omega^{1-\rho}) \;\big|\; \omega\big]$. This Noise Operator acts in a very simple way on the Fourier coefficients :

$$T_\rho : f = \sum_S \hat{f}(S)\, \chi_S \mapsto \sum_S \rho^{|S|} \hat{f}(S)\, \chi_S \,.$$

We have the following analog of lemma 3.1

Lemma 3.2 (Bonami-Gross-Beckner)**.** *For any $f : \Omega_n \to \mathbb{R}$,*

$$\|T_\rho f\|_2 \leq \|f\|_{1+\rho^2} \,.$$

The analogy with the classical lemma 3.1 is clear: the Heat flow is replaced here by the random Walk on the hypercube.

Before applying Hypercontractivity to Noise Sensitivity, let us sketch how this functional inequality helps to implement the above idea from [KKL88]. If a Boolean function f has small influences, its discrete derivatives $\nabla_k f$ have small support. Now these functions have values in $\{-1, 0, 1\}$, thus for any $1 < q < 2$ we have that $\|\nabla_k f\|_q = (\|\nabla_k f\|_2)^{2/q} \ll \|\nabla_k f\|_2$ (because of the small support of $\nabla_k f$). Now applying hypercontractivity (with $q = 1+\rho^2$), we obtain that $\|T_\rho \nabla_k f\|_2 \ll \|\nabla_k f\|_2$. Written on the Fourier side this means that

$$\sum_S \rho^{2|S|} \widehat{\nabla_k f}(S)^2 \ll \sum_S \widehat{\nabla_k f}(S)^2 \,,$$

and this happens only if most of the Spectral mass of $\nabla_k f$ is supported on high frequencies. It remains to make the above heuristics precise in the case which interests us here.

3.2 Applications to Noise sensitivity

Let us now see Hypercontractivity in action. As in the introduction we are interested in the noise sensitivity of a sequence of Boolean functions $f_n : \Omega_{m_n} \to \{0,1\}$. A deep theorem from [BKS99] can be stated as follows

Theorem 3.3 ([BKS99]). *Let $(f_n)_n$ be a sequence of Boolean functions. If $\mathbf{H}(f_n) \to 0$, when $n \to \infty$, then $(f_n)_n$ is* **Noise Sensitive** *i.e. for any $\epsilon > 0$, the correlation $\mathbb{E}\big[f_n(\omega) f_n(\omega^\epsilon)\big] - \mathbb{E}\big[f\big]^2$ converges to 0.*

The theorem is true independently of the speed of convergence of $\mathbf{H}(f_n)$. Nevertheless, if one assumes that there is some exponent $\delta > 0$, such that $\mathbf{H}(f_n) \leq (m_n)^{-\delta}$, then the proof is quite simple as was pointed out to us by Jeff Steif, and furthermore one obtains some "logarithmic bounds" on the sensitivity of f_n. We will restrict ourselves to this stronger assumption since it will be sufficient for our application to Percolation.

Remark 3.4. If the Boolean functions $(f_n)_n$ are assumed to be monotone, it is interesting to note that as we observed in (2.3), $\mathbf{H}(f_n) = \sum_{|S|=1} \hat{f}_n(S)^2$. So $\mathbf{H}(f_n)$ corresponds here to the level-1 Fourier weights. Thus in the monotone case, theorem 3.3 says that if asymptotically, there is no weight on the level one coefficients, then there is no weight on any finite level Fourier coefficient ! (Of course the Boolean hypothesis on f_n is also essential here). In particular, in the monotone case, the condition $\mathbf{H}(f_n) \to 0$ is equivalent to Noise Sensitivity.

Proof of Theorem 3.3 (under the stronger assumption $\mathbf{H}(f_n) \leq (m_n)^{-\delta}$ for some $\delta > 0$).

The spirit of the proof is similar to the one carried out in [KKL88], but the target is different here. Indeed in [KKL88], the goal was to obtain good Lower bounds on the Total influence in great generality; for example the (easy) sharp threshold encountered by the Majority function around $p = 1/2$ fits into their framework. Majority is certainly not a sensitive function, so theorem 3.3 requires more assumption than [KKL88] results. Nevertheless the strategy will be similar: we still use the "spectral decomposition" of f with respect to its "partial derivatives" $\nabla_k f$

$$\mathbf{I}(f) = \sum_k \|\nabla_k f\|_2^2 = \sum_k \|\widehat{\nabla_k f}\|_2^2 = 4 \sum_S |S| \hat{f}(S)^2 \,.$$

But now we want to use the fact that $\mathbf{H}(f) = \sum_k \|\nabla_k f\|_2^4$ is very small. If f is Boolean, this implies that the (almost) Boolean $\nabla_k f$ have very small support (and this can be made quantitative). Now, again we expect $\widehat{\nabla_k f}$ to be spread, but this time, we need more: not only $\nabla_k f$ have high frequencies, but in some sense "all their frequencies" are high leaving no mass (after summing up over the m_n variables) to the finite level Fourier coefficients. This is implemented using Hypercontractivity, following similar lines as in [KKL88].

Let us then consider a sequence of Boolean functions $f_n : \Omega_{m_n} \to \{0,1\}$ satisfying $\mathbf{H}(f_n) \leq (m_n)^{-\delta}$ for some exponent $\delta > 0$. We want to show that there is some

constant $M = M(\delta)$, such that

$$\sum_{0<|S|<M\log(m_n)} \hat{f}_n(S)^2 \to 0\,,$$

which gives a quantitative (logarithmic) noise sensitivity statement.

$$\begin{aligned}
\sum_{0<|S|<M\log(m_n)} \hat{f}_n(S)^2 &\leq 4 \sum_{0<|S|<M\log(m_n)} |S|\hat{f}_n(S)^2 = \sum_k \sum_{0<|S|<M\log(m_n)} \widehat{\nabla_k f}(S)^2 \\
&\leq \sum_k (\frac{1}{\rho^2})^{M\log(m_n)} \|T_\rho \nabla_k f\|_2^2 \\
&\leq \sum_k (\frac{1}{\rho^2})^{M\log(m_n)} \|\nabla_k f\|_{1+\rho^2}^2 \text{ By hypercontractivity.}
\end{aligned}$$

Now, since f is Boolean, one has $\|\nabla_k f\|_{1+\rho^2} = \|\nabla_k f\|_2^{2/(1+\rho^2)}$, hence

$$\begin{aligned}
\sum_{0<|S|<M\log(m_n)} \hat{f}_n(S)^2 &\leq \rho^{-2M\log(m_n)} \sum_k \|\nabla_k f\|_2^{4/(1+\rho^2)} = \rho^{-2M\log(m_n)} \sum_k \mathbf{I}_k(f)^{2/(1+\rho^2)} \\
&\leq \rho^{-2M\log(m_n)} (m_n)^{\rho^2/(1+\rho^2)} \Big(\sum_k \mathbf{I}_k(f)^2\Big)^{\frac{1}{1+\rho^2}} \text{ (By Hölder)} \\
&= \rho^{-2M\log(m_n)} (m_n)^{\rho^2/(1+\rho^2)}\, \mathbf{H}(f_n)^{\frac{1}{1+\rho^2}} \\
&\leq \rho^{-2M\log(m_n)} (m_n)^{\frac{\rho^2-\delta}{1+\rho^2}}\,.
\end{aligned}$$

Now by choosing $\rho \in (0,1)$ close enough to 0, and then by choosing $M = M(\delta)$ small enough, we obtain the desired logarithmic Noise Sensitivity.

Application to Percolation.
We want to apply the above result to the case of percolation crossings. Let D be some smooth domain of the plane (not necessarily simply connected) and let ∂_1, ∂_2 be two smooth arcs on the boundary ∂D. For all $n \geq 1$, let $D_n \subset \frac{1}{n}\mathbb{Z}^2$ be a domain approximating D, and call f_n, the indicator function of a left to right crossing in D_n from ∂_1^n to ∂_2^n (see figure 4.1 in section 4). Noise Sensitivity of percolation means that the sequence of events $(f_n)_{n\geq 1}$ is Noise Sensitive. Using theorem 3.3, it would be enough (and necessary) to show that $\mathbf{H}(f_n) \to 0$. But if we want a self-contained proof here, we would prefer to have at our disposal the stronger claim $\mathbf{H}(f_n) \leq n^{-\delta}$ for some $\delta > 0$ (here $m_n \asymp n^2$).

There are several ways to see why this stronger claim holds. The most natural one is to get good estimates on the probability for an edge e to be Pivotal. Indeed, recall that in the monotone case, $\mathbf{H}(f_n) := \sum_{\text{edges } e\in D_n} \mathbb{P}\big[e \text{ is Pivotal}\big]^2$ (see subsection 2.2). This probability, without considering boundary effects, is believed to be of order $n^{-5/4}$, which indeed makes $\mathbf{H}(f_n) \approx n^2 \cdot (n^{-5/4})^2 \approx n^{-1/2}$ decrease to 0 polynomially fast. This behavior is now known in the case of the triangular grid thanks to Schramm's SLEs and Smirnov's proof of conformal invariance (see [SW01] where the relevant critical exponent is computed).

At the time of [BKS99], of course, critical exponents were not available (and anyway, they still remain unavailable today on $\mathbb{Z}^2$) but Kesten had estimates which implied that for some $\epsilon > 0$, $\mathbb{P}\big[e \text{ is Pivotal}\big] \leq n^{-1-\epsilon}$ (which is enough to obtain $\mathbf{H}(f_n) \leq n^{-\delta}$)

Furthermore, an ingenious alternative way to obtain the polynomial convergence of $\mathbf{H}(f_n)$ was developed in [BKS99] which did not need Kesten's results. This alternative way, on which we will say a few words in section 4, in some sense prefigured the Randomized Algorithm approach that we will describe in the next section.

Some words on the general Theorem 3.3 and its proof.
The proof of the general result is a bit more involved than the one we outlined here. The main lemma is as follows :

Lemma 3.5. *There exist absolute constants C_k for all $k \geq 1$, such that for any monotone Boolean function f one has*

$$\sum_{|S|=k} \hat{f}(S)^2 \leq C_k \mathbf{H}(f) \left(-\log \mathbf{H}(f)\right)^{k-1}.$$

This lemma "mimics" a result from Talagrand's [Tal96]. Indeed proposition 2.2 in [Tal96] can be translated as follows: for any monotone Boolean function f, its level-2 Fourier weight (i.e. $\sum_{|S|=2} \hat{f}(S)^2$) is bounded by $O(1)\mathbf{H}(f)\log(1/\mathbf{H}(f))$. It obviously implies theorem 3.3 in the monotone case, the general case being deduced from it by a monotonization procedure. Hypercontractivity is used in the proof of this lemma.

3.3 Anomalous fluctuations, or Chaos

In this section, we will outline how Hypercontractivity was used in [BKS03] in order to prove that shape fluctuations in the model of First-Passage-Percolation are subgaussian.

3.3.1 The model of First Passage Percolation (FPP)

Let us start with the model and then state the theorem proved in [BKS03]. First Passage Percolation can be seen as a model of a **random metric** on $\mathbb{Z}^d$; it is defined simply as follows: take two different "lengths" a and b satisfying $0 < a < b$ and independently for each edge $e \in \mathbb{E}^d$, fix the length of e to be a with probability $1/2$, b else. In greater generality, the lengths of the edges are i.i.d. non-negative random variables, but here, following [BKS03], we will restrict ourselves to the above uniform distribution on $\{a, b\}$ to simplify the exposition (see [BR08] for an extension to more general laws). For any $\omega \in \{a, b\}^{\mathbb{E}^d}$, this defines a (random) metric, dist_ω, on $\mathbb{Z}^d$ satisfying for any $x, y \in \mathbb{Z}^d$,

$$\text{dist}_\omega(x, y) := \inf_{\gamma:\text{ path from } x \text{ to } y} l_\omega(\gamma),$$

where $l_\omega(\gamma)$ is the length of γ.

Using Sub-additivity, it is known that the renormalized ball $\frac{1}{n}B_\omega(0, n)$ converges towards a deterministic shape (which can be in certain cases computed explicitly).

3.3.2 Fluctuations around the limiting Shape

The fluctuations around the asymptotic limiting shape have received tremendous interest over the last 15 years. In the two-dimensional case, using very beautiful combinatorial bijections with Random Matrices, certain cases of **directed** Last Passage Percolation (where the law on the edges is taken to be geometric or exponential) have been understood very deeply. For example, it is known ([Joh00]) that the fluctuations of the Ball of radius n (i.e. the points whose Last Passage Time are below n) around n times its asymptotic deterministic shape, are of order $n^{1/3}$, and the law of these fluctuations properly renormalized follow the Tracy-Widom distribution (as do the fluctuations of the largest eigenvalue of GUE ensembles).

"Universality" is believed to hold for these models in the sense that the behavior of the fluctuations around the deterministic shape should not depend on the "microscopic" particularities of the model (for example the law on the edges). The shape itself does depend of course. In particular in the above model of (non-directed) First Passage Percolation in dimension $d = 2$, fluctuations are widely believed to be also of order $n^{1/3}$ following as well the Tracy-Widom Law. Still, the current state of understanding of this model is far from this conjecture.

Kesten first proved that the fluctuations of the Ball of radius n were at most $\sqrt{n}$ (which did not exclude yet Gaussian Behavior). Benjamini, Kalai and Schramm then strengthened this result by showing that the fluctuations were Sub-Gaussian. This does not yet reach the conjectured $n^{1/3}$-fluctuations, but their approach has the great advantage to be very general; in particular their result holds in any dimension $d \geq 2$.

Let us now state their main theorem concerning the fluctuations of the metric dist_ω.

Theorem 3.6 ([BKS03])**.** *For all a, b, d, there exists an absolute constant $C = C(a, b, d)$ such that in $\mathbb{Z}^d$*

$$\mathrm{var}(\mathrm{dist}_\omega(0, v)) \leq C \frac{|v|}{\log |v|}$$

for any point $v \in \mathbb{Z}^2$, $|v| \geq 2$.

3.3.3 Link with "Noise Sensitivity"

This result about Sub-Gaussian fluctuations might seem at first disjoint from our initial study of Noise Sensitivity, but they turn out to be intimately related. First of all the methods to understand one or the other, as we will see, follow very similar lines. But also, as is very nicely explained in [Cha08], the phenomenon of "Anomalous fluctuations" is in some sense equivalent to a certain "Noise sensitivity" of the geodesics of First-Passage-Percolation. More precisely, the Variance of the first passage time is of order $\mathbb{E}\big[|\gamma(\omega(0)) \cap \gamma(\omega(\tau))|\big]$, where $\tau \sim \mathcal{E}(1)$ is an exponential variable. Thus we see that if the metric, or rather the geodesic, is highly perturbed when the configuration is noised; then the distances happen to be very concentrated. Chatterjee calls this phenomenon **Chaos**. Of course, our short description here was

informal since in our present setup there might be many different geodesics between two fixed points. The above link between concentration and sensitivity discovered by Chatterjee works very nicely in the context of Maxima of Gaussian processes (which in that case arise a.s at a single point, or a single "geodesic" in a geometrical context); see [Cha08] for more details.

3.3.4 The simpler case of the Torus

Following the approach of [BKS03], we will first consider the case of the Torus. The reason for this is that it is a much simpler case: indeed in the torus, for the least-passage time that we will consider, any edge will have up to constant the same influence, while in the case of $\mathrm{dist}_\omega(0, v)$, edges near the endpoints 0 or v have a high influence on the outcome (in some sense there is more symmetry and invariance to play with in the case of the Torus).

Let $\mathbb{T}_m^d$ be the d-dimensional Torus $(\mathbb{Z}/m\mathbb{Z})^d$. As in the above (lattice) model, independently for each edge of $\mathbb{T}_m^d$, we choose its length to be either a or b. We are interested here in the smallest (random) length among closed paths γ "turning" around the torus along the first coordinate $\mathbb{Z}/m\mathbb{Z}$ (i.e. these paths γ, once projected onto the first cycle, have winding number one). In [BKS03], this is called the shortest *circumference*. For any configuration $\omega \in \{a, b\}^{E(\mathbb{T}_m^d)}$, call $\mathrm{Circ}_m(\omega)$ this shortest circumference.

Theorem 3.7. *There is a constant $C = C(a, b)$ (which does not depend on the dimension d), such that*

$$\mathrm{var}(\mathrm{Circ}_m(\omega)) \leq C \frac{m}{\log m}$$

Remark 3.8. A similar analysis as the one carried out below works in greater generality: if $G = (V, E)$ is some finite connected graph endowed with a random metric d_ω with $\omega \in \{a, b\}^{\otimes E}$, then one can obtain bounds on the fluctuation of the random diameter $D = D_\omega$ of (G, d_ω). See [BKS03, Theorem 2] for a precise statement in this more general context.

Sketch of proof of Theorem 3.7.
In order to highlight the similarities with the above case of Noise Sensitivity of percolation, we will not follow exactly [BKS03]; it will be more "hands-on" with the disadvantage of being less optimal (with constants and so on).

As before, for any edge e, let us consider the gradient along the edge e: $\nabla_e \mathrm{Circ}_m$; these gradient functions have values in $\{-(b-a), 0, b-a\}$, since changing the length of e can only have this effect on the circumference. By dividing our distances by the constant factor $b - a$, we can even assume our gradient functions to have values in $\{-1, 0, 1\}$. Doing so, we end up being in the same setup as in our previous study; influences are defined in the same way and so on. We sill see that our gradient functions (which are 'almost Boolean') have small support, and Hypercontractivity will imply the desired bounds.

Let us work in the general case of a function $f : \{-1, 1\}^n \to \mathbb{R}$, such that for any variable k, $\nabla_k f \in \{-1, 0, 1\}$. We are interested in $\mathrm{var}(f)$ (and we want to show that

if "influences are small" then $\mathrm{var}(f)$ is small). It is easy to check that the variance can be written

$$\mathrm{var}(f) = \frac{1}{4}\sum_k \sum_{\emptyset \neq S \subset [n]} \frac{1}{|S|}\widehat{\nabla_k f}(S)^2 \,.$$

We see on this expression, that if variables have very small influence, then as previously, the almost Boolean $\nabla_k f$ will be of High frequency. Heuristically, this should then imply that

$$\begin{aligned} \mathrm{var}(f) &\ll \sum_k \sum_{S \neq \emptyset} \widehat{\nabla_k f}(S)^2 \\ &= \sum_k \mathbf{I}_k(f) \,. \end{aligned}$$

We prove the following lemma on the link between the fluctuations of a real-valued function f on Ω_n and its influence vector.

Lemma 3.9. *Let $f : \Omega_n \to \mathbb{R}$ be a (real-valued) function such that each of its discrete derivative $\nabla_k f$, $k \in [n]$ have their values in $\{-1, 0, 1\}$. If we assume that the influences of f are small in the following sense: there exists some $\alpha > 0$ such that for any $k \in \{1, \dots, n\}$, $\mathbf{I}_k(f) \leq n^{-\alpha}$, then there is some constant $C = C(\alpha)$, such that*

$$\mathrm{var}(f) \leq \frac{C}{\log n} \sum_k \mathbf{I}_k(f) \,.$$

Before proving the lemma, let us see that in our special case of First Passage Percolation, the assumption on small influences is indeed verified. Since edges's length are in $\{a, b\}$, the smallest contour $\mathrm{Circ}_m(\omega)$ in $\mathbb{T}_m^d$ around the first coordinate lies somewhere in $[am, bm]$. Hence, if γ is a geodesic (a path in the Torus) satisfying $l(\gamma) = \mathrm{Circ}_m(\omega)$, then γ uses at most $\frac{b}{a}m$ edges. There might be several different geodesics minimizing the circumference. Let us choose randomly one of these in an 'invariant' way and call it $\tilde{\gamma}$. For any edge $e \in E(\mathbb{T}_m^d)$, if by changing the length of e, the circumference increases, then e has to be contained in any geodesic γ, and in particular in $\tilde{\gamma}$. This implies that $\mathbb{P}\big[\nabla_e \mathrm{Circ}_m(\omega) > 0\big] \leq \mathbb{P}\big[e \in \tilde{\gamma}\big]$. By symmetry we obtain that

$$\mathbb{P}\big[\nabla_e \mathrm{Circ}_m(\omega) \neq 0\big] \leq 2\mathbb{P}\big[e \in \tilde{\gamma}\big] \,.$$

As we have seen above, up to a constant factor $b - a$, $\nabla_e \mathrm{Circ}_m \in \{-1, 0, 1\}$; therefore $\mathbf{I}_e(\mathrm{Circ}_m) \leq O(1)\mathbb{P}\big[e \in \tilde{\gamma}\big]$. Now using the symmetries both of the Torus $\mathbb{T}_m^d$ and of our observable Circ_m, if $\tilde{\gamma}$ is chosen in an appropriate invariant way (uniformly among all geodesics would work), then it is clear that all the vertical edges (the edges which, once projected on the first cycle, project on a single vertex) have the same probability to lie in $\tilde{\gamma}$; same for horizontal edges. In particular:

$$\sum_{\text{“vertical edges” } e} \mathbb{P}\big[e \in \tilde{\gamma}\big] \leq \mathbb{E}\big[|\tilde{\gamma}|\big] \leq \frac{b}{a}m \,.$$

Since there are $O(1)m^d$ vertical edges, the influence of these is bounded by $O(1)m^{1-d}$. Same thing for horizontal edges. All together this gives the desired assumption needed in lemma 3.9. Applying this lemma, we indeed obtain that

$$\mathrm{var}(\mathrm{Circ}_m(\omega)) \leq O(1)\frac{m}{\log m}\,,$$

where the constant does not depend on the dimesnion d (since the dimension helps us here).

Proof of the Lemma. As for Noise sensitivity, the proof relies on implementing Hypercontractivity in the right way.

$$\begin{aligned}\mathrm{var}(f) &= \frac{1}{4}\sum_k \sum_{S\neq\emptyset} \frac{1}{|S|}\widehat{\nabla_k f}(S)^2 \\ &\leq \frac{1}{4}\sum_k \sum_{0<|S|<c\log n} \widehat{\nabla_k f}(S)^2 + \frac{O(1)}{\log n}\sum_k \mathbf{I}_k(f)\end{aligned}$$

Hence it is enough to bound the contribution of small frequencies, $0 < |S| < c\log n$, for some constant c which will be chosen later. As previously we have for any $\rho \in (0,1)$ and using Hypercontractivity,

$$\begin{aligned}\sum_k \sum_{0<|S|<c\log n} \widehat{\nabla_k f}(S)^2 &\leq \rho^{-2c\log n}\sum_k \|T_\rho \nabla_k f\|_2^2 \\ &\leq \rho^{-2c\log n}\sum_k \|\nabla_k f\|_{1+\rho^2}^2 \\ &= \rho^{-2c\log n}\sum_k \mathbf{I}_k(f)^{2/(1+\rho^2)} \\ &\leq \rho^{-2c\log n}\big(\sup_k \mathbf{I}_k(f)\big)^{\frac{1-\rho^2}{1+\rho^2}}\sum_k \mathbf{I}_k(f) \\ &\leq \rho^{-2c\log n} n^{-\alpha\frac{1-\rho^2}{1+\rho^2}}\sum_k \mathbf{I}_k(f) \ \text{ by our assumption}\,.\end{aligned} \tag{3.3}$$

Now fixing $\rho \in (0,1)$, and then choosing the constant c depending on ρ and α, the lemma follows (by optimizating on the choice of ρ, one could get better constants). □

3.3.5 Some hints for the proof of Theorem 3.6

The main difficulty here is that the quantity of interest: $f(\omega) := \mathrm{dist}_\omega(0,v)$ is not anymore invariant under a large class of graph automorphisms. This lack of symmetry makes the study of influences more difficult. (For example, as was noticed above, edges near the endpoints 0 or v will have high influence). To gain some more symmetry, the authors in [BKS03] rely on a nice "averaging" procedure. The idea

is as follows: instead of looking at the (random) distance form 0 to v, they first pick a point x randomly in the mesoscopic box $[-|v|^{1/4}, |v|^{1/4}]^d$ around the origin and then consider the distance from this point x towards $v+x$. Let $\tilde{f}$ denote this function $(\mathrm{dist}_\omega(x, v+x))$. $\tilde{f}$ uses extra randomness compared to f, but it is clear that $\mathbb{E}\big[f\big] = \mathbb{E}\big[\tilde{f}\big]$ and it is not hard to see that when $|v|$ is large, $\mathrm{var}(f) \asymp \mathrm{var}(\tilde{f})$. Therefore it is enough to study the fluctuations of the more symmetric $\tilde{f}$ (we already see here that thanks to this avergaing procedure; the endpoints 0 and v do not have anymore a high influence). In some sense, along geodesics, this procedure **"spreads"** the influence on the $|v|^{1/4}$-neighborhood of the geodesics. More precisely, if e is some edge, the influence of this edge is bounded by $2\mathbb{P}\big[e \in x + \gamma\big]$, where γ is chosen among geodesics from 0 to v. Now, as we have seen in the case of the Torus, geodesics are essentially one-dimensional (of length less than $O(1)|v|$); this is still true on the mesoscopic scale: For any box Q of radius $m := |v|^{1/4}$, $|\gamma \cap Q| \leq O(1)m$. Now by considering the mesoscopic box around e, it is like moving a "line" in a box of dimension d; the probability for an edge to be hit by that "line" is of order m^{1-d}. Therefore the influence of any edge e for the "spread" function $\tilde{f}$ is bounded by $O(1)|v|^{(1-d)/4} \leq O(1)|v|^{-1/4}$. This implies the needed assumption in lemma 3.9 and hence concludes the sketch of proof of theorem 3.7. See [BKS03] for a more detailed proof.

Remark 3.10. In this survey, we relied on lemma 3.9, since its proof is very similar to the Noise Sensitivity Proof. In [BKS03], the authors use (and reprove with better constants) an inequality from Talagrand ([Tal94]) which states that there is some universal constant $C > 0$ such that for any $f : \{0,1\}^n \to \mathbb{R}$,

$$\mathrm{var}(f) \leq C \sum_k \frac{\|\nabla_k f\|_2^2}{1 + \log(\|\nabla_k f\|_2 / \|\nabla_k f\|_1)} .$$

Conclusion : [BKS99, BKS03] developed multiple and very interesting techniques. The results of [BKS03] have since been extended to more general laws ([BR08]) but essentially, their control of the variance in $\Omega(n/\log n)$ is to this day still the best. The paper [BKS99] had a profound impact on the field. As we will see, some of the ideas present in [BKS99] already announced some ideas of the next section.

4 The Randomized Algorithm approach

In this part, we will describe the quantitative estimates on Noise Sensitivity obtained in [SS10b]. Their applications to the model of dynamical percolation will be described in the last section of this survey. But before we turn to the remarkable paper [SS10b], where Schramm and Steif introduced deep techniques to control Fourier spectrums of general functions, let us first mention and explain that the idea of using randomized algorithms was already present in essence in [BKS99], where they used an algorithm in order to prove that $\mathbf{H}(f)$ converges quickly (polynomially) towards 0.

4.1 First appearance of randomized algorithm ideas

In [BKS99], as we have seen above, in order to prove that percolation crossings are asymptotically noise sensitive, the authors needed the fact that $\mathbf{H}(f_n) = \sum_k \mathbf{I}_k(f_n)^2 \to 0$ (see theorem 3.3); if furthermore this L^2 quantity converges to zero more quickly than a polynomial of the number of variables: $(m_n)^{-\delta}$ for some $\delta > 0$, then the proof of theorem 3.3 is relatively simple as we outlined above. This fast convergence to zero of $\mathbf{H}(f_n)$ was guaranteed by the work of Kesten (in particular his work [Kes87] on hyperscaling from which follows the fact that the probability for a point to be pivotal until distance m is less than $O(1)m^{-1-\alpha}$ for some $\alpha > 0$).

Independently of Kesten's approach, the authors provided in [BKS99] a different way of looking at this problem (an approach more in the spirit of Noise Sensitivity). They noticed the remarkable property that if a **monotone** Boolean function f happens to be correlated very little with Majority functions (for all subsets of the bits); then $\mathbf{H}(f)$ has to be very small, and hence the function has to be sensitive to noise. They obtained a quantitative version of this statement that we briefly state here.

Let $f : \{-1, 1\}^n \to \{0, 1\}$ be a Boolean function. We want to use its correlations with Majority functions. Let us define these: for all $K \subset [n]$, define the Majority function on the subset K by $\mathbf{M}_K(x_1, \ldots, x_n) := \operatorname{sign} \sum_K x_i$ (where $\operatorname{sign} 0 := 0$ here). The correlation of the Boolean function f with these Majority functions is measured by

$$\Lambda(f) := \max_{K \subset [n]} |\mathbb{E}\big[f\mathbf{M}_K\big]| .$$

Being correlated very little with Majority functions corresponds to $\Lambda(f)$ being very small. The following quantitative theorem about correlation with Majority is proved in [BKS99]

Theorem 4.1. *There exists a universal constant $C > 0$ such that for any $f : \Omega_n \to \{0, 1\}$* **monotone**

$$\mathbf{H}(f) \leq C\Lambda(f)^2(1 - \log \Lambda(f)) \log n$$

(the result remains valid if f has values in $[0, 1]$ instead).

With this result at their disposal, in order to obtain fast convergence of $\mathbf{H}(f_n)$ to zero in the context of percolation crossings, the authors of [BKS99] investigated the correlations of percolation crossings f_n with Majority on subsets $K \subset [n]$. They showed that there exist $C, \alpha > 0$ universal constants, so that for any subset of the lattice K, $|\mathbb{E}\big[f_n\mathbf{M}_K\big]| \leq C\, n^{-\alpha}$. For this purpose, they used a nice appropriate Randomized Algorithm. We will not detail this algorithm used in [BKS99], since it was considerably strengthened in [SS10b]. We will now describe the approach of [SS10b] and then return to "correlation with Majority" using the stronger algorithm from [SS10b].

4.2 The Schramm/Steif approach

The authors in [SS10b] introduced the following beautiful and very general idea: suppose a real-valued function, $f : \Omega_n \to \mathbb{R}$ can be **exactly** computed with a

randomized algorithm A, so that every fixed variable is used by the algorithm A only with small probability; then this function f has to be of 'High frequency' with quantitative bounds which depend on how unlikely it is for any variable to be used by the algorithm.

4.2.1 Randomized algorithms, revealment and examples

Let us now define more precisely what types of randomized algorithms are allowed here. Take a function $f : \{-1,1\}^n \to \mathbb{R}$. We are looking for algorithms which compute the output of f by examining some of the bits (or variables) **one by one**, where the choice of the next bit may depend on the set of bits discovered so far, plus if needed, some additional randomness. We will call an algorithm satisfying this property a **Markovian (randomized) algorithm**. Following [SS10b], if A is a Markovian algorithm computing the function f, we will denote by $J \subset [n]$ the (random) set of bits examined by the algorithm.

In order to quantify the property that variables are unlikely to be used by an algorithm A, we define the **revealment** $\boldsymbol{\delta} = \boldsymbol{\delta}_A$ of the algorithm A to be the supremum over all variables $i \in [n]$ of the probability that i is examined by A. In other words,

$$\boldsymbol{\delta} = \boldsymbol{\delta}_A = \sup_{i \in [n]} \mathbb{P}\big[i \in J\big] .$$

We can now state one of the main theorems from [SS10b] (we will sketch its proof in the next subsection)

Theorem 4.2. *Let $f : \{-1,1\}^n \to \mathbb{R}$ be a function. Let A be a Markovian randomized algorithm for f having revealment $\boldsymbol{\delta} = \boldsymbol{\delta}_A$. Then for every $k = 1, 2, \ldots,$ The "level k"-Fourier coefficients of f satisfy*

$$\sum_{S \subset [n], |S| = k} \hat{f}(S)^2 \leq \boldsymbol{\delta}_A \, k \, \|f\|_2^2 .$$

Remark 4.3. If one is looking for a Markovian algorithm computing the output of the Majority function on n bits, then it is clear that the only way to proceed is to examine variables one at a time (the choice of the next variable being irrelevant since they all play the same role). The output will not be known until at least half of the bits are examined; hence the revealment for Majority is at least $1/2$.

In the case of percolation crossings, as opposed to the above case of Majority, one has to exploit the "richness" of the percolation picture in order to find algorithms which detect crossings while examining very few bits. A natural idea for a left to right crossing event in a large rectangle is to use an **exploration path**. The idea of an exploration path, which was highly influential in the introduction by Schramm of the SLE processes, was pictured in subsection 2.3 in the case of the triangular lattice.

More precisely, for any $n \geq 1$, Let D_n be a domain consisting of hexagons of meash $1/n$ approximating the square $[0,1]^2$, or more generally any smooth "quad"

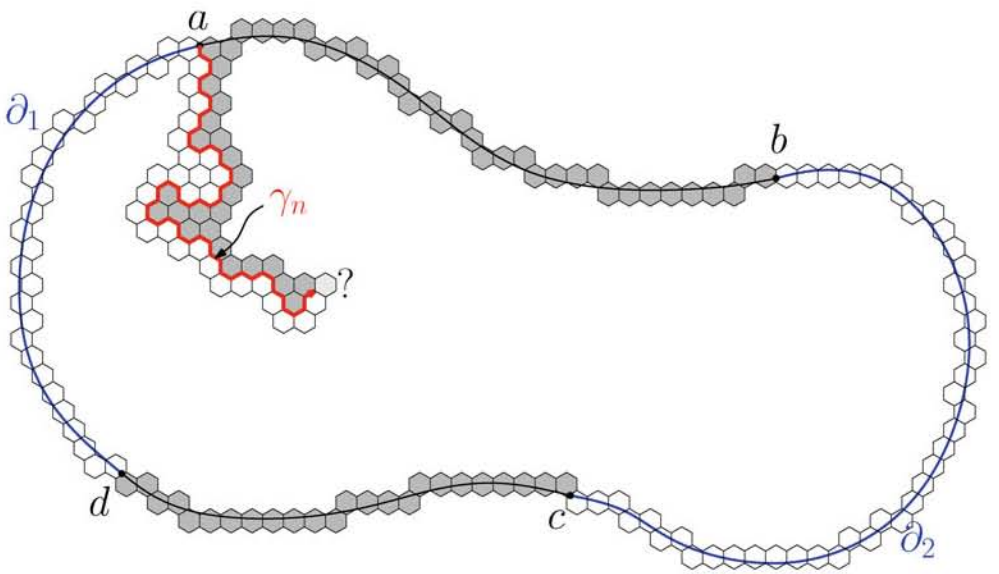


Figure 4.1: The random interface γ_n is discovered one site at a time (which makes our algorithm Markovian). The exploration stops when γ_n reaches either (bc) (in which case $f_n = 1$) or (cd) $(f_n = 0)$.

Ω with two prescribed arcs $\partial_1, \partial_2 \subset \Omega$, see figure 4.1. We are interested in the Left to Right crossing event (in the general setting, we look at the crossing event from ∂_1 to ∂_2 in D_n). Let f_n be the corresponding Boolean function and call γ_n the "exploration path" as in figure 4.1 (which starts at the upper left corner a). We run this exploration path until it reaches either the bottom side (in which case $f_n = 0$) or the right side (corresponding to $f_n = 1$).

This thus provides us with a markovian algorithm to compute f_n where the set $J = J_n$ of bits examined by the algorithm is the set of 'hexagons' touching γ_n on either sides. The nice property of both the exploration path γ_n and its $1/n$-neighborhood J_n, is that they have a scaling limit when the mesh $1/n$ goes to zero, this scaling limit being the well-known SLE_6 (this scaling limit of the exploration path was as we mentioned above one of the main motivations of Schramm to introduce these SLE processes). This scaling limit is a.s. a random fractal curve in $[0,1]^2$ (or Ω) of Hausdorff dimension $7/4$. This means that asymptotically, the exploration path uses a very small amount of all the bits included in this picture. With some more work (see [SW01, Wer07]), we can see that inside the domain (not near the corner or the sides), the probability for a point to be on γ_n is of order $n^{-1/4+o(1)}$, where $o(1)$ goes to zero as the mesh goes to zero.

One therefore expects the revealment $\boldsymbol{\delta}_n$ of this algorithm to be of order $n^{-1/4+o(1)}$. But the corner/boundaries have a non-trivial contribution here: for example in this setup, the single hexagon on the upper-left corner of the domain (where the interface starts) will be used systematically by the algorithm making the revealment equal to

one! There is an easy way to handle this problem: the idea in [SS10b] is to use some additional randomness and to start the exploration path from a random point x_n on the left-side of $[0,1]^2$. Doing so, this "smoothes" the singularity of the departure along the left boundary. There is a small counterpart to this: with this setup, one interface might not suffice to detect the existence of a left to right crossing and a second interface starting from the same random point x_n might be needed; see [SS10b] for more details. Using **arms exponents** from [SW01] and "quasimultiplicativity" of arms events ([Kes87, SS10b]), it can be checked that indeed the revealment of this modified algorithm is $n^{-1/4+o(1)}$.

Theorem 4.4. *Let $(f_n)_{n\geq 1}$ be the left to right crossing events in domains $(D_n)_{n\geq 1}$ approximating the unit square (or more generally a smooth domain Ω). Then there exists a sequence of Markovian algorithms, whose revealments $\boldsymbol{\delta}_n$ satisfy that for any $\epsilon > 0$,*

$$\boldsymbol{\delta}_n \leq C\, n^{-1/4+\epsilon}\,,$$

where $C = C(\epsilon)$ depends only on ϵ.

Therefore, applying theorem 4.2, one obtains

Corollary 4.5. *Let $(f_n)_n$ be the above sequence of crossing events. Let $\mathscr{S}_{f_n}$ denote the Spectral sample of these Boolean functions. Since $\|f_n\|_2 \leq 1$, we obtain that for any sequence $(L_n)_{n\geq 1}$,*

$$\mathbb{P}\big[0 < |\mathscr{S}_{f_n}| < L_n\big] \leq \boldsymbol{\delta}_n\, L_n^2\,. \tag{4.1}$$

In particular, this implies that for any $\epsilon > 0$, $\mathbb{P}\big[0 < |\mathscr{S}_{f_n}| < n^{1/8-\epsilon}\big] \to 0$.

This result gives precise lower bound information about the "spectrum of percolation" (or its "Energy spectrum"). It implies a "polynomial sensitivity" of crossing events in the sense that for any level of noise $\epsilon_n \gg n^{-1/8}$, we have that $\mathbb{E}\big[f_n(\omega) f_n(\omega^{\epsilon_n})\big] - \mathbb{E}\big[f_n\big]^2 \to 0$.

Remark 4.6.

- Note that equation (4.1) gives a good control on the lower tail of the spectral distribution, and as we will see in the last section, these lower-tail estimates are essential in the study of dynamical percolation.
- Similar results are obtained by the authors in [SS10b] in the case of the $\mathbb{Z}^2$-lattice, except that for this lattice, conformal invariance and convergence towards SLE_6 are not known; therefore critical exponents such as $1/4$ are not available; still [SS10b] obtains polynomial controls (thus strengthening [BKS99]) but with small exponents (their value coming from RSW estimates).

4.2.2 Link with Correlation with Majority

Before proving theorem 4.2, let us briefly return to the original motivation of these types of algorithm. Suppose we have at our disposal the above Markovian Algorithms for the left to right crossings f_n with small revealments $\boldsymbol{\delta}_n = n^{-1/4+o(1)}$; then

it follows easily that the events f_n are very little correlated with Majority functions; indeed let $n \geq 1$ and $K \subset D_n \approx [0,1]^2$ some fixed subset of the bits. By definition of the revealment, we have that $\mathbb{E}\big[|K \cap J_n|\big] \leq |K|\boldsymbol{\delta}_n$. This means that on average, the algorithm visits very few variables belonging to K. Since

$$|\mathbb{E}\big[f_n \mathbf{M}_K\big]| = |\mathbb{E}\big[f_n \mathbb{E}\big[\mathbf{M}_K \;\big|\; J_n\big]\big]| \leq \mathbb{E}\big[|\mathbb{E}\big[\mathbf{M}_K \;\big|\; J_m\big]|\big]\,,$$

and using the fact that on average, $|K \cap J_m|$ is small compared to $|K|$, it is easy to deduce that there is some $\alpha > 0$ such that

$$\Lambda(f_n) = \max_K |\mathbb{E}\big[f_n \mathbf{M}_K\big]| \leq n^{-\alpha}\,.$$

This, together with theorem 3.3 and theorem 4.1 implies a logarithmic noise sensitivity for Percolation. In [BKS99], they rely on another algorithm which instead of following interfaces, in some sense "invades" clusters attached to the left hand side. Since clusters are of fractal dimension 91/48, intuitively their algorithm, if boundary issues can be properly taken care of, would give a bigger revealment of order $n^{-5/48+o(1)}$ (the notion of revealment only appeared in [SS10b]). So the major breakthrough in [SS10b] is that they simplified tremendously the role played by the algorithm by introducing the notion of revealment and they noticed a more direct link with the Fourier transform. Using their correspondence between algorithm and Spectrum, they greatly improved the control on the Fourier spectrum (polynomial v.s. logarithmic). Furthermore, they improved the randomized algorithm.

4.2.3 Proof of Theorem 4.2

Let $f : \{-1,1\}^n \to \mathbb{R}$ be some real-valued function and consider A, a Markovian algorithm associated to it with revealment $\boldsymbol{\delta} = \boldsymbol{\delta}_A$. Let $k \in \{1,\dots,n\}$; we want to bound from above the size of the level-k Fourier coefficients of f (i.e. $\sum_{|S|=k} \hat{f}(S)^2$). For this purpose, one introduces the function $g = g^{(k)} = \sum_{|S|=k} \hat{f}(S)\chi_S$, which is the projection of f onto the subspace of level-k functions. By definition, one has $\sum_{|S|=k} \hat{f}(S)^2 = \|g\|_2^2$.

Very roughly speaking, the intuition goes as follows: if the revealment $\boldsymbol{\delta}$ is small, then for low level k: there are few "frequencies" in $g^{(k)}$ which will be "seen" by the algorithm. More precisely, for any fixed "frequency" S, if $|S| = k$ is small, then with high probability none of the bits in S will be visited by the algorithm. This means that $\mathbb{E}\big[g^{(k)} \;\big|\; J\big]$ (recall J denotes the set of bits examined by the algorithm) should be of small L^2 norm compared to $g^{(k)}$. Now since $f = \mathbb{E}\big[f \;\big|\; J\big] = \sum_k \mathbb{E}\big[g^{(k)} \;\big|\; J\big]$, most of the Fourier transform should be supported on high frequencies. There is some difficulty in implementing this intuition, since the conditional expectations $\mathbb{E}\big[g^{(k)} \;\big|\; J\big]$ are not orthogonal.

Following [SS10b] very closely, one way to implement this idea goes as follows:

$$\mathbb{E}\big[g^2\big] = \mathbb{E}\big[fg\big] = \mathbb{E}\big[f\mathbb{E}\big[g \;\big|\; J\big]\big] \leq \|f\|_2 \|\mathbb{E}\big[g \;\big|\; J\big]\|_2\,. \tag{4.2}$$

As hinted above, the goal is to control the L^2 norm of $\mathbb{E}\big[g \bigm| J\big]$. In order to achieve this, it will be helpful to interpret $\mathbb{E}\big[g \bigm| J\big]$ as the expectation of a random function g_J whose construction is explained below.

Recall that J is the set of bits examined by the randomized algorithm A. Since the randomized algorithm depends on the configuration $\omega \in \{-1,1\}^n$ and possibly some additional randomness: one can view J as a random variable on some extended probability space $\tilde{\Omega}$, whose elements can be represented as $\tilde{\omega} = (\omega, \tau)$ (τ corresponding here to the additional randomness).

For any function $h : \{-1,1\}^n \to \mathbb{R}$, one defines the **random** function h_J which corresponds to the function h where bits in J are fixed to match with what was examined by the algorithm. More exactly, if $J(\tilde{\omega}) = J(\omega, \tau)$ is the random set of bits examined, then the random function $h_J = h_{J(\tilde{\omega})}$ is the function on $\{-1,1\}^n$ defined by $h_{J(\omega,\tau)}(\omega') := h(\omega'')$, where $\omega'' = \omega$ on $J(\omega,\tau)$ and $\omega'' = \omega'$ on $\{1,\ldots,n\} \setminus J(\omega,\tau)$. Now, if the algorithm has examined the set of bits $J = J(\tilde{\omega})$, then with the above definition it is clear that the conditional expectation $\mathbb{E}\big[h \bigm| J\big]$ (which is a measurable function of $J = J(\tilde{\omega})$) corresponds to averaging h_J over configurations ω' (in other words we average on the variables outside of $J(\tilde{\omega})$); this can be written as

$$\mathbb{E}\big[h \bigm| J\big] = \int h_J := \hat{h}_J(\emptyset),$$

where the integration $\int$ is taken with respect to the uniform measure on $\omega' \in \Omega_n$. In particular $\mathbb{E}\big[h\big] = \mathbb{E}\big[\int h_J\big] = \mathbb{E}\big[\hat{h}_J(\emptyset)\big]$. Since $(h^2)_J = (h_J)^2$, it follows that

$$\|h\|_2^2 = \mathbb{E}\big[h^2\big] = \mathbb{E}\big[\int h_J^2\big] = \mathbb{E}\big[\|h_J\|_2^2\big]\,. \tag{4.3}$$

Recall from (4.2) that it only remains to control $\|\mathbb{E}\big[g \bigm| J\big]\|_2^2 = \mathbb{E}\big[\hat{g}_J(\emptyset)^2\big]$. For this purpose, we apply Parseval to the (random) function g_J: this gives (for any $\tilde{\omega} \in \tilde{\Omega}$),

$$\hat{g}_J(\emptyset)^2 = \|g_J\|_2^2 - \sum_{|S|>0} \hat{g}_J(S)^2.$$

Taking the expectation over $\tilde{\omega} = (\omega, \tau) \in \tilde{\Omega}$, this leads to

$$\begin{aligned}
\mathbb{E}\big[\hat{g}_J(\emptyset)^2\big] &= \mathbb{E}\big[\|g_J\|_2^2\big] - \sum_{|S|>0} \mathbb{E}\big[\hat{g}_J(S)^2\big] \\
&= \|g\|_2^2 - \sum_{|S|>0} \mathbb{E}\big[\hat{g}_J(S)^2\big] \quad \text{By (4.3)} \\
&= \sum_{|S|=k} \hat{g}(S)^2 - \sum_{|S|>0} \mathbb{E}\big[\hat{g}_J(S)^2\big] \left\{ \begin{array}{l} \text{since } g \text{ is supported} \\ \text{on level-}k \text{ coefficients} \end{array} \right. \\
&\leq \sum_{|S|=k} \mathbb{E}\big[\hat{g}(S)^2 - \hat{g}_J(S)^2\big] \left\{ \begin{array}{l} \text{by restricting on} \\ \text{level-}k \text{ coefficients} \end{array} \right.
\end{aligned}$$

Now, since g_J is built randomly from g by fixing the variables in $J = J(\tilde{\omega})$, and since g by definition does not have frequencies larger than k, it is clear that

$$\hat{g}_J(S) = \left\{ \begin{array}{ll} \hat{g}(S) = \hat{f}(S), & \text{if } S \cap J(\tilde{\omega}) = \emptyset \\ 0, & \text{else.} \end{array} \right.$$

Therefore, we obtain

$$\|\mathbb{E}\big[g \mid J\big]\|_2^2 = \mathbb{E}\big[\hat{g}_J(\emptyset)^2\big] \leq \sum_{|S|=k} \hat{g}(S)^2\, \mathbb{P}\big[S \cap J \neq \emptyset\big] \leq \|g\|_2^2\, k\, \boldsymbol{\delta}\,.$$

Combining with (4.2), this proves theorem 4.2. □

4.3 Is there any better algorithm ?

One might wonder whether there exist better algorithms which detect left to right crossings (better in the sense of **smaller revealment**). The existence of such algorithms would immediately imply sharper concentration results for the "Fourier Spectrum of percolation" than in Corollary 4.5.

This question of the "most effective" algorithm is very natural and has already been addressed in another paper of Oded et al. [PSSW07], where they study Random Turn Hex. Roughly speaking the "best" algorithm might be close to the following: assume j bits (forming the set S_j) have been explored so far; then choose for the next bit to be examined, the bit having the highest chance to be **pivotal** for the left-right crossing conditionally on what was previously examined (S_j). This algorithm stated in this way does not require additional randomness. Hence one would need to randomize it in some way in order to have a chance to obtain a small revealment. It is clear that this algorithm (following pivotal locations) in some sense is "smarter" than the one above, but on the other hand its analysis is highly nontrivial. It is not clear to guess what the revealment for that algorithm would be.

Before turning to the next section, let us observe that even if we had at our disposal the most effective algorithm (in terms of revealment), it would not yet necessarily imply the expected concentration behavior of the spectral measure of f_n around its mean (which for an $n \times n$ box on $\mathbb{T}$ turns out to be of order $n^{3/4}$). Indeed, in an $n \times n$ box, the algorithm will stop once it has found a crossing from left to right OR a dual crossing from top to bottom. In either case, the lattice or dual path is at least of length n; therefore we always have $|J| = |J_n| \geq n$. But if $\boldsymbol{\delta}_n$ is the revealment of any algorithm computing f_n, then by definition of the revealment, one has $\mathbb{E}\big[|J_n|\big] \leq O(1)n^2\boldsymbol{\delta}_n$ (there are $O(1)n^2$ variables); since $|J_n| \geq n$, this implies that $\boldsymbol{\delta}_n$ is necessarily bigger than $O(1)n^{-1}$. Now by Corollary 4.5, one has that

$$\sum_{0<|S|<n^\alpha} \hat{f}_n(S)^2 \leq n^{2\alpha}\, \boldsymbol{\delta}_n\,.$$

Therefore, the restriction that $\boldsymbol{\delta}_n$ has to be bigger than $O(1)n^{-1}$ implies that with the above algorithmic approach, one cannot hope to control the Fourier tail of percolation above $n^{1/2+o(1)}$ (while as we will see in the next section, the Spectral Measure of f_n is concentrated around $n^{3/4}$).

The above discussion raises the natural question of finding **good lower-bounds** on the "revealment of percolation crossings". It turns out that one can obtain much

better lower bounds on the smallest possible revealment than the straightforward one obtained above. In our present case (percolation crossings), the best known lower bound on the revealment follows from the following Theorem by O'donnell and Servedio

Theorem 4.7 ([OS07]). *Let $f : \Omega_n \to \{-1, 1\}$ be a* **monotone** *Boolean function. Any randomized algorithm A which computes f satisfies the following*

$$\boldsymbol{\delta}_A \geq \frac{\mathbf{I}(f)^2}{n} = \frac{(\sum_k \mathbf{I}_k(f))^2}{n} .$$

In our case, f_n depends on $O(n^2)$ variables, and if we are on the triangular grid $\mathbb{T}$, the total influence is known to be of order $n^{3/4+o(1)}$. Hence the above theorem implies that

$$\boldsymbol{\delta}_n \geq n^{-1/2+o(1)} .$$

Note that one could have recovered this lower-bound also from the case $k = 1$ in Theorem 4.2. Now, using Corollary 4.5, this lower-bound shows that one cannot hope to obtain concentration results on $\hat{f}_n$ above level $n^{1/4+o(1)}$. Note that this is still far from the expected $n^{3/4+o(1)}$.

In fact, Oded (and his coauthors) obtained several results which can be used to give lower bounds on revealments. Since these results are related (but slightly *tangential*) to this survey we list some of them in this last subsection

4.4 Related works of Oded an randomized algorithms

The first related result was obtained in [OSSS05]. Their main theorem can be stated as follows

Theorem 4.8 ([OSSS05]). *For any Boolean function $f : \Omega_n \to \{-1, 1\}$ and any Markovian randomized algorithm A computing f, one has*

$$\mathrm{Var}[f] \leq \sum_{i=1}^{n} \boldsymbol{\delta}_i \, \mathbf{I}_i(f) ,$$

where for each $i \in [n]$, $\boldsymbol{\delta}_i$ is the probability that the variable i is examined by A. (in particular, with these notations, $\boldsymbol{\delta}_A := \sup_i \boldsymbol{\delta}_i$.)

This beautiful result can be seen as a strengthening of Poincaré's inequality which states that $\mathrm{Var}[f] \leq \sum_i \mathbf{I}_i(f)$. The proof of the latter inequality is straightforward and well-known, and in some sense the proof of the above theorem pays attention to what is "lost" when one derives Poincaré's inequality.

This result has deep applications in *complexity theory*. It has also clear applications in our context since it provides lower bounds on revealments. For example, together with Theorem 4.7, it implies the second related result of Oded we wish to mention: the following Theorem from [BSW05]

Theorem 4.9 ([BSW05]). *Let $f : \Omega_n \to \{-1, 1\}$ be any* **monotone** *Boolean function, then any Markovian randomized algorithm A computing f satisfies the following*

$$\boldsymbol{\delta}_A \geq \frac{\mathrm{var}[f]^{2/3}}{n^{1/3}} .$$

In [BSW05] (where other results of this kind are proved), it is also shown that this theorem is **sharp** up to logarithmic terms.

Since the derivation of this estimate is very simple, let us see how it follows from Theorems 4.7 and 4.8.

$$\begin{aligned}
\mathrm{Var}[f] &\leq \sum_i \boldsymbol{\delta}_i \, \mathbf{I}_i(f) \quad \text{note that monotonicity is not needed here} \\
&\leq \boldsymbol{\delta}_A \sum_i \mathbf{I}_i(f) \\
&\leq \boldsymbol{\delta}_A \sqrt{n \boldsymbol{\delta}_A} \quad \text{by Theorem 4.7} \\
&= \boldsymbol{\delta}_A^{3/2} \sqrt{n} \, ,
\end{aligned}$$

which concludes the proof. □

Note that for our example of percolation crossings, this implies the following lower bound on the revealment

$$\boldsymbol{\delta}_n \geq C n^{-2/3} .$$

This lower bound in this case is not quite as good as the one given by Theorem 4.7, but is much better than the easy lower bound of $\Omega(1/n)$.

5 The "geometric" approach

As we explained in the previous section, the randomized algorithm approach cannot lead to the expected sharp behavior of the Spectrum of percolation. In [GPS10], Pete, Schramm and the author obtain, using a new approach which will be described in this section, a sharp control of the Fourier Spectrum of percolation. This approach implies among other things exact bounds on the sensitivity of percolation (the applications of this approach to dynamical percolation will be explained in the next section).

5.1 Rough description of the approach and "spectral measures"

The general idea is to study the "geometry" of the frequencies S. For any Boolean function $f : \{-1, 1\}^n \to \{0, 1\}$, with Fourier expansion $f = \sum_{S \subset [n]} \hat{f}(S) \, \chi_S$, one can consider the different frequencies $S \subset [n]$ as "random" subsets of the bits $[n]$. "Random" according to which measure ? Since we are interested in quantities like correlations

$$\mathbb{E}\big[f(\omega) f(\omega^\epsilon)\big] = \sum_{S \subset [n]} \hat{f}(S)^2 \, (1 - \epsilon)^{|S|} ,$$

it is very natural to introduce the **Spectral measure** $\hat{\mathbb{Q}} = \hat{\mathbb{Q}}_f$ on subsets of $[n]$ defined by

$$\hat{\mathbb{Q}}\big[S\big] := \hat{f}(S)^2,\ S \subset [n]\,. \tag{5.1}$$

By Parseval, the total mass of the so-defined spectral measure is

$$\sum_{S\subset[n]} \hat{f}(S)^2 = \|f\|_2^2\,.$$

One can therefore define a **Spectral probability measure** $\hat{\mathbb{P}} = \hat{\mathbb{P}}_f$ on the subsets of $[n]$ by

$$\hat{\mathbb{P}} := \frac{1}{\|f\|^2}\hat{\mathbb{Q}}\,.$$

The random variable under this probability corresponds to the random frequency and will be denoted by $\mathscr{S} = \mathscr{S}_f$ (i.e. $\hat{\mathbb{P}}\big[\mathscr{S} = S\big] := \hat{f}(S)^2/\|f\|^2$). We will call $\mathscr{S}_f$ the **Spectral sample** of f.

Remark 5.1. Note that one does not need the Boolean hypothesis here: $\hat{\mathbb{P}}_f$ is defined similarly for any real-valued $f : \Omega_n \to \mathbb{R}$.

In the remainder of this section, it will be convenient for simplicity to consider our Boolean functions from $\{-1,1\}^n$ into $\{-1,1\}$ (rather than $\{0,1\}$) thus making $\|f\|^2 = 1$ and $\hat{\mathbb{Q}} = \hat{\mathbb{P}}$.

Back to our context of percolation, in the following, for all $n \geq 1$, f_n will denote the Boolean function with values in $\{-1,1\}$ corresponding to the Left-Right crossing in a domain $D_n \subset \mathbb{Z}^2$ (or $\mathbb{T}$) approximating the square $[0,n]^2$ (or more generally $n\cdot\Omega$ where $\Omega \subset \mathbb{C}$ is some smooth quad with prescribed boundary arcs). To these Boolean functions, one associates their spectral samples $\mathscr{S}_{f_n}$.

In [GPS10], we show that most of the Spectral mass of f_n is concentrated around $n^2\alpha_4(n)$ (which in the triangular grid $\mathbb{T}$ is known to be of order $n^{3/4}$). More exactly we obtain the following result

Theorem 5.2 ([GPS10])**.** *If f_n, $n \geq 1$ are the above indicator functions (in $\{-1,1\}$) of the Left to Right crossings, then*

$$\limsup_{n\to\infty} \sum_{0<|S|<\epsilon\, n^2\,\alpha_4(n)} \hat{f}_n(S)^2 \underset{\epsilon\to 0}{\longrightarrow} 0\,, \tag{5.2}$$

or equivalently in terms of the Spectral probability measures:

$$\limsup_{n\to\infty} \hat{\mathbb{P}}\big[0 < |\mathscr{S}_{f_n}| < \epsilon\, n^2\alpha_4(n)\big] \underset{\epsilon\to 0}{\longrightarrow} 0\,. \tag{5.3}$$

This result is optimal in localizing where the spectral mass is, since as we will soon see, it is easy to show that the upper tail satisfy

$$\limsup_{n\to\infty} \hat{\mathbb{P}}\big[|\mathscr{S}_{f_n}| > \frac{1}{\epsilon}\, n^2\alpha_4(n)\big] \underset{\epsilon\to 0}{\longrightarrow} 0\,.$$

Hence as hinted long ago in figure 2.5, we indeed obtain that most of the Spectral mass is localized around $n^2\alpha_4(n)$ ($\approx n^{3/4}$ on $\mathbb{T}$). It is worth comparing this result

with the bound given by the algorithmic approach in the case of $\mathbb{T}$ which lead to a Spectral mass localized above $\approx n^{1/8}$. We will later give sharp bounds on the behavior of the Lower tail (i.e. at what "speed" does it decay to zero below $n^2\alpha_4(n)$).

Even though we are only interested in the size $|\mathscr{S}_{f_n}|$ of the spectral sample, the proof of this result will go through a detailed study of the "geometry" of $\mathscr{S}_{f_n}$. The study of this random object for its own sake was suggested by Gil Kalai and was also considered by Boris Tsirelson in his study of *Black noises*. The general belief, for quite some time already, was that asymptotically $\frac{1}{n}\mathscr{S}_{f_n}$ should behave like a random Cantor set in $[0,1]^2$.

In some particular cases of Boolean functions f, the geometry of $\mathscr{S}_f$ can be exactly analyzed: for example in [Tsi99], Tsirelson considers coalescing Random Walks and follows the trajectory under the coalescing flow of some particle. The position of the tagged particle at time n can be represented by a real-valued function g_n of the array of bits which define the coalescing flow. Tsirelson shows that the projection of $\mathscr{S}_{g_n}$ on the time axis has the same law as the zero-set of a Random Walk. Asymptotically (at least once projected on the x-axis), the spectral sample is indeed looking like a random Cantor set, and the zeros of a Random walk have a simple structure which enables one to prove a sharp localization of the Spectral mass of $\mathscr{S}_{g_n}$ around $n^{1/2}$.

In the case we are interested in, (i.e. $\mathscr{S}_{f_n}$), we do not have at our disposal such a simple description of $\mathscr{S}_{f_n}$. Until recently, there was some hope that $\mathscr{S}_{f_n}$ would asymptotically be very similar to the set of pivotal points of f_n, $\mathscr{P}_{f_n}$, and that one could study the behavior of $|\mathscr{S}_{f_n}|$ by focusing on the geometry of $\mathscr{P}_{f_n}$. We will see that indeed $\mathscr{S}_{f_n}$ and $\mathscr{P}_{f_n}$ share many common properties; nevertheless they are conjectured to be asymptotically **singular** with respect to each other. This is somewhat "unfortunate" since the study of $\mathscr{S}_{f_n}$ turns out to be far less tractable than the study of $\mathscr{P}_{f_n}$.

Therefore, since the law of $\mathscr{S}_{f_n}$ cannot be "exactly" analyzed, the strategy to prove theorem 5.2 will be to understand some useful properties of the law of $\mathscr{S}_{f_n}$ which will then be sufficient to imply the desired concentration results. Imagine one could show that $\frac{1}{n}\mathscr{S}_{f_n}$ behaves like a Cantor set with a lot of **"independence"** built in (something like a supercritical Galton-Watson Tree represented in $[0,1]^2$ as in Mandelbrot Fractal Percolation). Then our desired result would easily follow. In some sense, we will show that $\frac{1}{n}\mathscr{S}_{f_n}$ indeed behaves like a Cantor set, but we will only obtain a very weak independence statement (it is a posteriori surprising that such a weak control on the dependencies within $\mathscr{S}_{f_n}$ could lead to theorem 5.2).

In the next subsections, we will list a few simple observations on $\mathscr{S}_{f_n}$, some of them illustrating why our spectral sample $\frac{1}{n}\mathscr{S}_{f_n}$ should asymptotically look like a "random Cantor set" of $[0,1]^2$.

5.2 First and second moments on the size of the spectral sample

Let $f : \{-1,1\}^n \to \mathbb{R}$ be any real-valued function. We have the following lemma on the the Spectral measure $\hat{\mathbb{Q}} = \hat{\mathbb{Q}}_f$ (we state the lemma in greater generality than for functions into $\{-1,1\}$ since we will need it later).

Lemma 5.3. *For any subset of the bits $A \subset [n]$, one has*

$$\hat{\mathbb{Q}}\big[\mathscr{S}_f \subset A\big] = \mathbb{E}\big[\mathbb{E}\big[f \bigm| A\big]^2\big] = \|\mathbb{E}\big[f \bigm| A\big]\|^2 , \tag{5.4}$$

where $\mathbb{E}\big[f \bigm| A\big]$ is the conditional expectation of f knowing the bits in A.

The proof of this lemma is straightforward and follows from the definition of $\hat{\mathbb{Q}}_f$:

$$\begin{aligned} \hat{\mathbb{Q}}\big[\mathscr{S}_f \subset A\big] &:= \sum_{S \subset A} \hat{f}(S)^2 \\ &= \|\sum_{S \subset A} \hat{f}(S)\chi_S\|^2 \\ &= \|\mathbb{E}\big[f \bigm| A\big]\|^2 . \end{aligned}$$

Now let us return to our particular case of a Boolean function $f : \{-1,1\}^n \to \{-1,1\}$. Recall that its set of pivotal points $\mathscr{P} = \mathscr{P}_f(\omega)$ is the random set of variables k for which $f(\omega) \neq f(\sigma_k \cdot \omega)$.

Lemma 5.4 (Observation of Gil Kalai)**.** *Let $f : \{-1,1\}^n \to \{-1,1\}$ be a Boolean function, then*

$$\hat{\mathbb{E}}\big[|\mathscr{S}_f|\big] = \mathbb{E}\big[|\mathscr{P}_f|\big] \text{ and } \hat{\mathbb{E}}\big[|\mathscr{S}_f|^2\big] = \mathbb{E}\big[|\mathscr{P}_f|^2\big] . \tag{5.5}$$

This lemma is a very useful tool. Its poof is simple, let us sketch how to derive the first moment; by definition one has $\hat{\mathbb{E}}\big[|\mathscr{S}_f|\big] = \sum_{k \in [n]} \hat{\mathbb{P}}\big[k \in \mathscr{S}_f\big]$. Now for $k \in [n]$,

$$\begin{aligned} \hat{\mathbb{P}}\big[k \in \mathscr{S}_f\big] &= 1 - \hat{\mathbb{P}}\big[\mathscr{S}_f \subset [n] \setminus k\big] \\ &= \|f\|^2 - \|\mathbb{E}\big[f \bigm| [n] \setminus k\big]\|^2 \\ &= \|f - \mathbb{E}\big[f \bigm| \{k\}^c\big]\|^2 = \|f \cdot 1_{k \in \mathscr{P}_f}\|^2 \\ &= \mathbb{P}\big[k \in \mathscr{P}_f\big] . \end{aligned}$$

With a similar computation, one can see that for any $k, l \in [n]$, $\hat{\mathbb{P}}\big[k, l \in \mathscr{S}_f\big] = \mathbb{P}\big[k, l \in \mathscr{P}_f\big]$ thus proving the second moment relation. □

Remark 5.5. We mentioned already that $\frac{1}{n}\mathscr{S}_{f_n}$ and $\frac{1}{n}\mathscr{P}_{f_n}$ are believed to be asymptotically singular. This lemma shows that in terms of size, they have the same first and second moments, but it is not hard to check that their higher moments differ.

Consequence for the Spectral measures of percolation crossings f_n . It is a standard fact in critical percolation (see [GPS10]) that $\mathbb{E}\big[|\mathscr{P}_{f_n}|^2\big] \leq O(1)\mathbb{E}\big[|\mathscr{P}_{f_n}|\big]^2$ (this

follows from the quasimultiplicativity property). It then follows from lemma 5.4 that

$$\hat{\mathbb{E}}\big[|\mathscr{S}_{f_n}|^2\big] \leq C\,\hat{\mathbb{E}}\big[|\mathscr{S}f_n|\big]^2\,.$$

Hence, using Paley-Zygmund inequality, there is a universal constant $c > 0$ such that

$$\hat{\mathbb{P}}\big[|\mathscr{S}_{f_n}| > c\,\hat{\mathbb{E}}\big[|\mathscr{S}_{f_n}|\big]\big] > c\,.$$

Since $\hat{\mathbb{E}}\big[|\mathscr{S}_{f_n}|\big] = \mathbb{E}\big[|\mathscr{P}_{f_n}|\big] \asymp n^2\alpha_4(n)$, only using simple observations, one already obtains that at least a "positive fraction" of the Spectral mass is localized around $n^2\alpha_4(n)$. This property was known for some time already; but in order to be useful, a kind of **0-1 law** needs to be proved here. This turns out to be much more difficult and in some sense this was the task of [GPS10].

Notice also that $\hat{\mathbb{E}}\big[|\mathscr{S}_{f_n}|\big] \asymp n^2\alpha_4(n)$ easily implies, by Markov's inequality, the estimate on the upper tail behavior mentioned above i.e.

$$\limsup_{n\to\infty}\ \hat{\mathbb{P}}\big[|\mathscr{S}_{f_n}| > \frac{1}{\epsilon}\,n^2\alpha_4(n)\big] \underset{\epsilon\to 0}{\longrightarrow} 0\,.$$

5.3 "Scale invariance" properties of $\mathscr{S}_{f_n}$

The goal of this subsection is to identify some easy properties of $\mathscr{S}_{f_n}$ that reveal a certain scale invariance behavior. This is a first step towards what we mean by describing the "geometry" of $\mathscr{S}_{f_n}$.

If one compares with the first two approaches (i.e. hypercontractivity and algorithmic), both of them gave estimates on the size of $\mathscr{S}_{f_n}$ but did not give any information on the typical shape of $\mathscr{S}_{f_n}$. With these approaches, one cannot make the distinction between $\mathscr{S}_{f_n}$ being typically "localized" (all the points within a Ball of small radius) or on the other hand a Spectrum $\mathscr{S}_{f_n}$ being "self-similar". We will see, with simple observations, that at least a positive fraction of the Spectral measure $\hat{\mathbb{P}}_{f_n}$ is supported on self-similar sets.

Before stating these, let us consider a simple analog: let $n := 2^k$ and let us consider a Fractal Percolation on $[0,n]^2$ with parameter p. This is defined iteratively in a very simple way, as a Galton-Watson process: we start at scale $n = 2^k$; assume we are left with a certain set of squares of side-length 2^l, $l \geq 1$, if Q is one of these squares of sidelength 2^l, then divide this square into 4 dyadic subsquares and keep these independently with probability p. Let M denote the random discrete set that remains at the end of the procedure. By definition on has $\mathbb{E}\big[|M|\big] = p^k n^2 = n^{2+\log_2 p}$. Now, let $1 \ll r \ll n$ be some *mesoscopic* scale; assume $r = 2^l$, $1 < l < k$. If one is interested in the random set M only up to scale r (i.e. we don't keep the detailed information below scale r); this corresponds to the set $M_{(r)}$ where we stop the above induction once we reach squares of sidelength r (hence $M \subset M_{(r)}$). By definition one has that $\mathbb{E}\big[|M_{(r)}|\big] = (n/r)^{2+\log_2 p}$ (where $|M_{(r)}|$ denotes the number of r-squares in $M_{(r)}$).

In many ways, $\mathscr{S}_{f_n}$ should behave in a similar way as these Fractal Percolation sets (which are by construction self-similar). In order to test the self-similarity of $\mathscr{S}_{f_n}$, let us introduce some mesoscopic scale $1 \ll r \ll n$; one would like to study the

Spectral sample $\mathscr{S}_{f_n}$ only at the level of detail of the scale r and then claim that this r-"smoothed" picture of $\mathscr{S}_{f_n}$ should look like $\mathscr{S}_{f_{n/r}}$.

Let us precisely define an r-smoothed picture of $\mathscr{S}_{f_n}$. Divide the plane into a lattice of $r \times r$ squares (i.e. $r\,\mathbb{Z}^2$), and for any subset S of the window $[0,n]^2$, we define

$$S_{(r)} := \{ \text{ those } r \times r \text{ boxes that intersect } S\}.$$

We would like to see now that if $\mathscr{S} \sim \hat{\mathbb{P}}_{f_n}$, then $\mathscr{S}_{(r)}$ is similar to $\mathscr{S}_{f_{n/r}}$. We have the following estimate on the size of $\mathscr{S}_{(r)}$

Lemma 5.6.

$$\hat{\mathbb{E}}_{f_n}\big[|\mathscr{S}_{(r)}|\big] \asymp \frac{n^2}{r^2}\alpha_4(r,n) \asymp \hat{\mathbb{E}}\big[|\mathscr{S}_{f_{n/r}}|\big] \approx \left(\frac{n}{r}\right)^{3/4},$$

where the last approximate equality is known only on $\mathbb{T}$.

The proof is the same as the one for lemma 5.4, except that instead of summing over points, we sum over squares of radii r in the window $[0,n]^2$. There are $(n/r)^2$ such squares, and for any such square $Q = Q_r$, it remains to compute

$$\hat{\mathbb{P}}_{f_n}\big[Q \in \mathscr{S}_{(r)}\big] = \hat{\mathbb{P}}_{f_n}\big[\mathscr{S} \cap Q \neq \emptyset\big].$$

As in lemma 5.4, one gets $\hat{\mathbb{P}}_{f_n}\big[\mathscr{S} \cap Q \neq \emptyset\big] = 1 - \hat{\mathbb{P}}_{f_n}\big[\mathscr{S} \subset Q^c\big] = \|f - \mathbb{E}\big[f \mid Q^c\big]\|^2$. This expression is obviously bounded from above by the probability that Q is a pivotal square for f (i.e. that without the information in Q, one cannot yet predict f). This happens only if there is a four-arm event around Q. Neglecting the boundary issues, this is of probability $\alpha_4(r,n)$. It can be shown that not only $\hat{\mathbb{P}}_{f_n}\big[\mathscr{S} \cap Q \neq \emptyset\big] \leq \alpha_4(r,n)$ but in fact $\hat{\mathbb{P}}_{f_n}\big[\mathscr{S} \cap Q \neq \emptyset\big] \asymp \alpha_4(r,n)$, thus concluding the (sketch of) proof. □

Let us now state two observations about the "local" scale invariance of $\mathscr{S}_{f_n}$. First, if we consider one of the squares Q of the above lattice of $r \times r$ squares: how does $\mathscr{S}_{f_n}$ look **inside** Q when one conditions on $Q \cap \mathscr{S}_{f_n} \neq \emptyset$? This is answered (in the averaged sense) by the following lemma

Lemma 5.7. *Let Q be an $r \times r$ square in $[0,n]^2$, then*

$$\hat{\mathbb{E}}_{f_n}\big[|\mathscr{S} \cap Q| \;\big|\; Q \cap \mathscr{S} \neq \emptyset\big] \asymp r^2\alpha_4(r) \approx r^{3/4}. \tag{5.6}$$

The proof works as follows: there are about r^2 sites (or edges for $\mathbb{Z}^2$) in Q, and if x if one of these we have

$$\begin{aligned}\hat{\mathbb{P}}_{f_n}\big[x \in \mathscr{S} \;\big|\; \mathscr{S} \cap Q \neq \emptyset\big] = \frac{\hat{\mathbb{P}}_{f_n}\big[x \in \mathscr{S}\big]}{\hat{\mathbb{P}}_{f_n}\big[\mathscr{S} \cap Q \neq \emptyset\big]} &= \frac{\mathbb{P}\big[x \in \mathscr{P}_{f_n}\big]}{\hat{\mathbb{P}}_{f_n}\big[\mathscr{S} \cap Q \neq \emptyset\big]} \\ &\asymp \frac{\alpha_4(n)}{\alpha_4(r,n)},\end{aligned}$$

neglecting as usual boundary issues (and using the proof of the previous lemma for estimating $\hat{\mathbb{P}}_{f_n}\big[\mathscr{S} \cap Q \neq \emptyset\big]$). By quasi-multiplicativity, this is indeed of order $\alpha_4(r)$. □

Now, if one now conditions on the whole $\mathscr{S}_{f_n}$ to be contained in one of these mesoscopic box Q (of side-length r), then under this strong conditioning, one expects that inside Q, $\mathscr{S}_{f_n}$ should look like $\mathscr{S}_{f_r}$ in an $r \times r$ window. Using the same techniques one can obtain (the conditioning here requires some more work)

Lemma 5.8. *Let Q be an $r \times r$-square inside $[0,n]^2$, then*

$$\hat{\mathbb{E}}_{f_n}\big[|\mathscr{S}| \;\big|\; \emptyset \neq \mathscr{S} \subset Q\big] \asymp r^2 \alpha_4(r)\,.$$

All of these observations are relatively easy, since they are just *averaged* estimates instead of almost sure properties. Notice that one can easily extend these lemmas to their second moment analog (but this does not yet lead to almost sure properties).

Before turning to the strategy of the proof of theorem 5.2, let us state a last result in a different vain, which illustrates the asymptotic Cantor set behavior of $\mathscr{S}_{f_n}$.

Theorem 5.9 ([GPS10])**.** *In the case of the triangular lattice $\mathbb{T}$, as $n \to \infty$, $\frac{1}{n}\mathscr{S}_{f_n}$ converges in Law towards a random subset $\mathscr{S}$ of $[0,1]^2$ which, if not empty, is almost surely a perfect compact set (i.e. without isolated points and in particular infinite);*

The proof of this result combines ideas from Tsirelson as well as a "noise"-representation of percolation by Schramm and Smirnov ([SS10a]). The proof of theorem 5.2 also works in the "continuous" setting and implies that the scaling limit of $\frac{1}{n}\mathscr{S}_{f_n}$ is (if not empty) an almost sure compact set of Hausdorff dimension 3/4.

5.4 Strategy assuming a good control on the dependency structure

In this subsection, we will give the basic strategy for proving theorem 5.2 but without being very careful concerning the dependency structure within the random set $\mathscr{S}_{f_n}$. In the next subsection, not being able in [GPS10] to obtain a sharp understanding of this dependency structure, we will modify the basic strategy accordingly. Along the explanation of the strategy we will be in good shape to state a more precise theorem (than theorem 5.2) on the Lower Tail Behavior of $|\mathscr{S}_{f_n}|$.

The proof of theorem 5.2 will heavily rely on the self-similarity of $\mathscr{S}_{f_n}$. Before explaining the strategy which will be used for the Spectral sample, let us explain the same strategy applied to a much simpler example.

Consider a simple Random Walk on $\mathbb{Z}$, (X_k) starting uniformly on $\{-\lfloor\sqrt{n}\rfloor, \dots, \lfloor\sqrt{n}\rfloor\}$. Let $Z_n := \{j \in [0,n],\, X_j = 0\}$ be the zero-set of (X_k) until time n. (We do not start the walk from 0, since as in the case of the Spectral sample in general quads Ω, with positive probability one can have $Z_n = \emptyset$). It is clear that $\mathbb{E}\big[|Z_n|\big] \asymp n^{1/2}$ as well as $\mathbb{E}\big[|Z_n|^2\big] \asymp n$. One wishes to prove a statement for $|Z_n|$ similar to Theorem 5.2 and

possibly with detailed information on the Lower Tail of the distribution of $|Z_n|$ (all of this of course is classical and very well-known but this example we believe helps as an analogy).

Let us consider as above a mesoscopic scale $1 \ll r \ll n$. If $Z = Z_n$ is the zero-set of the Random-Walk, we define as above $Z_{(r)}$ by dividing the line into intervals of length r; $Z_{(r)}$ being the set of these intervals that intersect $Z = Z_n$. If $J = J_r$ is one of these intervals, then it is clear that

$$\mathbb{E}\big[|Z_n \cap J| \;\big|\; Z_n \cap J \neq \emptyset\big] \asymp r^{1/2}\,. \tag{5.7}$$

This is the analog of one of the (easy) above observations on $\mathscr{S}_{f_n}$. More importantly, the simple structure of the Zero-set of the Random Walk enables to obtain very good control on the dependency structure of Z_n in the following sense: for any interval $J = J_r$ of length r, conditioned on the whole zero-set Z_n outside of J, one still has good control on the size of Z_n inside J. More precisely, there is a universal constant $c > 0$ such that

$$\mathbb{P}\big[|Z_n \cap J| \geq r^{1/2} \;\big|\; J \cap Z_n \neq \emptyset,\, Z_n \cap [n] \setminus J\big] \geq c\,. \tag{5.8}$$

This type of control is much stronger than the estimate 5.7 (or its analog 5.6). What does it say about the distribution of $|Z_n|$? Each time Z_n intersects an interval J of length r, independently of everything outside, it has a positive chance to be of size $r^{1/2}$ inside. Therefore, if one is interested in the Lower Tail estimate $\mathbb{P}\big[|Z_n| < r^{1/2}\big]$, it seems that under the conditioning $\{|Z_n| < r^{1/2}\}$, the set Z_n will "typically" touch few intervals of length r; in other words $|Z_{(r)}|$ will have to be small. Now, how does the set $Z_{(r)}$ look when it is conditioned to be of small size ? Intuitively, it is clear that conditioned on $Z_{(r)}$ to be very small, say $|Z_{(r)}| = 2$, the two intervals J_1 and J_2 intersecting Z_n will (with high conditional probability) be very close to each other. Indeed else, one would would have to pay twice the unlikely event to touch the line and leave it right away. So for k "small", one expects that $\mathbb{P}\big[|Z_{(r)}| = k\big] \approx \mathbb{P}\big[|Z_{(r)}| = 1\big] \asymp (r/n)^{1/2}$ (there are $\frac{n}{r}$ r-intervals, and each of these has probability $\asymp (r/n)^{3/2}$ to be the only interval crossed by Z_n, the boundary issues being treated easily). Summarizing the discussion, $\mathbb{P}\big[|Z_n| < r^{1/2}\big]$ should be of order $\mathbb{P}\big[|Z_{(r)}| \text{ is "small"}\big]$ (since the different intervals touched by Z_n behave more or less independently), which should be of order $\mathbb{P}\big[|Z_{(r)}| = 1\big]$; hence one expects

$$\mathbb{P}\big[|Z_n| < \sqrt{r}\,\big] \leq O(1)\sqrt{\frac{r}{n}}\,. \tag{5.9}$$

In order to make these heuristics precise: the strong control on the dependency structure (5.8) implies

$$\mathbb{P}\big[0 < |Z_n| < r^{1/2} \;\big|\; |Z_{(r)}| = k\big] < (1-c)^k\,, \tag{5.10}$$

from which follows

$$\mathbb{P}\big[0 < |Z_n| < r^{1/2}\big] \;\leq\; \sum_{k \geq 1} \mathbb{P}\big[|Z_{(r)}| = k\big](1-c)^k\,.$$

Thanks to the exponential decay of $(1-c)^k$, it is enough to obtain a precise understanding of $\mathbb{P}\big[|Z_{(r)}| = k\big]$ for small values of k and to make sure that this estimate does not explode exponentially fast for largest values of k. More precisely we need to show that for any $k \geq 1$,

$$\mathbb{P}\big[|Z_{(r)}| = k\big] \;\leq\; g(k)\,\mathbb{P}\big[|Z_{(r)}| = 1\big] \asymp g(k)\,\sqrt{\frac{r}{n}}\,, \tag{5.11}$$

where $g(k)$ is a sub-exponentially fast growing function (which does not depend on r or n). Plugging this estimate into (5.10) leads to the precise Lower Tail estimate (5.9).

The advantage of this strategy is that by introducing the mesoscopic scale r, and assuming a good dependency structure at that scale, precise Lower Tail estimates boil down to rather rough estimates as (5.11) which need to be sharp only for the very bottom part of the Tail-distribution of $|Z_{(r)}|$.

Needless to say, there are simpler ways to prove the Lower Tail estimate (5.11) for the zero-set of Random Walks, but back to our Spectral Sample $\mathscr{S}_{f_n}$, this is the way we will prove theorem 5.2.

Recall from lemma 5.7 that if Q is an $r \times r$ square in $[0,n]^2$, then

$$\hat{\mathbb{E}}_{f_n}\big[|\mathscr{S} \cap Q| \;\big|\; Q \cap \mathscr{S} \neq \emptyset\big] \asymp r^2\alpha_4(r) \approx r^{3/4}\,. \tag{5.12}$$

One can prove a second moment analog of this expression which leads to

$$\hat{\mathbb{P}}_{f_n}\big[|\mathscr{S} \cap Q| \geq a\,r^2\alpha_4(r) \;\big|\; Q \cap \mathscr{S} \neq \emptyset\big] \geq a\,,$$

for some absolute constant $a > 0$. Since we believe that $\mathscr{S}_{f_n}$ behaves asymptotically like a 'nice' random Cantor set, it is natural to expect a nice dependency structure for $\mathscr{S}_{f_n}$ similar to (5.8) for Z_n. It turns out that such a precise understanding is very hard to achieve (and is presently unknown), but let's assume for the remainder of this subsection that we have such a nice independence at our disposal. Hence one ends up with

$$\hat{\mathbb{P}}\big[0 < |\mathscr{S}_{f_n}| < a\,r^2\alpha_4(r)\big] \leq \sum_{k\geq 1} \hat{\mathbb{P}}\big[|\mathscr{S}_{(r)}| = k\big](1-c)^k\,. \tag{5.13}$$

As for the zero-set of Random Walk, $\mathscr{S}_{(r)}$ typically likes to be spread; hence, if one conditions on $|\mathscr{S}_{(r)}|$ to be small, it should be, with high conditional probability, very concentrated in space. This hints that for k "small" one should have

$$\hat{\mathbb{P}}\big[|\mathscr{S}_{(r)}| = k\big] \approx \hat{\mathbb{P}}\big[|\mathscr{S}_{(r)}| = 1\big] \asymp \frac{n^2}{r^2}\alpha_4(r,n)^2\,.$$

This estimate follows from the fact that there are $(n/r)^2$ squares Q of side-length r, and for each of these, the probability $\hat{\mathbb{P}}_{f_n}\big[\emptyset \neq \mathscr{S} \subset Q\big]$ can be computed using lemma 5.3. It is up to constants the square of the probability that Q is pivotal for f_n (i.e. away from boundaries it is $\alpha_4(r,n)^2$).

Now to conclude the proof (assuming a nice dependency structure), one needs the following control on the probability that $|\mathscr{S}_{(r)}|$ is very-small.

Proposition 5.10 ([GPS10], section 4)**.** *There is a sub-exponentially fast growing function $g(k)$, $k \geq 1$, such that for any $1 \leq r \leq n$*

$$\hat{\mathbb{P}}_{f_n}\big[|\mathscr{S}_{(r)}| = k\big] \leq g(k)\,\frac{n^2}{r^2}\alpha_4(r,n)^2\,. \tag{5.14}$$

The proof of this proposition (which constitutes a non-negligeable part of [GPS10]) is of a 'combinatorial' flavor. It involves an induction over scales which roughly speaking uses the fact that if $\mathscr{S}_{(r)}$ is both spread and of small size, then one can detect several large empty annuli. The main idea is to "classify" the set of all possible $\mathscr{S}_{(r)}\,, |\mathscr{S}_{(r)}| = k$ into broader families, each family corresponding to a certain empty annuli structure. This classification into classes is easier to handle, since the families are built in such a way that they keep only the meaningful geometric information (i.e. here the large empty annuli) and in some sense "average" over the remaining "microscopic information". The inductive proof reveals that the families which contribute the most (under $\hat{\mathbb{P}}_{f_n}$) are the one with fewer empty annuli (this corresponds to the fact that very small Spectral sets tend to be localized).

Remark 5.11. Notice that since $g(k) \ll \exp(k)$, then summing (6.7) with $r := 1$ from $k = 1$ to $k = \log n$, yields to

$$\hat{\mathbb{P}}_{f_n}\big[0 < |\mathscr{S}| < \log n\big] \leq \log(n)\, g(\log n)\, n^2\alpha_4(n)^2 \ll \log n\, n^{2+\epsilon}\alpha_4(n)^2,$$

for any exponent $\epsilon > 0$. Since $n^2\alpha_4(n)^2 \approx n^{-1/2}$ on $\mathbb{T}$ (and on $\mathbb{Z}^2$, it is known to be $\leq n^{-\alpha}$, for some $\alpha > 0$), proposition 5.10 by itself reproves the results from [BKS99] with better quantitative bounds. This thus gives a proof more combinatorial than analytic (hypercontractivity).

Now proposition 5.10 applied to the expected (5.13) (where we assumed a good dependency knowledge) leads to following result on the Lower Tail behavior of $\mathscr{S}_{f_n}$

Theorem 5.12 ([GPS10])**.** *Both on the triangular lattice $\mathbb{T}$ and on $\mathbb{Z}^2$, one has*

$$\hat{\mathbb{P}}\big[0 < |\mathscr{S}_{f_n}| < r^2\alpha_4(r)\big] \asymp \frac{n^2}{r^2}\alpha_4(r,n)^2\,,$$

where the constants involved in $\asymp$ are absolute constants.

This theorem is sharp (up to constants), and it obviously implies the weaker theorem 5.2.
In the case of the triangular lattice, where critical exponents are known, this can be written in a more classical form for Lower Tail estimates

Proposition 5.13 ([GPS10])**.** *For every $\lambda \in (0,1]$, one has*

$$\limsup_{n\to\infty} \hat{\mathbb{P}}\big[0 < |\mathscr{S}_{f_n}| \leq \lambda\, \hat{\mathbb{E}}\big[|\mathscr{S}_{f_n}|\big]\big] \asymp \lambda^{2/3}\,.$$

5.5 The weak dependency control we achieved and how to modify the strategy

In the previous subsection, we gave a rough sketch of the proof of theorem 5.12 (and thus of theorem 5.2) assuming a good knowledge on the dependency structure for $\mathscr{S}_{f_n}$. Indeed it is natural to believe that if Q is some square of side-length r in $[0,n]^2$ and if W is some subset of $[0,n]^2$ not too close to Q (for example $d(Q,W) \geq r$ is enough), then for any subsets $\emptyset \neq S \subset Q$, the following strong independence statement should hold

$$\hat{\mathbb{P}}_{f_n}\big[\mathscr{S} \cap Q = S \;\big|\; \mathscr{S} \cap Q \neq \emptyset\big] \asymp \hat{\mathbb{P}}_{f_n}\big[\mathscr{S} \cap Q = S \;\big|\; \mathscr{S} \cap Q \neq \emptyset,\, \mathscr{S} \cap W\big]\,.$$

It is natural to believe such a statement is true since, for example, it is known to hold for the random set of pivotals points $\mathscr{P}_{f_n}$. Since our goal is to understand the size $|\mathscr{S}_{f_n}|$, and since we expect the Spectral sample $\mathscr{S}$ to be typically of size $r^2\alpha_4(r)$ in a square Q satisfying $\mathscr{S} \cap Q \neq \emptyset$ (see lemma 5.7), it would be enough to prove the following slightly weaker statement: there is some constant $c > 0$ such that

$$\hat{\mathbb{P}}_{f_n}\big[|\mathscr{S} \cap Q| \geq r^2\alpha_4(r) \;\big|\; \mathscr{S} \cap Q \neq \emptyset,\, \mathscr{S} \cap W\big] \geq c\,, \tag{5.15}$$

which is the exact analog of (5.8). Unfortunately, such a statement is out of reach at the moment (due to a lack of knowledge and control on the law of $\mathscr{S}_{f_n}$), but in [GPS10] one could prove the following weaker independence statement.

Proposition 5.14 ([GPS10])**.** *There is a universal constant $c > 0$, such that for any square Q of side-length r in $[0,n]^2$ and any $W \subset [0,n]^2$ only satisfying $d(Q,W) \geq r$, one has*

$$\hat{\mathbb{P}}_{f_n}\big[|\mathscr{S} \cap Q| \geq r^2\alpha_4(r) \;\big|\; \mathscr{S} \cap Q \neq \emptyset,\, \mathscr{S} \cap W = \emptyset\big] \geq c\,.$$

This results strengthens the easy estimate from lemma 5.7 and constitutes another non-negligible part of [GPS10] (we will not comment on its proof here). It says that as far as we looked for $\mathscr{S}$ in some region W but did not find anything there, one can still control the size of the Spectrum $\mathscr{S}$ in any disjoint square Q (knowing $\mathscr{S} \cap Q \neq \emptyset$).

This control on the dependency structure is much weaker than the expected (5.15), but it turns out that one can modify the strategy explained in the above subsection in order to deal with this lack of independence. Let us briefly summarize what is the idea behind the modified approach.

Let $1 \leq r \leq n$, we want to estimate the Lower-Tail quantity $\mathbb{P}\big[0 < |\mathscr{S}_{f_n}| < r^2\alpha_4(r)\big]$. As in the definition of $\mathscr{S}_{(r)}$, we divide the window $[0,n]^2$ into a lattice of $r \times r$ squares. Now, the **naive** idea is to "scan" the window $[0,n]^2$ with r-squares Q scanned **one at a time** (there are $\asymp (n/r)^2$ such squares to scan). We hope to show that with high probability, if $\mathscr{S} \neq \emptyset$, then at some point we will find some "good" square Q, with $|\mathscr{S} \cap Q| > r^2\alpha_4(r)$. The problem being that, if by the time one encounters a good square, we discovered some points but only few of them in the region W that was scanned so far, then due to our weak dependency control (one keeps a good control only as far as $\mathscr{S} \cap W = \emptyset$), we cannot push the scanning procedure any longer.

One way to handle this problem is to scan the window in a **dilute manner**: i.e. at any time, the region W we discovered so far should be some well-chosen "sparse" set. It is not hard to figure out what should be the right "intensity" of this set: assume we have discovered the region W so far and that we did not find any Spectrum there (i.e. $\mathscr{S} \cap W = \emptyset$). We want to continue the exploration (or scanning) inside some unexplored square Q such that $d(Q, W) \geq r$. Assume furthermore that we know $\mathscr{S} \cap Q \neq \emptyset$. Then, we are exactly under a conditional Law on $|\mathscr{S} \cap Q|$ that we are able to handle using our proposition 5.14. At that point we might scan the whole square Q and by proposition 5.14, we have a positive probability to succeed (that is, to find more than $r^2\alpha_4(r)$ points in the Spectral sample); but in case we only find few points (say $\log r$ points), then, as mentioned above, we will not be able to continue the procedure. Thus, a natural idea is to explore each site (or edge) $x \in Q$ only with probability $p_r := (r^2\alpha_4(r))^{-1}$ (independently of the other sites of the square and independently of $\mathscr{S}$). As such one only scans a random subset $\mathcal{Z}$ of Q. The advantage of this particular choice is that in the situation where $\emptyset \neq \mathscr{S} \cap Q$ is of small cardinality (compared to $r^2\alpha_4(r)$), then there is a small chance that $\mathcal{Z} \cap \mathscr{S} \neq \emptyset$, and we will be able to continue the scanning procedure in a some new square Q' (with $W' = W \cup \mathcal{Z}$). On the other hand, if one discovers $\mathscr{S} \cap \mathcal{Z} \neq \emptyset$, then with high probability one can guess that $|\mathscr{S} \cap Q|$ is of order $r^2\alpha_4(r)$, which is what we are looking for.

There are two difficulties one has to face with this modified strategy:
• The first technicality is that in the above description, we needed to choose new squares Q satisfying $d(Q, W) \geq r$, as such we might not be able to scan the whole window $[0, n]^2$ and we might miss where the Spectrum actually lies. There is an easy way out here: if Q is some new square in the scanning procedure only satisfying $Q \cap W = \emptyset$, then one still has a good control on the size of the Spectral sample under the above conditioning but now restricted to the concentric square $\bar{Q} \subset Q$ of side-length $r/2$. More precisely we have

$$\hat{\mathbb{P}}_{f_n}\big[|\mathscr{S} \cap \bar{Q}| \geq r^2\alpha_4(r) \;\big|\; \mathscr{S} \cap Q \neq \emptyset,\, \mathscr{S} \cap W = \emptyset\big] \geq c\,.$$

Hence, we can now explore all the squares Q of the $r \times r$ subgrid of $[0, n]^2$ one at a time and look for the spectrum only within some dilute sets $\mathcal{Z}$ inside their concentric square $\bar{Q}$.

• The second difficulty is harder to handle. It has to do with the fact that if one applies the above "naive" strategy, i.e. scanning the squares one by one, then one needs to keep track of the conditional probabilities $\hat{\mathbb{P}}_{f_n}\big[\mathscr{S} \cap Q \neq \emptyset \;\big|\; \mathscr{S} \cap W = \emptyset\big]$ as the explored set W is growing along the procedure (indeed in proposition 5.14, we need to condition on $\mathscr{S} \cap Q \neq \emptyset$). These conditional probabilities turn out to be hard to control and because of this, exploring squares one at a time is not doable. Instead, in [GPS10], we rely on a more abstract or "averaged" scanning procedure where all squares Q are considered in some sense simultaneously (see [GPS10], "a large deviation lemma", section 6).

To summarize, the proof of theorem 5.12 (which implies theorem 5.2), relies on three main steps:

1. Obtaining a Sharp control on the very beginning of the Lower Tail of the Spectrum $\mathscr{S}_{(r)}$ for all mesoscopic scales $1 \leq r \leq n$: this is given by proposition 5.10 (this step gives an independent proof of theorem 3.3).

2. Proving a sufficient control on the dependency structure within $\mathscr{S}_{f_n}$: proposition 5.14

3. Adapting the "naive" strategy of scanning mesoscopic squares Q of size r "one at a time" into an averaged procedure which avoids keeping track of conditional probabilities like $\hat{\mathbb{P}}_{f_n}\left[\mathscr{S} \cap Q \neq \emptyset \;\middle|\; \mathscr{S} \cap W = \emptyset\right]$ (section 6 in [GPS10]).

Explaining in details all these steps would take us too far afield in this survey, but we hope that this explanation as well as the analogy with the much simpler case of the zeros of Random Walks gives some intuition on how the proof works.

6 Applications to Dynamical Percolation

We already mentioned applications to First-Passage Percolation in the hypercontractivity section. Let us now present a very natural model where percolation undergoes a time-evolution: the model of **dynamical percolation** described below. The study of the "dynamical" behavior of percolation as opposed to its "static" behavior turns out to be very rich: interesting phenomena arise especially at the point where the phase transition takes place. We will see that in some sense, dynamical **planar** percolation at criticality is a very unstable (or chaotic) process. In order to understand this instability, sensitivity of percolation (and so its Fourier analysis) will play a key role.

6.1 The model of dynamical percolation

As mentioned earlier in the introduction of this survey, **dynamical percolation** was introduced in [HPS97] as a natural dynamic illustrating *thermal agitation* (similar as *Glauber-dynamics* for the Ising Model).

The dynamic introduced in [HPS97] (also invented independently by Itai Benjamini) is essentially the unique dynamic which preserves the i.i.d law of percolation and which is Markovian.

Due to the i.i.d structure of the percolation model, the dynamic is very simple: on $\mathbb{Z}^d$, at parameter $p \in [0,1]$, each edge $e \in \mathbb{E}^d$ is updated at rate one independently of the other edges. This means that at rate one, independently of everything else, one keeps the edge with probability p, and removes it with probability $1-p$. See [HPS97, SS10b] and especially [Ste09] for background on this model.

In [HPS97], it is shown that no surprises arise outside of the critical point $p_c(\mathbb{Z}^d)$ in dimension $d \geq 2$. If $(\omega_t^p)_{t \geq 0}$ denotes a trajectory of percolation configurations of parameter p on $\mathbb{Z}^d$ evolving as described above, they prove that if $p > p_c$, then almost surely, there is an infinite cluster in ALL configurations ω_t^p, $t \geq 0$ (as well, in the subcritical regime $p < p_c$, a.s. clusters remain finite along the dynamic).

What happens at the critical point $p_c(\mathbb{Z}^d)$ requires more care: in [HPS97], they prove that if the dimension is high enough ($d \geq 19$), then as in the subcritical regime, the clusters of $\omega_t^{p_c}$ remain finite as the time t goes on. Their proof relies essentially on the "mean-field" behavior of percolation rigorously obtained by Hara and Slade ([HS90]) in dimension $d \geq 19$ (conjectured to hold for $d > 6$ and $d = 6$ with 'logarithmic' corrections).

In lower dimensions, the phase transition is somehow more 'abrupt'. This can be seen by the fact that the density of the infinite cluster represented by $p \mapsto \theta(p)$ has an infinite derivative at the critical point (see [KZ87] in $d = 2$). Because of this phenomenon, one cannot easily rule out the sudden appearance of infinite clusters along the dynamic; consequently the study of dynamical percolation needs a more detailed study. We will restrict ourselves to the case of dimension $d = 2$, since in dimension $d = 3$ say, very little is known (even the static behavior of $3d$-percolation at the critical point is unknown, which makes its dynamical study at this point hopeless).

In the planar case $d = 2$, where there is a very rich literature on the critical behavior, it was asked back in [HPS97] whether at the critical point (say on $\mathbb{Z}^2$ at $p_c = 1/2$), such exceptional infinite clusters appear along the dynamic or not. More exactly, is it the case that there exists almost surely a (random) subset $\emptyset \neq \mathcal{E} \subset \mathbb{R}$ of **exceptional times**, such that for any $t \in \mathcal{E}$, there is an infinite cluster in $\omega_t^{p_c}$? We will see in the present section that there indeed exists such exceptional times. This illustrates the above mentioned unstable character of dynamical percolation at criticality. Note that from the continuity of the phase transition in $d = 2$, it easily follows that the exceptional set $\mathcal{E}$ is almost surely of Lebesgue measure zero.

This appearance of exceptional infinite clusters along the dynamic will be guaranteed by the fact that large scale connectivity properties should decorrelate very fast. Roughly speaking, if the large scale geometry of the configurations $\omega_t^{p_c}$ "changes" very quickly as time goes on, then it will have better chances to "catch" infinite clusters. As we argued earlier, the initial study of noise sensitivity in [BKS99] was partially motivated by this problem since the rapid decorrelation needed to prove the existence of exceptional times corresponds exactly to their notion of noise sensitivity. The results obtained in [BKS99] already illustrated the fact that at large scales, connectivity properties of $(\omega_t)_{t\geq 0}$ evolve very quickly. Nonetheless, their 'logarithmic' control of noise sensitivity was not sufficient to imply the appearance of infinite clusters. The first proof of the above conjecture (about the existence of exceptional times) in the context of a two-dimensional lattice was achieved by Schramm and Steif in [SS10b] for the triangular lattice $\mathbb{T}$. Their proof relies on their algorithmic approach of the 'Fourier spectrum' of percolation explained in section 4. Before explaining in more details the results from [SS10b] and [GPS10], let us present what is the strategy to prove the existence of exceptional times as well as a variant model of dynamical percolation in $d = 2$ introduced by Benjamini and Schramm where the decorrelation structure is somehow easier to handle (does not require any Fourier analysis).

6.2 Second moment analysis and the "exceptional planes of percolation"

Suppose we are considering some model of percolation at criticality; for example $\mathbb{Z}^2$ or the triangular lattice $\mathbb{T}$ at $p_c = 1/2$; but it could also be Boolean Percolation (a Poisson Point Process of Balls in the plane) at its critical intensity. Let's consider a trajectory of percolation configurations $(\omega_t)_{t\geq 0}$ for the model under consideration evolving as above (in the case of the Boolean Model, one could think of many natural dynamics, for example each ball moving independently according to a Brownian motion). One would like to prove the existence of **exceptional times** t for which there is an infinite cluster in ω_t. It turns out to be more convenient to show the existence of exceptional times where the origin itself is connected to infinity via some open path; furthermore it is enough to show that with positive probability there is some time $t \in [0,1]$ where $0 \overset{\omega_t}{\longleftrightarrow} \infty$.

For any large radius $R \gg 1$, let us then introduce Q_R to be the set of times where the origin 0 is connected to distance R:

$$Q_R := \{t \in [0,1] : 0 \overset{\omega_t}{\longleftrightarrow} R\}.$$

Proving the existence of exceptional times boils down to proving that with positive probability $\cap_{R>0} Q_R \neq \emptyset$. Even though the sets Q_R are not closed, with some additional technicality (see [SS10b]), it is enough to prove that $\inf_{R>1} \mathbb{P}\big[Q_R \neq \emptyset\big] > 0$.

In order to achieve this, we introduce the random variable X_R corresponding to the amount of time where 0 is connected to distance R

$$X_R := \int_0^1 1_{0 \overset{\omega_t}{\longleftrightarrow} R}\, dt.$$

Our goal is therefore to prove that there is some universal constant $c > 0$, so that $\mathbb{P}\big[X_R > 0\big] > c$ for any radius R. As usual, this is achieved by the second moment method:

$$\mathbb{P}\big[Q_R \neq \emptyset\big] = \mathbb{P}\big[X_R > 0\big] \geq \frac{\mathbb{E}\big[X_R\big]^2}{\mathbb{E}\big[X_R^2\big]},$$

it remains to prove the existence of some constant $C > 0$ such that for all $R > 1$, $\mathbb{E}\big[X_R^2\big] < C\,\mathbb{E}\big[X_R\big]^2$. Now, notice that the second moment can be written

$$\begin{aligned}\mathbb{E}\big[X_R^2\big] &= \iint_{\substack{0\leq s\leq 1\\ 0\leq t\leq 1}} \mathbb{P}\big[0 \overset{\omega_s}{\longleftrightarrow} R,\, 0 \overset{\omega_t}{\longleftrightarrow} R\big]\, ds dt \\ &\leq 2\int_0^1 \mathbb{P}\big[0 \overset{\omega_0}{\longleftrightarrow} R, 0 \overset{\omega_t}{\longleftrightarrow} R\big]\, dt\,. \end{aligned} \tag{6.1}$$

One can read from this expression that a fast decorrelation over time of the event $0 \overset{\omega}{\longleftrightarrow} R$ will imply the existence of exceptional times. The study of the correlation structure of these events usually goes through Fourier analysis unless there is a more direct way to get a hand on these (we will give such an example of dynamical percolation below). Fourier analysis helps as follows: if f_R denotes

the indicator function of the event $\{0 \overset{\omega}{\longleftrightarrow} R\}$ then, as a Boolean function (of the bits inside the disk of radius R into $\{0,1\}$), it has a Fourier-Walsh expansion. Now, as was explained in subsection 2.1, the desired correlations are simply expressed in terms of the Fourier coefficients of f_R

$$\begin{aligned}\mathbb{P}\big[0 \overset{\omega_0}{\longleftrightarrow} R,\, 0 \overset{\omega_t}{\longleftrightarrow} R\big] &= \mathbb{E}\big[f_R(\omega_0) f_R(\omega_t)\big] \\ &= \mathbb{E}\big[\big(\sum_{S\subset B(0,R)} \hat{f}_R(S)\chi_S(\omega_0)\big)\big(\sum_{S\subset B(0,R)} \hat{f}_R(S)\chi_S(\omega_t)\big)\big] \\ &= \mathbb{E}\big[f_R\big]^2 + \sum_{\emptyset\not\subset S\subset B(0,R)} \hat{f}_R(S)^2 \exp(-t|S|)\end{aligned}$$

We see in this expression, that for small times t, the frequencies contributing in the correlation between $\{0 \overset{\omega_0}{\longleftrightarrow} R\}$ and $\{0 \overset{\omega_t}{\longleftrightarrow} R\}$ are of 'small' size $|S| \lesssim 1/t$. In particular, in order to detect the existence of exceptional times, one needs to achieve good control on the **Lower tail** of the **Fourier Spectrum** of f_R.

We will give some more details on this setup in the coming subsection, but before going more deeply into the Fourier side, let us briefly present the first planar percolation model for which this phenomenon of exceptional infinite clusters was proved: *exceptional planes of percolation*, [BS98b].

The model can be described as follows: sample balls of radius $r > 0$ in $\mathbb{R}^4$ according to a Poisson Point Process with Lebesgue intensity 1 on $\mathbb{R}^4$. Let $\mathcal{U}_r$ denote the union of these balls. Now, for each plane $P \subset \mathbb{R}^4$, let $\omega_P^r = \omega_P$ be the percolation configuration corresponding to the set $\mathcal{U}_r \cap P$. For a fixed plane P, ω_P^r follows the law of a Boolean Percolation Model with random i.i.d radii (in $[0, r]$).

This provides us with a natural planar dynamical percolation where the dynamic is here parametrized by planes $P \subset \mathbb{R}^4$ instead of time. It is known ([MR96]) that there is a critical parameter $r_c > 0$ so that the corresponding planar model is critical (with a.s. no infinite cluster at the critical radius r_c). This means that at criticality, for "almost all" planes $P \subset \mathbb{R}^4$, there is no infinite cluster in ω_P. Whence the natural question: are there are exceptional planes of percolation or not ?

Theorem 6.1 (Benjamini and Schramm [BS98b])**.** *At the critical radius r_c, there exist exceptional planes $P \in \mathbb{R}^4$ for which $\omega_P := \mathcal{U}_{r_c} \cap P$ has an infinite cluster.*

The proof of this result follows the above setup. To prove the existence of such exceptional planes, it will be enough to restrict ourselves to the subclass of planes P which include the origin and we will ask that the origin itself is connected to infinity within ω_P. If P_0 is some plane of reference, by the second moment method, the key quantity to estimate is the correlation $\mathbb{P}\big[0 \overset{\omega_{P_0}}{\longleftrightarrow} R,\, 0 \overset{\omega_P}{\longleftrightarrow} R\big]$; this correlation being then integrated uniformly over planes P containing the origin.

The big advantage of this model is that if $P \neq P_0$, since the radius r_c of the balls in $\mathbb{R}^4$ is finite, percolation configurations ω_P and ω_{P_0} will "interact" together only within a finite radius $R_{P_0,P}$ which depends on the angle $\theta_{P_0,P}$ formed by the two planes (the closer they are, the further away they interact). We can therefore

bound the correlation as follows (using similar notations as in subsection 2.3)

$$\begin{aligned}\mathbb{P}\big[0 \overset{\omega_{P_0}}{\longleftrightarrow} R,\, 0 \overset{\omega_P}{\longleftrightarrow} R\big] &\leq \mathbb{P}\big[0 \overset{\omega_{P_0}}{\longleftrightarrow} R_{P_0,P},\, R_{P_0,P} \overset{\omega_{P_0}}{\longleftrightarrow} R,\, R_{P_0,P} \overset{\omega_P}{\longleftrightarrow} R\big] \\ &\leq \alpha_1(R_{P_0,P})\alpha_1(R_{P_0,P}, R)^2,\end{aligned}$$

by independence and since configurations do not interact after $R_{P_0,P}$. Now, if one had at our disposal a RSW theorem similar as what is known for $\mathbb{Z}^2$ percolation, this would imply a quasimultiplicativity result on the one-arm event:

$$\alpha_1(r_1, r_3) \asymp \alpha_1(r_1, r_2)\alpha_1(r_2, r_3),$$

for all $r_1 \leq r_2 \leq r_3$. It would then follow that the correlation is bounded by $O(1)\alpha_1(R_{P_0,P})^{-1}\alpha_1(R)^2$ (see [BS98b] where they relied on a more complicated argument since RSW was not available).

In the second moment argument, we need to bound the second moment by the squared of the first one (here $\alpha_1(R)^2$). Hence it remains to show that $\int_P \alpha_1(R_{P_0,P})^{-1}\, d\lambda(P) < \infty$. This can be done using the expression of $R_{P_0,P}$ in terms of the angle $\theta_{P_0,P}$ as well as some control on the decay of $\alpha_1(u)$ as $u \to \infty$. See [BS98b] where the computation is carried out.

6.3 Dynamical percolation, later contributions

In this subsection, we present the results from [SS10b] and [GPS10] on dynamical percolation.

6.3.1 Consequences of the algorithmic approach ([SS10b])

As we described in section 4, Schramm and Steif introduced a powerful technique in [SS10b] to control the Fourier coefficients of general Boolean functions $f : \{-1, 1\}^n \to \{0, 1\}$. We have already seen in Corollary 4.5 that their approach enabled one to obtain a polynomial control on the sensitivity of percolation crossings, thus strengthening previous results from [BKS99]. But more importantly, using their technique (mainly theorem 4.2) they obtained deep results on dynamical percolation on the triangular grid $\mathbb{T}$. Their main result can be summarized in the following theorem (see [SS10b] for their other contributions)

Theorem 6.2 (Schramm and Steif, [SS10b])**.** *For dynamical percolation on the triangular lattice $\mathbb{T}$ at the critical point $p_c = 1/2$, one has*

- *Almost surely, there exist exceptional times $t \in [0, \infty]$ such that ω_t has an infinite cluster.*
- *Furthermore, if $\mathcal{E} \subset [0, \infty]$ denotes the (random) set of these exceptional times, then the Hausdorff dimension of $\mathcal{E}$ is an almost sure constant in $[1/6, 31/36]$.*

They conjectured that the dimension of these exceptional times is a.s. $31/36$. Let us now explain how they obtained such a result.

As we outlined above, in order to prove the existence of exceptional times, one needs to prove fast decorrelation of $f_R(\omega) := 1_{\{0 \overset{\omega}{\longleftrightarrow} R\}}$. Recall that the correlation between the configuration at time 0, ω_0 and the configuration at time t, ω_t can be written

$$\begin{aligned}
\mathbb{P}\big[0 \overset{\omega_0}{\longleftrightarrow} R,\, 0 \overset{\omega_t}{\longleftrightarrow} R\big] &= \mathbb{E}\big[f_R(\omega_0) f_R(\omega_t)\big] \\
&= \mathbb{E}\big[f_R\big]^2 + \sum_{\emptyset \not\subset S \subset B(0,R)} \hat{f}_R(S)^2 \exp(-t|S|) \\
&\lesssim \mathbb{E}\big[f_R\big]^2 + O(1) \sum_{k=1}^{1/t} \sum_{|S|=k} \hat{f}_R(S)^2 \,. \qquad (6.2)
\end{aligned}$$

The correlation for small times t is thus controlled by the Lower tail of the Fourier spectrum of f_R. The great advantage of the breakthrough approach in [SS10b] is that their technique is particularly well suited to the study of the Lower tail of $\hat{f}_R$. Indeed theorem 4.2 says that if one can find a markovian algorithm computing f_R with small revealment $\boldsymbol{\delta}$, then for any $l \geq 1$, one has

$$\sum_{k=1}^{l} \sum_{|S|=k} \hat{f}_R(S)^2 \leq l^2\, \boldsymbol{\delta} \|f_R\|_2^2 = l^2\, \boldsymbol{\delta}\, \alpha_1(R)\,,$$

since by definition of f_R, $\|f_R\|_2^2 = \mathbb{P}\big[0 \overset{\omega}{\longleftrightarrow} R\big] = \alpha_1(R)$ (which in the triangular lattice is of order $R^{-5/48}$). What remains to be done is to find an algorithm minimizing the revealment as much as possible. But there is a difficulty here, similar as the one encountered in the proof of the sub-Gaussian fluctuations of FPP in $\mathbb{Z}^2$: any algorithm computing f_R will have to examine sites around the origin 0 making its revealment close to one. This is not only a technical difficulty here, the deeper reason comes from the following interpretation of the correlation $\mathbb{P}\big[0 \overset{\omega_0}{\longleftrightarrow} R,\, 0 \overset{\omega_t}{\longleftrightarrow} R\big]$: this correlation can be seen as $\mathbb{P}\big[0 \overset{\omega_0}{\longleftrightarrow} R\big]\, \mathbb{P}\big[0 \overset{\omega_t}{\longleftrightarrow} R \;\big|\; 0 \overset{\omega_0}{\longleftrightarrow} R\big]$. Now, assume R is very large and t very small; if one conditions on the event $\{0 \overset{\omega_0}{\longleftrightarrow} R\}$, since few sites are updated, the open path in ω_0 from 0 to distance R will still be preserved in ω_t at least up to some distance $L(t)$ while further away, large scale connections start being "noise sensitive". In some sense the geometry associated to the event $\{0 \overset{\omega}{\longleftrightarrow} R\}$ is "frozen" on a certain scale between time 0 and time t. Therefore, it is natural to divide our correlation analysis into two scales: the ball of radius $r = r(t)$ and the annulus from $r(t)$ to R (we might need to take $r(t) \gg L(t)$ since the control on the Fourier tail given by the algorithmic approach is not sharp). Obviously the "frozen radius" $r = r(t)$ increases as $t \to 0$. As in the exceptional planes case, let us then bound our correlation by

$$\begin{aligned}
\mathbb{P}\big[0 \overset{\omega_0}{\longleftrightarrow} R,\, 0 \overset{\omega_t}{\longleftrightarrow} R\big] &\leq \mathbb{P}\big[0 \overset{\omega_0}{\longleftrightarrow} r\big]\, \mathbb{P}\big[r \overset{\omega_0}{\longleftrightarrow} R,\, r \overset{\omega_t}{\longleftrightarrow} R\big] \\
&\leq \alpha_1(r)\, \mathbb{E}\big[f_{r,R}(\omega_0) f_{r,R}(\omega_t)\big], \qquad (6.3)
\end{aligned}$$

where $f_{r,R}$ is the indicator function of the event $r \overset{\omega}{\longleftrightarrow} R$. Now, as above

$$\begin{aligned}\mathbb{E}\big[f_{r,R}(\omega_0)f_{r,R}(\omega_t)\big] &\lesssim \mathbb{E}\big[f_{r,R}\big]^2 + O(1)\sum_{k=1}^{1/t}\sum_{|S|=k}\hat{f}_{r,R}(S)^2 \\ &\lesssim \alpha_1(r,R)^2 + O(1)\,t^{-2}\,\boldsymbol{\delta}_{f_{r,R}}\;\alpha_1(r,R)\,.\end{aligned}$$

The Boolean function $f_{r,R}$ somehow avoids the singularity at the origin, and it is possible to find algorithms for this function with small revealments. It is an interesting exercise to adapt the exploration algorithm which was used above in the Left-rigth crossing case to our current radial case. We will sketch at the end of this subsection, why using a natural exploration-type algorithm, one can obtain a revealment for $f_{r,R}$ of order $\boldsymbol{\delta} \asymp \alpha_2(r)\alpha_1(r,R)$. Assuming this, one ends up with

$$\begin{aligned}\mathbb{P}\big[0 \overset{\omega_0}{\longleftrightarrow} R,\, 0 \overset{\omega_t}{\longleftrightarrow} R\big] &\lesssim \alpha_1(r)\left(\alpha_1(r,R)^2 + t^{-2}\alpha_2(r)\alpha_1(r,R)^2\right) \\ &\lesssim \left(\alpha_1(r)^{-1} + t^{-2}\frac{\alpha_2(r)}{\alpha_1(r)}\right)\alpha_1(R)^2\,,\end{aligned}$$

using the quasi-multiplicativity property 2.5. Now using the knowledge of the critical exponents on $\mathbb{T}$ without being rigorous (these exponents are known only up to logarithmic corrections), one gets

$$\mathbb{P}\big[0 \overset{\omega_0}{\longleftrightarrow} R,\, 0 \overset{\omega_t}{\longleftrightarrow} R\big] \lesssim (1 + t^{-2}r^{-1/4})r^{5/48}\alpha_1(R)^2\,,$$

Optimizing in $r = r(t)$, one chooses $r(t) := t^{-8}$ which implies the following bound on the correlation:

$$\mathbb{P}\big[0 \overset{\omega_0}{\longleftrightarrow} R,\, 0 \overset{\omega_t}{\longleftrightarrow} R\big] \lesssim t^{-5/6}\alpha_1(R)^2\,. \tag{6.4}$$

Now, since $\int_0^1 t^{-5/6}dt < \infty$, by integrating the above correlation over the unit interval, one has from (6.1) that $\mathbb{E}\big[X_R^2\big] \leq C\mathbb{E}\big[X_R\big]^2$ for some universal $C > 0$ (since, by definition of X_R, $\mathbb{E}\big[X_R\big] = \alpha_1(R)$). This thus proves the existence of exceptional times.

If one had obtained a much weaker control on the correlation than that in (6.4), for example a bound of $t^{-1}\log(t)^{-2}\alpha_1(R)^2$, this would have still implied the existence of exceptional times: one can thus exploit more the good estimate provided by equation (6.4). This estimate in fact easily implies the second part of theorem 6.2, i.e. that almost surely the Hausdorff dimension of the set of exceptional times is greater than 1/6 (the upper bound of 31/36 is rather easy to obtain, it is a first moment analysis and follows from the behavior of the density function θ near $p_c = 1/2$, see [SS10b] and [Ste09]). The proof of the lower bound on the Hausdorff dimension which is based on the estimate (6.4), is classical and essentially consists into defining a (random) Frostman measure on the set $\mathcal{E}$ of exceptional times; one concludes by bounding its expected α-energies for all $\alpha < 1/6$.

See [SS10b] for rigorous proofs (taking care of the logarithmic corrections etc..).

Let us conclude this subsection by briefly explaining why one can achieve a revealment of order $\alpha_2(r)\alpha_1(r,R)$ for the Boolean function $f_{r,R}(\omega) := 1_{\{r \overset{\omega}{\longleftrightarrow} R\}}$. We

use an algorithm that mimics the one we used in the "chordal" case except the present setup is "radial". As in the chordal case, we randomize the starting point of our exploration process: let's start from a site taken uniformly on the 'circle' of radius R. Then, let's explore the picture with an exploration path γ directed towards the origin; this means that as in the chordal case, when the interface encounters an open (resp closed) site, it turns say on the right (resp left), the only difference being that when the exploration path closes a loop around the origin, it continues its exploration inside the connected component of the origin (see [Wer07] for more details on the radial exploration path). It is known that this discrete curve converges towards **radial** SLE_6 on $\mathbb{T}$, when the mesh goes to zero. It turns out that the so-defined exploration path gives all the information we need. Indeed, if the exploration path closes a clockwise loop around the origin, this means that there is a closed circuit around the origin making $f_{r,R}$ equal to zero. On the other hand, if the exploration path does not close any clockwise loop until it reaches radius r, it means that $f_{r,R} = 1$. Hence, we run the exploration path until either it closes a clockwise loop or it reaches radius r. Now, what is the revealment for this algorithm ? Neglecting boundary issues (points near radius r or R), if x is a point at distance u from 0, with $2r < u < R/2$, in order for x to be examined by the algorithm, it is needed that the exploration path did not close any clockwise loop before radius u; hence there is an open path from u to R, this already costs a factor $\alpha_1(u, R)$. Now at the level of the scale u, what is the probability that x lies on the interface ? The interface is asymptotically of dimension $7/4$, so in the annulus $A(u/2, 2u)$, about $u^{7/4+o(1)}$ points lie on γ and the probability that the point x is in γ is of order $u^{-1/4}$. In fact it is exactly up to constant $\alpha_2(u)$. Hence, the probability $\mathbb{P}\big[x \in J\big]$, for $|x| = u$ is of order $\alpha_2(u)\alpha_1(u, R)$. Now recall that the revealment is the supremum over all sites of this probability; it is easy to see that the smaller the scale (r), the higher this probability is; therefore neglecting boundary issues, one obtains that $\boldsymbol{\delta} \approx \alpha_2(r)\alpha_1(r, R)$.

6.3.2 Consequences of the Geometric approach ([GPS10])

We now present the applications to dynamical percolation of the Sharp estimates on the Fourier spectrum obtained in [GPS10]. Since the results in [GPS10] are sharp on the triangular lattice $\mathbb{T}$ as well as on $\mathbb{Z}^2$, the first notable consequence is the following result

Theorem 6.3 ([GPS10])**.** *If $(\omega_t)_{t\geq 0}$ denotes some trajectory of dynamical percolation on the square grid $\mathbb{Z}^2$ at $p_c = 1/2$, then almost surely, there exist exceptional times $t \in [0, \infty)$, such that ω_t has an infinite cluster.*

Furthermore there is some $\alpha > 0$, such that if $\mathcal{E} \subset [0, \infty)$ denotes the random set of these exceptional times, then almost surely, the Hausdorff dimension of $\mathcal{E}$ is greater than α.

Remark 6.4. The paper [SS10b] came quite close to proving that exceptional times exist for dynamical critical bond percolation on $\mathbb{Z}^2$, but there is still a missing gap if one wants to use randomized algorithm techniques in this case.

On the triangular lattice $\mathbb{T}$, since the critical exponents are known, the sharp control on $\hat{f}_R$ in [GPS10] and especially on its **Lower-tail** enables one to obtain

detailed information on the structure of the (random) set of exceptional times $\mathcal{E}$. One can prove for instance

Theorem 6.5 ([GPS10]). *If $\mathcal{E} \subset [0,\infty)$ denotes the (random) set of exceptional times of dynamical percolation on $\mathbb{T}$, then almost surely the Hausdorff dimension of $\mathcal{E}$ equals $\frac{31}{36}$.*

This theorem strengthens theorem 6.2 from [SS10b]. Finally, one obtains the following interesting phenomenon

Theorem 6.6 ([GPS10]). *Almost surely on $\mathbb{T}$, there exist exceptional times "of the second kind" $t \in [0,\infty)$, for which an infinite (open) cluster* **coexists** *in ω_t with an infinite dual (closed) cluster.*

Furthermore, if $\mathcal{E}^{(2)}$ denotes the set of these exceptional times; then almost surely, the Hausdorff dimension of $\mathcal{E}^{(2)}$ is greater than $1/9$.

We conjectured in [GPS10] that $\mathcal{E}^{(2)}$ should be almost surely of dimension $2/3$; but unfortunately, the methods in [GPS10] do not apply well to non-monotone functions, in particular in this case, to the indicator function of a two-arm event.

Let us now focus on theorem 6.5. As in in the previous subsection, in order to have detailed information on the dimension of the exceptional set $\mathcal{E}$, one needs to obtain a sharp control on the correlations $\mathbb{P}\big[0 \overset{\omega_0}{\longleftrightarrow} R,\, 0 \overset{\omega_t}{\longleftrightarrow} R\big]$. More exactly, the almost sure dimension $31/36$ would follow from the following estimate

$$\mathbb{P}\big[0 \overset{\omega_0}{\longleftrightarrow} R,\, 0 \overset{\omega_t}{\longleftrightarrow} R\big] \lesssim t^{-5/36}\, \alpha_1(R)^2\,, \tag{6.5}$$

which gives a more precise control on the correlations than the one given by (6.4) obtained in [SS10b] (which implied a lower bound on the dimension of $\mathcal{E}$ equal to $1/6$). As previously, proving such an estimate will be achieved through a sharp understanding of the Fourier Spectrum of $f_R(\omega) := 1_{0 \overset{\omega}{\longleftrightarrow} R}$ (but this time, in order to get a sharp control, we will rely on the "geometric approach" explained in section 5). Indeed, recall from (6.2) that

$$\mathbb{P}\big[0 \overset{\omega_0}{\longleftrightarrow} R,\, 0 \overset{\omega_t}{\longleftrightarrow} R\big] \lesssim \hat{\mathbb{Q}}\big[0 < |\mathscr{S}_{f_R}| < 1/t\big]\,. \tag{6.6}$$

Before considering the precise behavior of the Lower-tail of $\hat{\mathbb{Q}}_{f_R}$, let us detail a heuristic argument which explains why one should expect the above control on the correlations (estimate 6.5). Recall ω_t can be seen as a noised configuration of ω_0. What we learned from section 5 is that in the Left-Right case, once the noise ϵ is high enough so that "many pivotal points" are touched, the system starts being *noise sensitive*. This means that the Left to Right crossing event under consideration for ω^ϵ starts being independent of the whole configuration ω. On the other hand (and this side is much easier), if the noise is such that pivotal points are (with high probability) not flipped, then the system is *stable*. Applied to our radial setting, if R is some large radius and r some intermediate scale then, conditioned on the event $\{0 \overset{\omega}{\longleftrightarrow} R\}$, it is easy to check that on average, there are about $r^{3/4}$ pivotal points at scale r (say, in the annulus $A(r,2r)$). Hence, if the level of noise t is such

that $tr^{3/4} \ll 1$, then the radial event is "stable" below scale r. On the other hand if $tr^{3/4} \gg 1$, then many pivotal points will be touched and the radial event should be sensitive above scale r . As such, there is some limiting scale $L(t)$ separating a frozen phase (where radial crossings are stable) from a sensitive phase. The above argument gives $L(t) \approx t^{-4/3}$; this limiting scale is often called the **characteristic length**: below that scale, one does not feel the effects of the dynamic from ω_0 to ω_t, while above that scale, connectivity properties start being uncorrelated between ω_0 and ω_t.

Remark 6.7. Note that in our context of radial events, one has to be careful with the notion of being *noise sensitive* here. Indeed our sequence of Boolean functions $(f_R)_{R>1}$ satisfy $\mathrm{Var}(f_R) \to 0$; therefore, the sequence trivially satisfies our initial definition of being noise sensitive (definition 1.2). For such a sequence of Boolean functions corresponding to events of probability going to zero, what we mean by being noise sensitive here is rather that the correlations satisfy $\mathbb{E}\big[f(\omega)f(\omega^\epsilon)\big] \leq O(1)\mathbb{E}\big[f\big]^2$. This changes the intuition by quite a lot: for example, previously, for a sequence of functions with L^2-norm one, analyzing the sensitivity was equivalent to localizing where most of the Fourier mass was; while if $\|f_R\|_2$ goes to zero, and if one wishes to understand for which levels of noise $\epsilon_R > 0$, one has $\mathbb{E}\big[f_R(\omega)f_R(\omega^{\epsilon_R})\big] \leq O(1)\mathbb{E}\big[f_R\big]^2$, then it is the size of the Lower tail of $\hat{f}_R$ which is now relevant.

To summarize and conclude our heuristical argument, as in the above subsection, one can bound our correlations as follows

$$\mathbb{P}\big[0 \overset{\omega_0}{\longleftrightarrow} R,\, 0 \overset{\omega_t}{\longleftrightarrow} R\big] \leq \mathbb{P}\big[0 \overset{\omega_0}{\longleftrightarrow} L(t)\big]\, \mathbb{P}\big[L(t) \overset{\omega_0}{\longleftrightarrow} R,\, L(t) \overset{\omega_t}{\longleftrightarrow} R\big]\,.$$

Not much is lost in this bound since connections are stable below the correlation length $L(t)$. Now, above the correlation length, the system starts being noise sensitive, hence one expects $\mathbb{P}\big[L(t) \overset{\omega_0}{\longleftrightarrow} R,\, L(t) \overset{\omega_t}{\longleftrightarrow} R\big] \leq O(1)\mathbb{P}\big[L(t) \overset{\omega}{\longleftrightarrow} R\big]^2 = O(1)\alpha_1(L(t), R)^2$. Using quasi-multiplicativity and the values of the critical exponents, one ends up with

$$\begin{aligned}\mathbb{P}\big[0 \overset{\omega_0}{\longleftrightarrow} R,\, 0 \overset{\omega_t}{\longleftrightarrow} R\big] &\leq O(1)\alpha_1(L(t))^{-1}\alpha_1(R)^2 \\ &\lesssim t^{-5/36}\alpha_1(R)^2\,,\end{aligned}$$

as desired.

From now on, we wish to briefly explain how to adapt the geometric approach of section 5 to our present radial setting. Recall that the goal is to prove an analog of theorem 5.12 for the radial crossing event $f_R(\omega) = 1_{0 \overset{\omega}{\longleftrightarrow} R}$. Before stating such an analogous statement (which would seem at first artificial),we need to understand some properties of $\mathscr{S}_{f_R}$.

First of all, similarly as in section 5, it is easy to check that $\hat{\mathbb{P}}\big[|\mathscr{S}_{f_R}|\big] \asymp R^2\alpha_4(R)$ (recall that $\hat{\mathbb{P}}_{f_R}$ is the renormalized Spectral measure $\frac{1}{\|f_R\|^2}\hat{\mathbb{Q}}_{f_R} = \frac{1}{\alpha_1(R)}\hat{\mathbb{Q}}_{f_R}$). The second moment is also easy to compute. One can conclude from these estimates that with positive probability (under $\hat{\mathbb{P}}_{f_R}$), $|\mathscr{S}_{f_R}|$ is of order $R^2\alpha_4(R)$ ($\approx R^{3/4}$ on $\mathbb{T}$).

Now, following the strategy explained in section 5, we divide the ball of radius R into a grid of mesoscopic squares of radii r (with $1 < r < R$). It is easy to check that for any such r-square Q, one has

$$\hat{\mathbb{P}}_{f_R}\big[|\mathscr{S} \cap Q| \;\big|\; \mathscr{S} \cap Q \neq \emptyset\big] \asymp r^2 \alpha_4(r)\,.$$

Furthermore, and this part is very similar as in the chordal case, one can obtain a "weak" control on the dependencies within $\mathscr{S}_{f_R}$, namely that for any r-square Q and any subset W such that $Q \cap W = \emptyset$, one has

$$\hat{\mathbb{P}}_{f_R}\big[|\mathscr{S} \cap \bar{Q}| \geq r^2 \alpha_4(r) \;\big|\; \mathscr{S} \cap Q \neq \emptyset,\, \mathscr{S} \cap W = \emptyset\big] \geq c\,,$$

where $\bar{Q}$ is the concentric square in Q of side-length $r/2$ and $c > 0$ is some absolute constant.

As in section 5, if $S \subset [-R, R]^2$, we denote by $S_{(r)}$ the set of those $r \times r$ squares which intersect S. Thanks to the above weak independence statement, it remains (see section 5) to study the lower-tail of the number of mesoscopic squares touched by the Spectral sample, i.e. the lower-tail of $|\mathscr{S}_{(r)}|$ (only the very bottom-part of this distribution needs to be understood in a sharp way). Hence, one has to understand how $\mathscr{S}_{(r)}$ typically looks when it is conditioned to be of very small size (say of size less than $\log R/r$, i.e. much smaller than the average size of $\mathscr{S}_{(r)}$ which is of order $(R/r)^{3/4}$). The intuition differs here from the chordal case (analyzed in section 5); indeed recall that in the chordal case, if $|\mathscr{S}_{(r)}|$ was conditioned to be of small size, then it was typically "concentrated in space" somewhere "randomly" in the square. In our radial setting, if $|\mathscr{S}_{(r)}|$ is conditioned to be of small size, it also tends to be localized, but instead of being localized somewhere randomly through the disk of radius R, it is localized around the origin. The reason of this localization around the origin comes from the following estimate: if Q_0 denotes the $r-$square surrounding the origin

$$\begin{aligned}
\hat{\mathbb{P}}\big[Q_0 \notin \mathscr{S}_{(r)}\big] &\leq \hat{\mathbb{P}}\big[\mathscr{S}_{f_R} \subset B(0, \frac{r}{2})^c\big] \\
&= \frac{1}{\alpha_1(R)} \hat{\mathbb{Q}}\big[\mathscr{S}_{f_R} \subset B(0, \frac{r}{2})^c\big] \\
&= \frac{1}{\alpha_1(R)} \mathbb{E}\big[\mathbb{E}\big[f_R \;\big|\; B(0,R) \setminus B(0,r)\big]^2\big] \;\text{ By lemma 5.3} \\
&\leq \frac{O(1)}{\alpha_1(R)}\, \alpha_1(r,R)\alpha_1(r)^2 \leq O(1)\, \alpha_1(r)
\end{aligned}$$

(In the last line, we used the fact that $\mathbb{E}\big[f_R \;\big|\; B(0,R) \setminus B(0,r)\big] \leq \alpha_1(r)\, 1_{r \leftrightarrow R}$). Since $\alpha_1(r)$ goes to 0 as $r \to \infty$, this means that for a mesoscopic scale r such that $1 \ll r \ll R$, $\mathscr{S}_{(r)}$ "does not like" to avoid the origin. In some sense, this shows that the origin is an attractive point for $\mathscr{S}_{f_R}$.

Still, this does not yet imply that the origin will be attractive also for $\mathscr{S}_{(r)}$ conditioned to be of small size. If one assumes that, as in the chordal case, $\mathscr{S}_{(r)}$ tends to be localized when it is conditioned to be of small size (i.e. that for k small, $\hat{\mathbb{P}}\big[|\mathscr{S}_{(r)}| = k\big] \approx \hat{\mathbb{P}}\big[|\mathscr{S}_{(r)}| = 1\big]$), then it not hard to see that it has to be

localized around the origin: indeed, one can estimate by hand that once conditioned on $|\mathscr{S}_{(r)}| = 1$, $\mathscr{S}_{(r)}$ will be (with high conditional probability) close to the origin. To summarize, assuming that $\mathscr{S}_{(r)}$ tends to "clusterize" when it is conditioned to be of small size, one has for k "small"

$$\begin{aligned}
\hat{\mathbb{Q}}\big[|\mathscr{S}_{(r)}| = k\big] &\approx \hat{\mathbb{Q}}\big[|\mathscr{S}_{(r)}| = 1\big] \asymp \hat{\mathbb{Q}}\big[\mathscr{S}_{(r)} = Q_0\big] \\
&\asymp \hat{\mathbb{Q}}\big[\emptyset \neq \mathscr{S}_{f_R} \subset B(0,r)\big] \\
&\asymp \mathbb{E}\big[\mathbb{E}\big[f_R \mid B(0,r)\big]^2\big] - \hat{\mathbb{Q}}\big[\mathscr{S}_{f_R} = \emptyset\big] \\
&\asymp O(1)\alpha_1(r)\alpha_1(r,R)^2 - \alpha_1(R)^2 \\
&\asymp \frac{1}{\alpha_1(r)}\alpha_1(R)^2 \text{ using quasi-multiplicativity}\,.
\end{aligned}$$

Using an inductive proof (similar as the one used in the chordal case, but where the origin plays a special role), one can prove that it is indeed the case that $\mathscr{S}_{(r)}$ tends to be localized when it is conditioned to be of small size. More quantitatively, the inductive proof leads to the following estimate on the lower-tail of $|\mathscr{S}_{(r)}|$

Proposition 6.8 ([GPS10], section 4)**.** *There is a sub-exponentially fast growing function $g(k)$, $k \geq 1$, such that for any $1 \leq r \leq n$*

$$\hat{\mathbb{Q}}_{f_R}\big[|\mathscr{S}_{(r)}| = k\big] \leq g(k)\,\frac{1}{\alpha_1(r)}\alpha_1(R)^2\,. \tag{6.7}$$

Using the same technology as in section 5 (i.e. the weak independence control, the above proposition on the lower tail of the mesoscopic sizes and a large deviation lemma which helps implementing the "scanning procedure"), one obtains that $\hat{\mathbb{Q}}\big[0 < |\mathscr{S}_{f_R}| < r^2\alpha_4(r)\big] \asymp \hat{\mathbb{Q}}_{f_R}\big[|\mathscr{S}_{(r)}| = 1\big] \asymp \frac{1}{\alpha_1(r)}\alpha_1(R)^2$, which is exactly a sharp lower-tail estimate on $\hat{f}_R$. More exactly one has the following theorem (analog of theorem 5.12 in the chordal case)

Theorem 6.9 ([GPS10])**.** *Both on the triangular lattice $\mathbb{T}$ and on $\mathbb{Z}^2$, if f_R denotes the radial event up to radius R, one has*

$$\hat{\mathbb{Q}}\big[|\mathscr{S}_{f_R}| < r^2\alpha_4(r)\big] \asymp \frac{1}{\alpha_1(r)}\alpha_1(R)^2\,,$$

where the constants involved in $\asymp$ are absolute constants.

On the triangular lattice $\mathbb{T}$, using the knowledge of the critical exponents, this can be read as follows

$$\hat{\mathbb{Q}}\big[|\mathscr{S}_{f_R}| < r^{3/4}\big] \approx r^{5/48}\alpha_1(R)^2\,,$$

or equivalently for any $u \lesssim R^{3/4}$,

$$\hat{\mathbb{Q}}\big[|\mathscr{S}_{f_R}| < u\big] \approx u^{5/36}\alpha_1(R)^2\,.$$

Recall our goal was to obtain a sharp bound on the correlation between $f_R(\omega_0)$ and $f_R(\omega_t)$. From our control on the lower-tail of $\hat{f}_R$ and equation (6.6) one obtains

$$\begin{aligned}\mathbb{P}\big[0 \overset{\omega_0}{\longleftrightarrow} R,\, 0 \overset{\omega_t}{\longleftrightarrow} R\big] &\lesssim \hat{\mathbb{Q}}\big[0 < |\mathscr{S}_{f_R}| < 1/t\big] \\ &\lesssim t^{-\frac{5}{36}}\, \alpha_1(R)^2\,,\end{aligned}$$

which, as desired, implies that the exceptional set $\mathcal{E}$ is indeed a.s. of dimension $\frac{31}{36}$ (as claimed above, the upper bound is much easier using the knowledge on the density function $\theta_{\mathbb{T}}$).

References

[BR08] Michel Benaïm and Raphaël Rossignol. Exponential concentration for first passage percolation through modified Poincaré inequalities. *Ann. Inst. Henri Poincaré Probab. Stat.*, 44(3):544–573, 2008.

[BKS99] Itai Benjamini, Gil Kalai, and Oded Schramm. Noise sensitivity of Boolean functions and applications to percolation. *Inst. Hautes Études Sci. Publ. Math.*, (90):5–43 (2001), 1999.

[BKS03] Itai Benjamini, Gil Kalai, and Oded Schramm. First passage percolation has sublinear distance variance. *Ann. Probab.*, 31(4):1970–1978, 2003.

[BS98a] Itai Benjamini and Oded Schramm. Conformal invariance of Voronoi percolation. *Comm. Math. Phys.*, 197(1):75–107, 1998.

[BS98b] Itai Benjamini and Oded Schramm. Exceptional planes of percolation. *Probab. Theory Related Fields*, 111(4):551–564, 1998.

[BSW05] Itai Benjamini, Oded Schramm, and David B. Wilson. Balanced Boolean functions that can be evaluated so that every input bit is unlikely to be read. In *STOC'05: Proceedings of the 37th Annual ACM Symposium on Theory of Computing*, pages 244–250. ACM, New York, 2005.

[BOL87] M. Ben-Or and N. Linial. Collective coin flipping. *Randomness and Computation (S. Micali, ed.)*, 1987.

[BKK+92] Jean Bourgain, Jeff Kahn, Gil Kalai, Yitzhak Katznelson, and Nathan Linial. The influence of variables in product spaces. *Israel J. Math.*, 77(1-2):55–64, 1992.

[Car92] John Cardy. Critical percolation in finite geometries. *J. Phys. A*, 25(4):L201–L206, 1992.

[Cha08] Sourav Chatterjee. Chaos, Concentration, and Multiple Valleys. preprint, 2008.

[FK96] Ehud Friedgut and Gil Kalai. Every monotone graph property has a sharp threshold. *Proc. Amer. Math. Soc.*, 124(10):2993–3002, 1996.

[GPS10] Christophe Garban, Gábor Pete, and Oded Schramm. The Fourier spectrum of critical percolation. *Acta Math.*, to appear, 2010.

[Gri99] Geoffrey Grimmett. *Percolation*. Grundlehren der mathematischen Wissenschaften 321. Springer-Verlag, Berlin, second edition, 1999.

[Gri08] Geoffrey Grimmett. *Probability On Graphs*. 2008.

[Gro75] Leonard Gross. Logarithmic Sobolev inequalities. *Amer. J. Math.*, 97(4):1061–1083, 1975.

[HPS97] Olle Häggström, Yuval Peres, and Jeffrey E. Steif. Dynamical percolation. *Ann. Inst. H. Poincaré Probab. Statist.*, 33(4):497–528, 1997.

[HS90] Takashi Hara and Gordon Slade. Mean-field critical behaviour for percolation in high dimensions. *Comm. Math. Phys.*, 128(2):333–391, 1990.

[Joh00] Kurt Johansson. Shape fluctuations and random matrices. *Comm. Math. Phys.*, 209(2):437–476, 2000.

[KKL88] Jeff Kahn, Gil Kalai, and Nathan Linial. The influence of variables on boolean functions. *29th Annual Symposium on Foundations of Computer Science*, (68–80), 1988.

[KS06] Gil Kalai and Shmuel Safra. Threshold phenomena and influence: perspectives from mathematics, computer science, and economics. In *Computational complexity and statistical physics*, St. Fe Inst. Stud. Sci. Complex., pages 25–60. Oxford Univ. Press, New York, 2006.

[Kes87] Harry Kesten. Scaling relations for 2D-percolation. *Comm. Math. Phys.*, 109(1):109–156, 1987.

[KZ87] Harry Kesten and Yu Zhang. Strict inequalities for some critical exponents in two-dimensional percolation. *J. Statist. Phys.*, 46(5-6):1031–1055, 1987.

[LPSA94] Robert Langlands, Philippe Pouliot, and Yvan Saint-Aubin. Conformal invariance in two-dimensional percolation. *Bull. Amer. Math. Soc. (N.S.)*, 30(1):1–61, 1994.

[LSW02] Gregory F. Lawler, Oded Schramm, and Wendelin Werner. One-arm exponent for critical 2D percolation. *Electron. J. Probab.*, 7:no. 2, 13 pp. (electronic), 2002.

[Mar74] G. A. Margulis. Probabilistic characteristics of graphs with large connectivity. *Problemy Peredači Informacii*, 10(2):101–108, 1974.

[MR96] Ronald Meester and Rahul Roy. *Continuum percolation*, volume 119 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1996.

[MOO05] Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. *Ann. Math.*, to appear, 2005.

[Nel66] Edward Nelson. A quartic interaction in two dimensions. In *Mathematical Theory of Elementary Particles (Proc. Conf., Dedham, Mass., 1965)*, pages 69–73. M.I.T. Press, Cambridge, Mass., 1966.

[Nol09] Pierre Nolin. Near-critical percolation in two dimensions. *Electron. J. Probab.*, 13(55):1562—1623, 2009.

[O'D] Ryan O'Donnell. History of the hypercontractivity theorem. *http://boolean-analysis.blogspot.com/*.

[O'D03] Ryan O'Donnell. *Computational Applications Of Noise Sensitivity*. PhD thesis, M.I.T., 2003.

[OS07] Ryan O'Donnell and Rocco A. Servedio. Learning monotone decision trees in polynomial time. *SIAM J. Comput.*, 37(3):827–844 (electronic), 2007.

[OSSS05] Ryan O'Donnell, Michael Saks, Oded Schramm, and Rocco Servedio. Every decision tree has an influential variable. *FOCS*, 2005.

[PSSW07] Yuval Peres, Oded Schramm, Scott Sheffield, and David B. Wilson. Random-turn hex and other selection games. *Amer. Math. Monthly*, 114(5):373–387, 2007.

[Rus78] Lucio Russo. A note on percolation. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 43(1):39–48, 1978.

[Rus82] Lucio Russo. An approximate zero-one law. *Z. Wahrsch. Verw. Gebiete*, 61(1):129–139, 1982.

[Sch00] Oded Schramm. Scaling limits of loop-erased random walks and uniform spanning trees. *Israel J. Math.*, 118:221–288, 2000.

[Sch07] Oded Schramm. Conformally invariant scaling limits: an overview and a collection of problems. In *International Congress of Mathematicians. Vol. I*, pages 513–543. Eur. Math. Soc., Zürich, 2007.

[SS10a] Oded Schramm and Stanislav Smirnov. On the scaling limits of planar percolation. *Ann. Probab.*, to appear, 2010.

[SS10b] Oded Schramm and Jeffrey Steif. Quantitative noise sensitivity and exceptional times for percolation. *Ann. Math.*, 171(2):619–672, 2010.

[Smi01] Stanislav Smirnov. Critical percolation in the plane: conformal invariance, Cardy's formula, scaling limits. *C. R. Acad. Sci. Paris Sér. I Math.*, 333(3):239–244, 2001.

[SW01] Stanislav Smirnov and Wendelin Werner. Critical exponents for two-dimensional percolation. *Math. Res. Lett.*, 8(5-6):729–744, 2001.

[Ste09] Jeffrey Steif. A survey of dynamical percolation. *Fractal geometry and stochastics, IV, Birkhauser*, pages 145–174, 2009.

[Tal94] Michel Talagrand. On Russo's approximate zero-one law. *Ann. Probab.*, 22(3):1576–1587, 1994.

[Tal95] Michel Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Inst. Hautes Études Sci. Publ. Math.*, (81):73–205, 1995.

[Tal96] Michel Talagrand. How much are increasing sets positively correlated? *Combinatorica*, 16(2):243–258, 1996.

[Tsi99] Boris Tsirelson. Fourier-Walsh coefficients for a coalescing flow (discrete time). 1999.

[Tsi04] Boris Tsirelson. Scaling limit, noise, stability. In *Lectures on probability theory and statistics*, volume 1840 of *Lecture Notes in Math.*, pages 1–106. Springer, Berlin, 2004.

[TV98] B. S. Tsirelson and A. M. Vershik. Examples of nonlinear continuous tensor products of measure spaces and non-Fock factorizations. *Rev. Math. Phys.*, 10(1):81–145, 1998.

[Wer07] Wendelin Werner. *Lectures on two-dimensional critical percolation.* IAS Park City Graduate Summer School, 2007.

NOISE SENSITIVITY OF BOOLEAN FUNCTIONS AND APPLICATIONS TO PERCOLATION

by ITAI BENJAMINI, GIL KALAI, ODED SCHRAMM

ABSTRACT

It is shown that a large class of events in a product probability space are highly sensitive to noise, in the sense that with high probability, the configuration with an arbitrary small percent of random errors gives almost no prediction whether the event occurs. On the other hand, weighted majority functions are shown to be noise-stable. Several necessary and sufficient conditions for noise sensitivity and stability are given.

Consider, for example, bond percolation on an $n + 1$ by n grid. A configuration is a function that assigns to every edge the value 0 or 1. Let ω be a random configuration, selected according to the uniform measure. A crossing is a path that joins the left and right sides of the rectangle, and consists entirely of edges e with $\omega(e) = 1$. By duality, the probability for having a crossing is 1/2. Fix an $\varepsilon \in (0, 1)$. For each edge e, let $\omega'(e) = \omega(e)$ with probability $1 - \varepsilon$, and $\omega'(e) = 1 - \omega(e)$ with probability ε, independently of the other edges. Let $p(\tau)$ be the probability for having a crossing in ω, conditioned on $\omega' = \tau$. Then for all n sufficiently large, $\mathbf{P}\{\tau : |p(\tau) - 1/2| > \varepsilon\} < \varepsilon$.

CONTENTS

1. Introduction 6
   1.1. Noise sensitivity -- three examples 6
   1.2. Influences of variables 7
   1.3. Weighted majority 9
   1.4. Stability 10
   1.5. Fourier-Walsh expansion 10
   1.6. Some related and future work 11
   1.7. The structure of this paper 12
2. Sensitivity to noise 13
3. Correlation with majority 20
   3.1. Uniform weights 20
   3.2. General weights 23
4. An application to percolation 28
5. Some conjectures and problems concerning percolation 31
   5.1. Other sensitivity conjectures 31
   5.2. Stronger sensitivity conjectures 32
   5.3. Dynamical percolation 32
   5.4. Limits and conformal invariance 33
   5.5. Fourier-Walsh coefficients of percolation 34
   5.6. Other models of statistical mechanics 35
6. Some further examples 35
   6.1. Tribes 35
   6.2. Recursive majority on the ternary tree 35
   6.3. Number of runs 36
   6.4. Majority of triangles 36
7. Relations with complexity theory 36
   7.1. AC0 and influences 36
   7.2. TC0 and noise sensitivity 37
8. Random walks 38
9. Changing a fixed size set of bits 39

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_12, © Springer Science+Business Media, LLC 2011

1. Introduction

1.1. *Noise sensitivity – three examples*

Consider the Hamming cube $\Omega_n = \{0, 1\}^n$ endowed with the uniform probability measure $\mathbf{P}$. Let $\mathcal{A} \subset \Omega_n$ be some event. Given a random $x = (x_1, ..., x_n) \in \Omega_n$, suppose that $y = (y_1, ..., y_n)$ is a random perturbation of x; that is, for every $j \in \{1, ..., n\}$, $y_j = x_j$ with probability $1 - \varepsilon$, independently for distinct j's. Here $\varepsilon \in (0, 1)$ is some small fixed constant. This random perturbation of x will be denoted $N_\varepsilon(x)$. We may think of $N_\varepsilon(x)$ as x with some noise.

Based on the knowledge of $N_\varepsilon(x)$, we would like to predict the event $x \in \mathcal{A}$. Since the joint distribution $(x, N_\varepsilon(x))$ is the same as that of $(N_\varepsilon(x), x)$, an equivalent problem is to predict $N_\varepsilon(x) \in \mathcal{A}$ knowing x. The event $\mathcal{A}$ is **noise sensitive** if for all but a small set of x, knowing x does not significantly help in predicting the event $N_\varepsilon(x) \in \mathcal{A}$. More formally, $\mathcal{A}$ is noise sensitive, if for some small $\delta > 0$,

$$\gamma(\mathcal{A}, \varepsilon, \delta) := \mathbf{P}\left\{x : \left|\mathbf{P}(N_\varepsilon(x) \in \mathcal{A} \mid x) - \mathbf{P}(\mathcal{A})\right| > \delta\right\} < \delta. \tag{1.1}$$

Set

$$\phi(\mathcal{A}, \varepsilon) = \inf\left\{\delta > 0 : \gamma(\mathcal{A}, \varepsilon, \delta) < \delta\right\},$$

which is the infimum of all $\delta > 0$ such that (1.1) holds. This will be called the **sensitivity gauge** of $\mathcal{A}$. A sequence of events $\mathcal{A}_m \subset \Omega_{n_m}$ will be called **asymptotically noise sensitive** if

$$\lim_{m \to \infty} \phi(\mathcal{A}_m, \varepsilon) = 0, \qquad \forall \varepsilon \in (0, 1/2).$$

Remark **1.1.** — As shown in Section 2, $\mathcal{A}_m$ are asymptotically noise sensitive if and only if

$$\lim_{m \to \infty} \operatorname{var}[\mathbf{P}(N_\varepsilon(x) \in A_m | x)] = 0. \tag{1.2}$$

A simple example of a sequence of events which are not noise sensitive is dictatorship. The first bit dictator is the event $\mathcal{D}_n = \left\{(x_1, ..., x_n) \in \Omega_n : x_1 = 1\right\}$. To verify that $\{\mathcal{D}_n\}$ is not asymptotically noise sensitive, consider some event $\mathcal{A} \subset \Omega_n$. Then for $k > n$ we may obviously consider $\mathcal{A}$ as a subset of Ω_k, by ignoring the extra variables. Note that this does not change the value of $\phi(\mathcal{A}, \varepsilon)$. Consequently, $\phi(\mathcal{D}_n, \varepsilon) = \phi(\mathcal{D}_1, \varepsilon) \neq 0$ for all $n > 1$.

Let us examine now the example of majority. Pick some $\varepsilon \in (0, 1/2)$. Let $\mathcal{M}_n \subset \Omega_n$ denote the majority event, that is,

$$\mathcal{M}_n = \left\{ (x_1, ..., x_n) \in \Omega_n : \sum_j x_j \geq n/2 \right\}.$$

The probability that $\sum_j x_j - n/2 > \sqrt{n}$ is bounded from below as $n \to \infty$. Given such an x, the probability that $N_\varepsilon(x) \in \mathcal{M}_n$ is greater than $\mathbf{P}[\mathcal{M}_n] + \delta_1$ for some constant $\delta_1 > 0$, depending on ε. We conclude that majority is not asymptotically noise sensitive as $n \to \infty$.

Majority and dictatorship are not only noise insensitive, they are actually "noise stable", in a sense defined in Subsection 1.4 below.

It turns out that the noise insensitivity of majority and dictatorship is atypical, and many natural and interesting events are asymptotically noise sensitive.

Our third example is bond percolation on an $m+1$ by m rectangle in the ordinary square grid $\mathbf{Z}^2$. A configuration is an element in $\Omega = \{0, 1\}^E$, where E is the set of edges in this rectangle. Let $\omega \in \Omega$ be a random configuration, selected according to the uniform measure. A **crossing** is a path that joins the left and right sides of the rectangle, and consists entirely of edges e with $\omega(e) = 1$. Let $\mathcal{C}_m$ be the event that there is some crossing of this rectangle. By duality, it is not hard to see that $\mathbf{P}[\mathcal{C}_m] = 1/2$.

Theorem **1.2.** — *The crossing events $\mathcal{C}_m$ are asymptotically noise sensitive; that is, $\phi(\mathcal{C}_m, \varepsilon) \to 0$ as $m \to \infty$.*

This theorem will appear as a corollary of a general result. To introduce the more general statement, we need the notion of influence.

1.2. *Influences of variables*

Set $[n] = \{1, ..., n\}$. Given $x \in \Omega$ and $j \in [n]$, let $\sigma_j x = (x'_1, ..., x'_n)$, where $x'_k = x_k$ when $k \neq j$ and $x'_j = 1 - x_j$. The **influence** of the k-th variable on a function $f: \Omega \to \mathbf{R}$ is defined by

$$\mathbf{I}_k(f) = \| f(\sigma_k x) - f(x) \|_1 . \tag{1.3}$$

In other words, $\mathbf{I}_k(f)$ is the expected absolute value of the change in f when the k'th bit x_k is flipped. We shall often not distinguish between an event $\mathcal{A}$ and its indicator function $\chi_{\mathcal{A}}$. In particular, for events $\mathcal{A}$, $\mathbf{I}_k(\mathcal{A}) = \mathbf{I}_k(\chi_{\mathcal{A}})$. Note that $\mathbf{I}_k(\mathcal{A})$ is the probability that precisely one of the two elements x, $\sigma_k x$ is in $\mathcal{A}$.

This notion of influence was introduced by Ben-Or and Linial [4]. Kahn, Kalai and Linial [23] (see also [10, 31]) showed that for every $\mathcal{A} \subset \Omega_n$ with $\mathbf{P}[\mathcal{A}] = 1/2$ there is a $j \in [n]$ with $\mathbf{I}_j(\mathcal{A}) \geq c \log n/n$, for some constant $c > 0$, and that there always

exists a set $S \subset [n]$ with $|S| \leqslant c(\varepsilon)n/\log n$ whose cumulative influence is $> 1 - \varepsilon$; that is, the measure of the set of inputs for variables in $[n] - S$ which determine the value of f is less than ε.

Put

$$\mathbf{I}(f) = \sum_k \mathbf{I}_k(f),$$

$$\mathbf{H}(f) = \sum_k \mathbf{I}_k(f)^2.$$

Theorem **1.3.** — *Let $\mathscr{A}_m \subset \Omega_{n_m}$ be a sequence of events and suppose that $\mathbf{H}(\mathscr{A}_m) \to 0$ as $m \to \infty$. Then $\{\mathscr{A}_m\}$ is asymptotically noise sensitive.*

Equivalently, there is some continuous function Φ satisfying $\Phi(0, \varepsilon) = 0$ such that $\phi(\mathscr{A}, \varepsilon) \leqslant \Phi(\mathbf{H}(\mathscr{A}), \varepsilon)$ for every event $\mathscr{A}$ in some Ω_n.

On Ω_n, we use the usual lattice order: $(x_1, ..., x_n) \leqslant (y_1, ..., y_n)$ iff $x_j \leqslant y_j$ for all $j \in [n]$. A function $f: \Omega_n \to \mathbf{R}$ is **monotone** if $f(x) \leqslant f(y)$ whenever $x \leqslant y$. An event $\mathscr{A} \subset \Omega_n$ is **monotone** if its indicator function $\chi_{\mathscr{A}}$ is monotone.

For monotone events, Theorem 1.3 has a converse:

Theorem **1.4.** *Let $\mathscr{A}_m \subset \Omega_{n_m}$ be a sequence of monotone events with*

$$\inf_m \mathbf{H}(\mathscr{A}_m) > 0.$$

Then $\{\mathscr{A}_m\}$ is not asymptotically noise sensitive.

The assumption that the events $\mathscr{A}_m$ are monotone is necessary here. (For example, take $\mathscr{A}_m$ to be a uniform random subset of Ω_m, or parity: $\mathscr{A}_m := \{x \in \Omega_m : \|x\|_1 \text{ is odd}\}$.)

Suppose that $\mathscr{A}$ is a monotone event where the influences of all the variables are the same. The influence $\mathbf{I}_1(\mathscr{A})$ then measures the sensitivity of $\mathscr{A}$ to flips of a single variable. Note that, quite paradoxically, $\mathscr{A}$ is *least* sensitive to noise when $\mathbf{I}_1(\mathscr{A})$ is largest.

We now give a quantitative version of Theorem 1.3 under the assumption that $\mathbf{H}(\mathscr{A}_m)$ goes to zero fast enough.

Theorem **1.5.** — *Let $\mathscr{A} \subset \Omega_n$, and suppose that $\mathbf{H}(\mathscr{A}) \leqslant n^{-a}$, where $a \in (0, 1/2]$. Then there exist $c_1, c_2 > 0$, depending only on a so that*

$$\phi(\mathscr{A}, \varepsilon) \leqslant c_1 n^{-c_2 \varepsilon}, \qquad \forall \varepsilon \in (0, 1/4).$$

Consequently, if $\mathscr{A}_m \subset \Omega_{n_m}$ is a sequence of events satisfying $\mathbf{H}(\mathscr{A}_m) \leqslant (n_m)^{-a}$ and ε_m is a sequence in $(0, 1/4)$ such that $\varepsilon_m \log n_m \to \infty$, then $\phi(\mathscr{A}_m, \varepsilon_m) \to 0$.

1.3. *Weighted majority*

It turns out that for monotone events noise insensitivity is also closely related to correlation with majority functions.

Let $K \subset [n]$ and define the **majority function** on K by

$$M_K(x) = \operatorname{sign} \sum_{j \in K} (2x_j - 1);$$

that is,

$$M_K(x) = \begin{cases} -1 & \text{if } \sum_{j\in K} x_j < |K|/2 \ ; \\ 0 & \text{if } \sum_{j\in K} x_j = |K|/2 \ ; \\ 1 & \text{if } \sum_{j\in K} x_j > |K|/2 . \end{cases} \tag{1.4}$$

For $f: \Omega_n \to \mathbf{R}$ set

$$\Lambda(f) = \max\Big\{ |\mathbf{E}(f M_K)| : K \subset [n] \Big\}.$$

Theorem **1.6.** — *Let* $f: \Omega_n \to [0, 1]$ *be monotone. Then*

$$\mathbf{H}(f) \leqslant C\Lambda(f)^2 \Big(1 - \log \Lambda(f)\Big) \log n,$$

where C *is some universal constant.*

Consequently, if $\mathscr{A}_m \subset \Omega_{n_m}$ *is a sequence of monotone events with*

$$\lim_{m\to\infty} \Lambda(\mathscr{A}_m)^2 \Big(1 - \log \Lambda(\mathscr{A}_m)\Big) \log n_m = 0. \tag{1.5}$$

Then $\{\mathscr{A}_m\}$ *is asymptotically noise sensitive.*

One cannot get rid of the $\log n_m$ factor (see Remark 3.10), except by using weighted majority functions. For positive weights $\mathbf{w} = (w_1, w_2, ..., w_n)$ consider a weighted majority function, which is defined by

$$M_{\mathbf{w}}(x_1, x_2, ..., x_n) = \operatorname{sign}\Big(\sum (2x_j - 1) w_j\Big).$$

Finally write

$$\tilde{\Lambda}(A) = \max\Big\{ |\mathbf{E}(\chi_A M_{\mathbf{w}})| : \mathbf{w} \in [0, 1]^n \Big\}.$$

Theorem **1.7.** — *Let* $\mathscr{A}_m \subset \Omega_{n_m}$ *be a sequence of monotone events. Then* $\{\mathscr{A}_m\}$ *is asymptotically noise sensitive if and only if* $\lim_{m\to\infty} \tilde{\Lambda}(A_m) = 0$.

For a monotone event $\mathscr{A} \subset \Omega_n$, which is symmetric in the n variables, its correlation with unweighted majority is enough to determine if it is noise sensitive.

1.4. *Stability*

We now define the notion of **stability**, which is the opposite of noise sensitivity. Suppose $\mathcal{A} \subset \Omega_n$, and let $x \in \Omega_n$ be random-uniform. For $\varepsilon > 0$, let $N_\varepsilon \mathcal{A}$ denote the event $N_\varepsilon(x) \in \mathcal{A}$. It is then clear that $\mathbf{P}[\mathcal{A} \triangle N_\varepsilon \mathcal{A}] \to 0$ as $\varepsilon \to 0$. ($\mathcal{B} \triangle \mathcal{A}$ denotes the symmetric difference, $(\mathcal{B} - \mathcal{A}) \cup (\mathcal{A} - \mathcal{B})$.) The faster $\mathbf{P}[\mathcal{A} \triangle N_\varepsilon \mathcal{A}]$ tends to zero, the more noise-stable $\mathcal{A}$ is. More precisely, let $\{\mathcal{A}_i\}$ be a collection of events, where $\mathcal{A}_i \subset \Omega_{n_i}$. We say that $\{\mathcal{A}_i\}$ are **uniformly stable** if the limit $\lim_{\varepsilon \to 0} \mathbf{P}[x \in \mathcal{A}_i \triangle N_\varepsilon \mathcal{A}_i] = 0$ is uniform in i.

For $\mathbf{w} \in \mathbf{R}^n$ and $s \in \mathbf{R}$, let $\mathcal{M}_{\mathbf{w},s}$ be the (generalized) weighted majority event

$$\mathcal{M}_{\mathbf{w},s} := \left\{ x \in \mathbf{R}^n : \sum_{j=1}^{n} (2x_j - 1) w_j > s \right\} \subset \Omega_n.$$

Let $\mathfrak{M}$ denote the collection of such events:

$$\mathfrak{M} := \left\{ \mathcal{M}_{\mathbf{w},s} : n = 1, 2, \ldots, \mathbf{w} \in \mathbf{R}^n, \ \mathbf{s} \in \mathbf{R} \right\}.$$

In Section 3 we show that

Theorem **1.8.** - *$\mathfrak{M}$ is uniformly stable. Moreover, for every $\mathcal{M} \in \mathfrak{M}$*

$$\mathbf{P}[\mathcal{M} - N_\varepsilon \mathcal{M}] \leqslant C\varepsilon^{1/4},$$

where C is a universal constant independent of $\mathcal{M}$.

Note that an infinite sequence $\{\mathcal{A}_i\}$ with $\mathbf{P}[\mathcal{A}_i]$ bounded away from 0 and 1 cannot be asymptotically noise sensitive and uniformly stable. We also observe (Lemma 3.8) that when $\{\mathcal{A}_i\}$, $(\mathcal{A}_i \subset \Omega_{n_i})$, is asymptotically noise sensitive and $\{\mathcal{B}_i\}$, $(\mathcal{B}_i \subset \Omega_{n_i})$, is uniformly stable, then $\mathcal{A}_i$ and $\mathcal{B}_i$ are asymptotically uncorrelated. One can say, somewhat imprecisely, that the noise sensitive events are asymptotically in the orthocomplement of the uniformly stable events.

Stability and sensitivity are two extremes. However, there are events that are neither sensitive nor stable. For example, if $\mathcal{C}$ is the event of a percolation crossing, as described above, and $\mathcal{M}$ is the majority event, then $\mathcal{C} \cap \mathcal{M}$ is neither asymptotically noise sensitive, nor uniformly stable.

1.5. *Fourier-Walsh expansion*

For a boolean function f on $\{0, 1\}^n$, consider the Fourier-Walsh expansion $f = \sum_{S \subset [n]} \widehat{f}(S) u_S$, where, $u_S(T) = (-1)^{|S \cap T|}$. Here and in the following, we identify any vector $x \in \Omega^n$ with the subset $\{j \in [n] : x_j = 1\}$, of $[n] = \{1, 2, \ldots, n\}$. Consequently, $|x|$ denotes the cardinality of that set; that is, $|x| = \|x\|_1$ for $x \in \Omega_n$.

Theorem **1.9.** -- *Let $\mathcal{A}_m \subset \Omega_{n_m}$ be a sequence of events, and set $g_m = \chi_{\mathcal{A}_m}$. Then $\{\mathcal{A}_m\}$ is asymptotically noise sensitive iff for every finite k*

$$\lim_m \sum \left\{ \widehat{g}_m(S)^2 : S \subset [n],\ 1 \leqslant |S| \leqslant k \right\} = 0. \tag{1.6}$$

$\{\mathcal{A}_m\}$ is uniformly stable iff

$$\lim_{k\to\infty} \sup_m \sum \left\{ \widehat{g}_m(S)^2 : S \subset [n],\ |S| \geqslant k \right\} = 0. \tag{1.7}$$

It can be easily shown that for $f = \chi_{\mathcal{A}}$

$$I(f) = 4 \sum_{S\subset[n]} \widehat{f}(S)^2 |S|.$$

(This follows from (2.5) below with $p = 2$.) We will introduce another quantity

$$J(f) = \sum_{\emptyset \neq S\subset[n]} \widehat{f}(S)^2/|S|.$$

Also set for $\mathcal{A} \subset \Omega_n$, $n > 1$,

$$\alpha(\mathcal{A}) = \log I(\mathcal{A})/\log n,$$
$$\beta(\mathcal{A}) = -\log J(\mathcal{A})/\log n.$$

For events $\mathcal{A}$ we clearly have $0 \leqslant \beta(\mathcal{A})$, and $\beta(\mathcal{A}) \leqslant \alpha(\mathcal{A})$, provided that $\mathbf{P}[\mathcal{A}] = 1/2$. When $\mathcal{A}$ is monotone $\alpha(\mathcal{A}) \leqslant 1/2$.

Perhaps some words of explanation are needed. $I(\mathcal{A})$ measures the sum of the influences of the variables. For monotone events it is maximal for majority, where $I(\mathcal{A}) \simeq \sqrt{n}$ and thus $\alpha(\mathcal{A}) \to 1/2$. In the terminology used in percolation theory, $I(\mathcal{A})$ is the expected number of pivotal edges.

For the crossing events $\mathcal{C}$ of percolation (in arbitrary dimensions) it is conjectured that $I(\mathcal{C})$ behaves like a certain fractional power (a *critical exponent*) of n. It is conjectured that in dimension 2, as n tends to infinity, $\alpha(\mathcal{C})$ tends to $3/8$. Thus, this critical exponent generalizes and has a Fourier-analysis interpretation for arbitrary Boolean functions.

$\alpha(\mathcal{A})$ is large if there are substantial Fourier coefficients $\widehat{f}(S)$ for large $|S|$. In contrast, $\beta(\mathcal{A})$ is large if there are *no* substantial Fourier coefficients $\widehat{f}(S)$ for S of small positive size. We conjecture that for the crossing events for percolation, as n tends to infinity $\beta(\mathcal{C})$ tends to a positive limit. We are curious to know whether this limit is strictly smaller than the limit for $\alpha(\mathcal{C})$.

1.6. *Some related and future work*

There are interesting connections between noise sensitivity and isoperimetric inequalities of the form described by Talagrand in [32]. These connections and

applications for first passage percolation problems will be discussed in a subsequent paper [6].

Our notion of noise sensitivity is related to the study of noises by Tsirelson [34, 35]. "Noise", in Tsirelson's sense, is a type of σ-field filtration. Uniform stability seems to correspond, in the limit, to the noise being white, while asymptotic sensitivity seems to correspond to the noise being black.

1.7. *The structure of this paper*

Theorems 1.3 and 1.4 are proved in the next section. Our proofs combines combinatorial reasonings with applying certain inequalities for the Fourier coefficients of Bonami and Beckner which were used already in [23]. However, to get the results in the sharpest forms we have to rely on a sophisticated "bootstrap" method of [33] and on the main results of that paper which rely on this method. Talagrand's remarkable paper [33] has thus much influence on the present work.

Weighted and unweighted majority functions are considered in Section 3. An applications to percolation is described in Section 4 followed by some related open problems in Section 5. In Section 6, we will work out two examples (due to Ben-Or and Linial). In one of these $\alpha(\mathcal{A}) \to 1 - \log_2 3$ and $\beta(\mathcal{A}) \to 1 - \log_2 3$. In Section 7 we consider relations with complexity theory. A simple description of noise-sensitivity in terms of random walks is given in Section 8. In Section 9 we consider perturbations with a different sort of noise, where the number of bits that are changed is fixed. The conclusions are similar to those above, but there is an amusing and slightly unexpected twist.

For simplicity we consider here the uniform measure on Ω_n. More generally, one may consider the product measure $\mathbf{P}_p$, where $\mathbf{P}_p\{x : x_j = 1\} = p$. Our results and proof apply in this setting. (All that is needed is to replace the Fourier-Walsh transform by its analog as given in Talagrand's paper [31] and the proofs go through without change.) However, the case when p itself depends on n is interesting, but will not be considered here.

Since the first version of this paper was distributed, a few of the problems we posed were settled by several people, not always in the direction anticipated by us. These developments are mentioned briefly in a few "late remarks" throughout the paper.

Acknowledgments.

It is a pleasure to thank Noga Alon, Ehud Friedgut, Ravi Kannan, Harry Kesten, Yuval Peres, Michel Talagrand and Avi Wigderson for helpful discussions.

2. Sensitivity to noise

We now put the noise operator N_ε defined in the introduction into a somewhat more general framework. That will allow us to deal, for example, with the situation where the 1 bits are immune to noise but the 0 bits are noise prone.

Consider the following method for selecting a random point $x \in \Omega_n$. Let $q_1, ..., q_n$ be independent random variables in [0,1], with $\mathbf{E}q_j = 1/2$, for $j = 1, ..., n$, and let $\omega \in [0, 1]^n$ be random uniform. Set

$$x_j = \begin{cases} 1, & \text{if } 1 - \omega_j < q_j, \\ 0, & \text{otherwise.} \end{cases}$$

Then x is distributed according to the uniform measure of Ω_n; it will be denoted by $N(\omega, q)$.

Let ν be the measure on $[0, 1]^n$ such that $\nu(X) = \mathbf{P}[(q_1, ..., q_n) \in X]$. We think of x as being chosen in two stages. In the first stage, $q = (q_1, ..., q_n)$ is selected according to ν. This q gives a product measure $\mathbf{P}_q$ on $\{0, 1\}^n$ that satisfies $\mathbf{P}_q\{\tau \in \Omega_n : \tau(j) = 1\} = q_j$. Then x is chosen according to the measure $\mathbf{P}_q$.

For example, suppose $z \in \Omega_n$. Define $q = q(z) \in [0, 1]^n$ by $q_j = 1 - \varepsilon$ if $z_j = 1$ and $q_j = \varepsilon$ if $z_j = 0$. Then for every $z \in \Omega_n$, the perturbation $N_\varepsilon(z)$ has the same distribution as $N(\omega, q(z))$. The ν giving this distribution of q will be denoted ν_ε.

However, the construction $N(\omega, q)$ is more general than that given by the noise operator N_ε. As hinted above, one can create a situation where 1 bits are robust, but 0 bits are prone to noise. More precisely, take $q_j = 1$, with probability $1/2 - \varepsilon$ and $q_j = \varepsilon/(1/2 + \varepsilon)$ with probability $1/2 + \varepsilon$.

Another interesting example is obtained by taking each q_j to be 1, with probability $(1 - \varepsilon)/2$, 0, with probability $(1 - \varepsilon)/2$, and $1/2$ with probability ε.

Let $f: \{0, 1\}^n \to \mathbf{R}$ be some function. In the following, f will be taken to be the characteristic function $\chi_{\mathcal{A}}$ of some event $\mathcal{A} \subset \{0, 1\}^n$, or $f = \chi_{\mathcal{A}} - \mathbf{P}(\mathcal{A})$. What information does the first stage in the selection of $x = N(\omega, q)$, namely the selection of q, give about the value of $f(x)$? If we know that $q = z$, then our prediction for $f(x)$ would be

$$G(f, z) = \mathbf{E}(f(x) \mid q = z).$$

The expected value of $G(f, q)$ is obviously $\mathbf{E}(f)$. Let

$$Z(f, \nu) = \mathbf{E}_q G(f, q)^2 = \int G(f, z)^2 \, d\nu(z).$$

This is just the second moment of $G(f, q)$. If $Z(f, \nu) - (\mathbf{E}f)^2$ is small, then for "most" values of q there is no prediction for $f(x)$ that is significantly better than the a priori knowledge of $\mathbf{E}f$. We often write $G(\mathcal{A}, \cdot)$ and $Z(\mathcal{A}, \cdot)$ in place of $G(\chi_{\mathcal{A}}, \cdot)$ and $Z(\chi_{\mathcal{A}}, \cdot)$.

Lemma **2.1.** — *The number* $Z(f, \nu)$ *depends only on* f *and the variances* ζ_j *of the variables* q_j. *Its expression in terms of the Fourier coefficients is,*

$$Z(f, \nu) = \sum_{S \in \Omega_n} \hat{f}(S)^2 \prod_{j \in S} 4\zeta_j.$$

Proof. —

$$\begin{aligned}
G(f, z) &= \mathbf{E}(f(x) \mid q = z) \\
&= \sum_{T \subset [n]} f(T) \prod_{j \in T} z_j \prod_{j \notin T} (1 - z_j) \\
&= \sum_{T} \sum_{S} \hat{f}(S)(-1)^{|T \cap S|} \prod_{j \in T} z_j \prod_{j \notin T} (1 - z_j) \\
&= \sum_{S} \hat{f}(S) \left(\sum_{T' \subset S} (-1)^{|T'|} \prod_{j \in T'} z_j \prod_{j \notin T'} (1 - z_j) \right) \cdot \\
&\qquad \left(\sum_{T'' \subset [n]-S} \prod_{j \in T''} z_j \prod_{j \notin T''} (1 - z_j) \right) \\
&= \sum_{S} \hat{f}(S) \left(\prod_{j \in S} ((1 - z_j) - z_j) \right) \left(\prod_{j \notin S} ((1 - z_j) + z_j) \right) \\
&= \sum_{S} \hat{f}(S) \prod_{j \in S} (1 - 2z_j).
\end{aligned}$$

Therefore,

$$\begin{aligned}
Z(f, \nu) &= \mathbf{E}G(f, q)^2 \\
&= \sum_{S} \sum_{S'} \hat{f}(S)\hat{f}(S')\mathbf{E} \left(\prod_{j \in S} (1 - 2q_j) \prod_{j \in S'} (1 - 2q_j) \right) \\
&= \sum_{S} \sum_{S'} \hat{f}(S)\hat{f}(S') \prod_{j \in S \cap S'} \mathbf{E}(1 - 2q_j)^2 \prod_{j \in S \triangle S'} \mathbf{E}(1 - 2q_j).
\end{aligned}$$

Since $\mathbf{E}q_j = 1/2$, summands with $S \neq S'$ vanish. The lemma follows. □

For every $\varepsilon \in [0, 1]$, $x \in \Omega_m$ and $f: \Omega_n \to \mathbf{R}$ set

$$Q_\varepsilon f(x) = \mathbf{E}f(N_\varepsilon(x))$$

(here the expectation is only with respect to the noise). Also let

$$\mathrm{var}(f, \varepsilon) = \mathrm{var}(Q_\varepsilon f) = Z(f, \nu_\varepsilon) - (\mathbf{E}f)^2.$$

Note that for singletons $S = \{i\} \subset [n]$, we have $Q_\varepsilon u_S = (1 - 2\varepsilon)u_S$. If $S_1, S_2 \subset [n]$ are disjoint and $x \in \Omega_n$ is fixed, then $N_\varepsilon(x) \cap S_1$ and $N_\varepsilon(x) \cap S_2$ are independent.

Consequently, $Q_\varepsilon(u_{S_1} u_{S_2}) = (Q_\varepsilon u_{S_1})(Q_\varepsilon u_{S_2})$. We may conclude that $Q_\varepsilon u_S = (1-2\varepsilon)^{|S|} u_S$ for every $S \subset [n]$, and linearity gives

$$Q_\varepsilon f = \sum_{S \subset [n]} \hat{f}(S)(1-2\varepsilon)^{|S|} u_S. \tag{2.1}$$

One consequence of this, which can also be obtained from Lemma 2.1, is

$$\mathrm{var}(f, \varepsilon) = \sum_{\emptyset \neq S \subset [n]} \hat{f}(S)^2 (1-2\varepsilon)^{2|S|}. \tag{2.2}$$

Now we relate $\mathrm{var}(\mathscr{A}, \varepsilon)$ with the sensitivity gauge $\phi(\mathscr{A}, \varepsilon)$:

Proposition **2.2.** — *For every* $\mathscr{A} \subset \Omega_n$

$$\frac{1}{2}\,\mathrm{var}(\mathscr{A}, \varepsilon) \leqslant \phi(\mathscr{A}, \varepsilon) \leqslant \mathrm{var}(\mathscr{A}, \varepsilon)^{1/3}.$$

Proof. — Let $\delta = \phi(\mathscr{A}, \varepsilon)$, and set

$$Y = \left\{ y \in \Omega_n : \left| Q_\varepsilon \chi_{\mathscr{A}}(y) - \mathbf{P}[\mathscr{A}] \right| \geqslant \delta \right\}.$$

Then, by the definition of ϕ, $\mathbf{P}[Y] \geqslant \delta$. Consequently,

$$\mathrm{var}(\mathscr{A}, \varepsilon) \geqslant \delta^2 \mathbf{P}[Y] \geqslant \delta^3 = \phi(\mathscr{A}, \varepsilon)^3.$$

For the other direction set

$$Y' = \left\{ y \in \Omega_n : \left| Q_\varepsilon \chi_{\mathscr{A}}(y) - \mathbf{P}[\mathscr{A}] \right| > \delta \right\}.$$

Then $\mathbf{P}[Y'] \leqslant \delta$. For $y \in Y'$, the trivial estimate $\left| Q_\varepsilon \chi_{\mathscr{A}} - \mathbf{P}[\mathscr{A}] \right| \leqslant 1$ holds. Therefore,

$$\mathrm{var}(\mathscr{A}, \varepsilon) \leqslant \mathbf{P}[Y'] + \delta^2 \leqslant 2\phi(\mathscr{A}, \varepsilon). \quad \square$$

Proof of **1.9.** — The first part is immediate from Prop. 2.2 and (2.2). For the proof of the second part, observe that (2.1) implies that (1.7) is equivalent to $\|g_m - Q_\varepsilon g_m\|_2 \to 0$ uniformly as $\varepsilon \to 0$. Since $|g_m|$ and $|Q_\varepsilon g_m|$ are bounded, this is equivalent to $\|g_m - Q_\varepsilon g_m\|_1 \to 0$ uniformly, which is the same as uniform stability for $\{\mathscr{A}_m\}$. $\square$

Remark **2.3.** — Another consequence of 2.2 and (2.2) is that for constant $\varepsilon, \varepsilon' \in (0, 1/2)$, we have $\phi(\mathscr{A}_m, \varepsilon) \to 0$ iff $\phi(\mathscr{A}_m, \varepsilon') \to 0$. Consequently, to verify that $\mathscr{A}_m$ is asymptotically noise sensitive, it is enough to prove $\mathrm{var}(\mathscr{A}_m, \varepsilon) \to 0$ with any fixed $\varepsilon \in (0, 1/2)$.

By Theorem 1.9, to establish Theorem 1.3 we need to show that the L^2 weight of the Fourier coefficients with $|S|$ small is negligible. For a function $g = \sum \widehat{g}(S) u_S$ let

$$T_\eta g = Q_{\frac{1-\eta}{2}} g = \sum \widehat{g}(S) \eta^{|S|} u_S.$$

Observe that $T_0(g) = \mathbf{E}g$ and $T_1 g = g$. Also note that

$$\|T_{1-2\varepsilon} g\|_2^2 = \mathrm{var}(g, \varepsilon) + \widehat{g}(\emptyset)^2, \tag{2.3}$$

by (2.2).

The following hyper-contractive inequality of Bonami and Beckner [7, 3], which was crucial in [23], will be useful.

Lemma **2.4** (Bonami, Beckner). — $\|T_\eta f\|_2 \leqslant \|f\|_{1+\eta^2}$.

The following is a slightly weaker version of Theorem 1.3, which is sufficient for the applications to percolation. It is presented here, since we can give an almost self-contained proof of it.

Theorem **2.5.** — *Suppose that $\mathscr{A}_m \subset \Omega_{n_m}$ is a sequence of events and*

$$\lim_{m\to\infty} \frac{\log \mathbf{H}(\mathscr{A}_m)}{\log \log n_m} = -\infty. \tag{2.4}$$

Then $\{\mathscr{A}_m\}$ is asymptotically noise sensitive.

Proof. — Abbreviate $\mathscr{A}$ for $\mathscr{A}_m$ and n for n_m, and set $f := \chi_{\mathscr{A}}$. Let $\widetilde{f} := \chi_{\mathscr{A}} - \mathbf{P}[\mathscr{A}]$. Thus, $\widehat{\widetilde{f}}(\emptyset) = 0$ and $\widehat{\widetilde{f}}(S) = \widehat{f}(S)$, when $S \neq \emptyset$.

Recall that $\sigma_j x = (x'_1, \ldots, x'_n)$, where $x'_i = x_i$ if $i \neq j$ and $x'_j = 1 - x_j$. Let

$$f_j(x) = f(x) - f(\sigma_j x), \qquad j = 1, 2, \ldots, n,$$

and note that

$$\widehat{f_j}(S) = \begin{cases} 0, & \text{if } j \notin S, \\ 2\widehat{f}(S), & \text{if } j \in S. \end{cases}$$

Since f_j takes only the values -1, 0, 1, equation (1.3) gives for every $p \geqslant 1$,

$$\|f_j\|_p = I_j(f)^{1/p}. \tag{2.5}$$

We set $\eta := 1 - 2\varepsilon$, where $\varepsilon \in (0, 1/2)$ and

$$F_{\mathscr{A}}(\eta) := \mathrm{var}(\mathscr{A}, \varepsilon) = \|T_\eta \widetilde{f}\|_2^2 = \sum_{S \neq \emptyset} \widehat{f}(S)^2 \eta^{2|S|}.$$

By Remark 2.3 and Theorem 2.2, it is enough to prove that $F_{\mathscr{A}}(1/2) \to 0$ as $m \to \infty$. We have

$$
\begin{aligned}
F_{\mathscr{A}}(\eta) &\leqslant \sum_S \hat{f}(S)^2 |S| \eta^{2|S|} = \frac{1}{4} \sum_{j=1}^{n} \|T_\eta f_j\|_2^2 \\
&\leqslant \frac{1}{4} \sum_{j=1}^{n} \|f_j\|_{1+\eta^2}^2 \qquad \text{(by Lemma 2.4)} \\
&\leqslant \sum_{j=1}^{n} I_j(\mathscr{A})^{2/(1+\eta^2)} \qquad \text{(by (2.5))} \\
&\leqslant n^{\eta^2/(1+\eta^2)} \mathbf{H}(\mathscr{A})^{1/(1+\eta^2)} \qquad \text{(by the means inequality).}
\end{aligned} \tag{2.6}
$$

Take some $\eta_1 \in (0, 1/2)$, to be later specified, and set $\lambda := \log F_{\mathscr{A}}(\eta_1)/\log \eta_1$. If $\eta \geqslant \eta_1$, then

$$
\begin{aligned}
F_{\mathscr{A}}(\eta) &\leqslant \sum_{1 \leqslant |S| \leqslant \lambda/2} \hat{f}(S)^2 \eta^{2|S|} + \eta^{\lambda} \sum_S \hat{f}(S)^2 \\
&\leqslant (\eta/\eta_1)^{\lambda} F_{\mathscr{A}}(\eta_1) + \eta^{\lambda} = 2\eta^{\lambda}.
\end{aligned} \tag{2.7}
$$

Assume that $\mathbf{H}(\mathscr{A}) \in (0, e^{-2})$, and let $a := \min\left\{-\log \mathbf{H}(\mathscr{A})/\log n, 1/2\right\}$. We may choose $\eta_1 := \sqrt{a/2}$. Then $\mathbf{H}(\mathscr{A}) \leqslant n^{-a}$, and therefore (2.6) and the definition of λ give

$$
\lambda \geqslant \frac{a \log n}{3 \log(1/a)}. \tag{2.8}
$$

The definition of a together with (2.4) and (2.8) show that $\lambda \to \infty$ as $m \to \infty$. Hence (2.7) implies $F_{\mathscr{A}_m}(1/2) \to 0$ as $m \to \infty$, which completes the proof. □

Proof of **1.5.** - - The above calculations together with Prop. 2.2 show that

$$
\phi(\mathscr{A}, \varepsilon) \leqslant \operatorname{var}(\mathscr{A}, \varepsilon)^{1/3} = F_{\mathscr{A}}(1 - 2\varepsilon)^{1/3} \leqslant 2^{1/3} \left(1 - 2\varepsilon\right)^{\frac{a \log n}{9 \log(1/a)}},
$$

for $\varepsilon \in (0, 1/4)$, when we assume $\mathbf{H}(\mathscr{A}) \leqslant n^{-a}$, $a \in (0, 1/2]$. The theorem follows immediately. □

For the proof of Theorem 1.3, we will need the following.

Theorem **2.6.** — *For each $k = 1, 2, \ldots$, there is a constant $C_k < \infty$ with the following property. Let $\mathscr{A} \subset \Omega_n$ be a monotone event and $f = \chi_{\mathscr{A}}$. Then*

$$
\sum_{|S| = k} \hat{f}(S)^2 \leqslant C_k \mathbf{H}(A) \left(-\log \mathbf{H}(A)\right)^{k-1}
$$

This inequality was proved by Talagrand [33] for $k=2$. (Talagrand considers an extension of this relation for two events, and our generalization applies for that extension as well.)

Proof of **2.6.** — To prove the theorem one can follow Talagrand's proof almost word-by-word. We will only describe the changes needed to adapt the proof. One modification required is that the inequality

$$\mathbf{P}\left\{S' : \left|\sum_{|S|=k} \alpha_S u_S(S')\right| \geqslant t\right\} \leqslant e^2 \exp\left(-t^{2/k}\left(e^2 \sum \alpha_S^2\right)^{-1/k}\right) \tag{2.9}$$

must be used in place of the sub-Gaussian estimate that appears as Prop. 2.1 in [33]. Set $q = t^{2/k}/\left(e^2 \sum \alpha_S^2\right)^{1/k}$. For $q \leqslant 2$ the inequality (2.9) is trivial, while for $q > 2$ it follows by substituting q into

$$\left\|\sum_{|S|=k} \alpha_S u_S\right\|_q \leqslant (q-1)^{k/2} \left(\sum \alpha_S^2\right)^{1/2}, \qquad \forall q \geqslant 2, \tag{2.10}$$

which appears in [31] as (2.4) and is a consequence of the dual version of the Bonami-Beckner inequality.

Set $A_k := \{x \in \mathscr{A} : \sigma_k x \notin \mathscr{A}\}$, and note that $2\mathbf{P}[A_k] = I_k(\mathscr{A})$. In the proof for the case $k=2$, Talagrand considers in Section 3 of [33] partitions $I \cup J = [n]$, and estimates $\sum\{I_j(\mathscr{A})^2 : j \in L(s)\}$, where $L(s)$ is the set of $j \in J$ such that

$$\sum_{i \in I} \left(\int_{A_j} u_{\{i\}}(x)\right)^2 \geqslant s^2 \mathbf{P}[A_j]^2 .$$

To generalize Talagrand's argument for $k > 2$, one gives a similar estimate to $\sum\{I_j(\mathscr{A})^2 : j \in L_{k-1}(s)\}$, where $L_{k-1}(s)$ is the set of $j \in J$ such that

$$\sum \left\{ \left(\int_{A_j} u_i(x)\right)^2 : i \subset I,\ |i| = k-1 \right\} \geqslant s^2 \mathbf{P}[A_j]^2 .$$

We omit the details, since from this point on only straightforward changes are required to adapt Talagrand's beautiful (but rather mysterious) argument. □

In the case of monotone events, Theorem 1.3 follows immediately from Theorems 2.6 and 1.9. In order to get rid of the monotonicity assumption, we introduce the shifting operator.

Let $j \in \{1, ..., n\}$, and let $f: \Omega_n \to \mathbf{R}$. For $x \in \Omega_n$, set

$$\kappa_j f(x) := \begin{cases} \max\{f(x), f(\sigma_j x)\}, & \text{if } x_j = 1, \\ \min\{f(x), f(\sigma_j x)\}, & \text{if } x_j = 0. \end{cases}$$

The operator κ_j is called the ***j*-shift**. The following lemma describes some useful properties of shifts.

Lemma **2.7** (Shifting). — *Let* $f: \Omega_n \to \mathbf{R}$, *and let* $j, i \in \{1, ..., n\}$. *Then*

1. $\kappa_1 \kappa_2 ... \kappa_n f$ *is monotone.*
2. $I_i(\kappa_j f) \leqslant I_i(f)$.
3. $\mathrm{var}(\kappa_j f, \varepsilon) \geqslant \mathrm{var}(f, \varepsilon)$ *for each* $\varepsilon \in [0, 1]$.

Proof. — Suppose for the moment that $i \neq j$. For any $a, b \in \{0, 1\}$ and $x \in \Omega_n$, let $x_{a,b}$ be x with the i'th coordinate set to a and the j'th coordinate set to b. Note that $\kappa_j f$ is monotone nondecreasing in the variable x_j. Hence $\kappa_i \kappa_j f(x_{1,1})$ is the maximum of f on $\{x_{0,0}, x_{0,1}, x_{1,0}, x_{1,1}\}$ and $\kappa_i \kappa_j f(x_{0,0})$ is the minimum. It follows that $\kappa_i \kappa_j f = \kappa_j \kappa_i \kappa_j f$. This relation easily implies the first claim of the lemma.

For the second part, we may assume with no loss of generality that $j \neq i$, because $I_i(\kappa_i f) = I_i(f)$. A case by case analysis shows that

$$\begin{aligned} &|f(x_{0,0}) - f(x_{1,0})| + |f(x_{0,1}) - f(x_{1,1})| \\ &\qquad \geqslant |\kappa_j f(x_{0,0}) - \kappa_j f(x_{1,0})| + |\kappa_j f(x_{0,1}) - \kappa_j f(x_{1,1})|, \end{aligned}$$

and the second part follows by summing over $x \in \Omega_n$.

For the last part, set

$$\begin{aligned} g(y) &= \mathbf{E}[f(\mathrm{N}_\varepsilon(x)) \mid x = y], \\ \tilde{g}(y) &= \mathbf{E}[\kappa_j f(\mathrm{N}_\varepsilon(x)) \mid x = y]. \end{aligned}$$

Note that $g(y) + g(\sigma_j(y)) = \tilde{g}(y) + \tilde{g}(\sigma_j y)$, but $|g(y) - g(\sigma_j(y))| \leqslant |\tilde{g}(y) - \tilde{g}(\sigma_j y)|$. This implies $g(y)^2 + g(\sigma_j(y))^2 \leqslant \tilde{g}(y)^2 + \tilde{g}(\sigma_j y)^2$. By summing over y, we obtain $\mathbf{E}(g^2) \leqslant \mathbf{E}(\tilde{g}^2)$. Since $\mathbf{E}g = \mathbf{E}\tilde{g}$, the last claim of the lemma now follows. □

Proof of **1.3.** — Let $\mathscr{A} \subset \Omega_n$. Set $g = \kappa_1 \kappa_2 ... \kappa_n \chi_{\mathscr{A}}$. Then by Lemma 2.7, g is monotone, $\mathbf{H}(g) \leqslant \mathbf{H}(\mathscr{A})$ and for each $\varepsilon > 0$ we have $\mathrm{var}(g, \varepsilon) \geqslant \mathrm{var}(\mathscr{A}, \varepsilon)$. Moreover, g takes only the values 0 and 1. By applying Theorem 2.6 for g, and using Theorem 2.2, Theorem 1.3 immediately follows. □

Proof of **1.4.** — Observe that for a monotone $f: \Omega_n \to \mathbf{R}$

$$I_j(f) = 2|\hat{f}(\{j\})|, \tag{2.11}$$

and therefore

$$\mathbf{H}(f) = 4\sum_j \hat{f}(\{j\})^2. \tag{2.12}$$

Hence 1.4 follows from Theorem 1.9. □

Note that (2.12) implies the well-known inequality

$$\mathbf{H}(\mathcal{A}) \leqslant 1 \tag{2.13}$$

for monotone events $\mathcal{A}$.

Remark **2.8.** — It is tempting to look for a simpler proof of Theorem 1.3, along the following lines. Using (2.5) with $p=2$, we find that

$$\mathbf{H}(f) = \sum_{j=1}^{n} \left(\sum_{S \supset \{j\}} 4\hat{f}(S)^2 \right)^2 = 16 \sum_{S,S'} \hat{f}(S)^2 \hat{f}(S')^2 |S \cap S'|, \tag{2.14}$$

where $f = \chi_{\mathcal{A}}$ for some event $\mathcal{A} \subset \Omega_n$. This expression is more complicated than (2.12), but is still valid when $\mathcal{A}$ is not monotone. The fact that f is the indicator function of an event is summarized by the equation $f^2 = f$. In terms of the Fourier transform, this translates to a convolution equation

$$\hat{f} * \hat{f} = \hat{f}. \tag{2.15}$$

(By replacing f with $2f-1$, this transforms to the simpler looking $\hat{f} * \hat{f} = \chi_{\{\emptyset\}}$.) One may suspect that there should be a direct argument that uses only (2.15) and (2.14) to prove that for every $k = 1, 2, \ldots$

$$\sum_{|S|=k} \hat{f}(S)^2 \to 0$$

when $\mathbf{H}(f) \to 0$. Then Theorem 1.3 would follow from Theorem 1.9.

3. Correlation with majority

3.1. *Uniform weights*

Fix some $n \in \mathbf{N}$. Recall the definition (1.4) of the majority function M_K, and set $M = M_n = M_{[n]}$.

Theorem **3.1.** — *Let* $f: \Omega_n \to [0, 1]$ *be monotone. Then*

$$I(f) \leqslant C\sqrt{n}\,\mathbf{E}(fM)\left(1 + \sqrt{-\log \mathbf{E}(fM)}\right),$$

where C *is some universal constant.*

Proof. — Write $\bar{f}(k)$ for the average of f on the set $\left\{x : \sum_j x_j = k\right\}$:

$$\bar{f}(k) = \binom{n}{k}^{-1} \sum_{|x|=k} f(x).$$

Then

$$\mathbf{E}(f\mathrm{M}) = 2^{-n} \sum_{k > \frac{n}{2}} \binom{n}{k} \left(\bar{f}(k) - \bar{f}(n-k)\right). \tag{3.1}$$

Recall that $s_j x = (y_1, \ldots, y_n)$ where $y_j = 1 - x_j$ and $y_i = x_i$ for $i \neq j$. Then

$$\mathrm{I}(f) = 2^{-n} \sum_x \sum_j |f(x) - f(s_j x)|.$$

Since f is monotone, $f(x) - f(s_j x) \geqslant 0$ when $x_j = 1$ and $f(x) - f(s_j x) \leqslant 0$ when $x_j = 0$. Hence the expression for $\mathrm{I}(f)$ simplifies,

$$\begin{aligned} \mathrm{I}(f) &= 2^{-n} \sum_x f(x)\left(2|x| - n\right) \\ &= 2^{-n} \sum_k \binom{n}{k} \bar{f}(k)(2k - n) \\ &= 2^{-n} \sum_{k > \frac{n}{2}} \binom{n}{k} \left(\bar{f}(k) - \bar{f}(n-k)\right)(2k - n). \end{aligned} \tag{3.2}$$

For any $\lambda \geqslant 0$ write $k(\lambda) = (n + \lambda\sqrt{n})/2$. Since $0 \leqslant \bar{f}(k) \leqslant 1$, by comparing (3.2) and (3.1), we obtain the following estimate.

$$\begin{aligned} \mathrm{I}(f) &\leqslant (2k(\lambda) - n)\,\mathbf{E}(f\mathrm{M}) + 2^{-n} \sum_{k > k(\lambda)} \binom{n}{k} \left(\bar{f}(k) - \bar{f}(n-k)\right)(2k - n) \\ &\leqslant \lambda\sqrt{n}\,\mathbf{E}(f\mathrm{M}) + 2^{-n} \sum_{k > k(\lambda)} \binom{n}{k}(2k - n). \end{aligned} \tag{3.3}$$

Because there are constants $\mathrm{C}_1, \mathrm{C}_2 > 0$ such that

$$2^{-n} \binom{n}{k}(2k - n) \leqslant \mathrm{C}_1 \exp\left(-\frac{(2k-n)^2}{\mathrm{C}_2 n}\right) \tag{3.4}$$

holds for every n and k, by choosing $\lambda = C_3\sqrt{-\log \mathbf{E}(f\mathrm{M})}$, where C_3 is a sufficiently large constant, we get

$$2^{-n} \sum_{k>k(\lambda)} \binom{n}{k} (2k-n) \leqslant C_4 \sqrt{n}\, \mathbf{E}(f\mathrm{M}),$$

and the theorem follows from (3.3). □

Given a set $\mathrm{K} \subset [n]$, let M_K denote the majority function on the set K; that is,

$$\mathrm{M}_\mathrm{K}(x) = \begin{cases} -1 & \text{if } \sum_{j\in\mathrm{K}} x_j < |\mathrm{K}|/2\,; \\ 0 & \text{if } \sum_{j\in\mathrm{K}} x_j = |\mathrm{K}|/2\,; \\ 1 & \text{if } \sum_{j\in\mathrm{K}} x_j > |\mathrm{K}|/2\,. \end{cases}$$

Also set,

$$\mathrm{I}_\mathrm{K}(f) = \sum_{k\in\mathrm{K}} \mathrm{I}_k(f).$$

Corollary **3.2.** — *Let* $\mathrm{K} \subset [n]$ *and suppose that* $f\colon \Omega_n \to [0, 1]$ *is monotone. Then*

$$\mathrm{I}_\mathrm{K}(f) \leqslant \mathrm{C}\sqrt{|\mathrm{K}|}\, \mathbf{E}(f\mathrm{M}_\mathrm{K}) \left(1 + \sqrt{-\log \mathbf{E}(f\mathrm{M}_\mathrm{K})}\right),$$

where C *is some universal constant.*

Proof. — Set $m = |\mathrm{K}|$, and assume, that $\mathrm{K} = \{1, \ldots, m\}$. Given $z \in \Omega_m$, set

$$f_\mathrm{K}(z) = 2^{m-n} \sum_{y\in\Omega_{n-m}} f(z, y).$$

Then f_K is monotone and $\mathrm{I}(f_\mathrm{K}) = \mathrm{I}_\mathrm{K}(f)$. Consequently, the corollary follows from Theorem 3.1. □

Proof of **1.6.** — Assume, with no loss of generality, that

$$\mathrm{I}_{j+1}(f) \leqslant \mathrm{I}_j(f) \tag{3.5}$$

for all $j \in \{1, \ldots, n-1\}$. Cor. 3.2 implies that

$$\sum_{j=1}^{k} \mathrm{I}_j(f) \leqslant C_1 \Lambda(f) \left(1 + \sqrt{-\log \Lambda(f)}\right) \sqrt{k} \tag{3.6}$$

for some constant C_1 and every $k \in [n]$. Subject to these constraints and (3.5), $\mathbf{H}(f)$ is maximized if equality occurs in (3.6) for every k. Therefore,

$$\mathbf{H}(f) \leqslant C_1^2 \Lambda(f)^2 \left(1 + \sqrt{-\log \Lambda(f)}\right)^2 \sum_{k=1}^{n} \left(\sqrt{k} - \sqrt{k-1}\right)^2$$

$$= O(1)\Lambda(f)^2 \left(1 - \log \Lambda(f)\right) \sum_{k=1}^{n} k^{-1}$$

$$= O(1)\Lambda(f)^2 \left(1 - \log \Lambda(f)\right) \log n.$$

This proves the first part of Theorem 1.6. The second part now follows from Theorem 1.3. □

Theorem 1.6 tells us that if $\Lambda(\mathscr{A}_m) \to 0$ fast enough for monotone events $\mathscr{A}_m$, then they are asymptotically noise sensitive. Conversely, if a sequence of (not necessarily monotone) events satisfies $\inf_m \Lambda(\mathscr{A}_m) > 0$, then it is not asymptotically noise sensitive. This can be proven directly, and also follows from Lemma 3.8 below.

It is interesting to note that

Theorem **3.3.** — *Majority maximizes* I *among monotone events* $\mathscr{A} \subset \Omega_n$.

This follows from [15], although the explicit statement does not appear there. It also follows from the classical Kruskal-Katona theorem. See also [18, Lem. 6.1].

3.2. *General weights*

We will investigate now some relations between noise-sensitivity and weighted majority functions. Several of the properties we need for weighted majority functions are easy to establish if the distribution of weights allows us to use a normal approximation for $f(x) = \sum_j w_j x_j$. But, as it turns out, working with arbitrary weights is harder.

Our first goal is to show that weighted majority functions are uniformly noise stable. This will imply the "only if" part of Theorem 1.7. For this, the following easy (and quite standard) lemma will be needed.

Lemma **3.4.** — *Let* $w = (w_1, ..., w_n) \neq 0$ *and* $f(x) = \sum_j w_j(2x_j - 1)$. *Then*

$$\mathbf{P}[|f| \geqslant t\|w\|_2] \leqslant 3t^{-4}, \tag{3.7}$$

and

$$\mathbf{P}[|f| \leqslant 0.3\|w\|_2] \leqslant 0.92. \tag{3.8}$$

A much stronger estimate than (3.7) is known (see [28]).

Proof. — Without loss of generality, we assume that $\|w\|_2 = 1$. Then $\mathbf{E}[f^4] = 3\|w\|_2^4 - 2\|w\|_4^4 \leqslant 3$. Hence (3.7) follows:

$$\mathbf{P}[|f| \geqslant t] = \mathbf{P}[f^4 \geqslant t^4] \leqslant t^{-4}\mathbf{E}[f^4] = 3t^{-4}.$$

This implies

$$\mathbf{E}[\mathbf{1}_{\{f^2>t\}} f^2] = \mathbf{P}[f^2 > t] + \int_{s \geqslant t} \mathbf{P}[f^2 > s]\, ds \leqslant 3t^{-2} + t^{-3}.$$

Hence

$$\mathbf{E}[\mathbf{1}_{\{f^2 \leqslant t\}} f^2] = \mathbf{E}[f^2] - \mathbf{E}[f^2 \mathbf{1}_{\{f^2>t\}}] \geqslant 1 - 3t^{-2} - t^{-3}.$$

We choose $t = 10$, and obtain

$$\begin{aligned} 10 - 9.9\mathbf{P}[f^2 \leqslant 1/10] &= 10\mathbf{P}[f^2 > 1/10] + \mathbf{P}[f^2 \leqslant 1/10]/10 \\ &\geqslant \mathbf{E}[\mathbf{1}_{\{f^2 \leqslant 10\}} f^2] \geqslant 9/10, \end{aligned}$$

which gives (3.8). □

Lemma **3.5.** — *Let $b > 0$, let $v_1, \ldots, v_d \geqslant b$, and let $g = \sum_{j=1}^d z_j v_j$, where $\mathbf{P}[z_j = 1] = \mathbf{P}[z_j = -1] = 1/2$, and the z_j are independent. Then for every $t \geqslant 1$ and every $s \in \mathbf{R}$,*

$$\mathbf{P}[|g - s| \leqslant tb] \leqslant c \cdot t/\sqrt{d}, \tag{3.9}$$

where c is some universal constant.

This lemma is a consequence of Theorem 2.14 in [28], for example. However, since the proof of that theorem is arduous, we now present a simple combinatorial proof.

Proof. — Let x be a random uniform element in Ω_d, and let π be a random uniform permutation of $\{1, 2, \ldots, d\}$. Let C be the collection of sets S that have the form $S = \{j : \pi(j) < r\}$ for some $r \in \mathbf{R}$. Then there is a unique $y \in C$ with $|y| = |x|$. Observe that y is a random uniform element of Ω_d. Consequently, the distribution of g is the same as the distribution of $h(y) := \sum_{j=1}^d (1 - 2y_j)v_j$, where y_j is 1 or 0 when $j \in y$ or $j \notin y$, respectively. Since C is totally ordered by inclusion, there is at most one $S \in C$ such that $|h(S) - s| < b/2$. So when π is fixed, the probability that $|h(S) - s| \leqslant b/3$ is at most $\max\{\mathbf{P}[|x| = r] : r \in \mathbf{R}\} = O(1)/\sqrt{d}$. This establishes (3.9) for $t = 1/3$. The result for general $t \geqslant 1$ follows by applying the result for $t = 1/3$ for an appropriate succession of values of s. □

Proof of **1.8.** — Let $w = (w_1, \ldots, w_n) \neq 0$ and $s_0 \in \mathbf{R}$. Let $f(x) := \sum_{j=1}^n w_j(2x_j - 1)$, and consider the event $\mathcal{M} := \{x \in \Omega : f(x) > s_0\}$. Take $\varepsilon > 0$, and let $J \subset [n]$ be

a random subset, where each $j \in [n]$ is in J with probability ε, independently. Set $Y(J) := \sum_{j \in J} w_j(2x_j - 1)$. Then $2Y(J)$ has the distribution of $f - N_\varepsilon f$. Let $\delta \in (0, 1)$ and set

$$a := \inf\{t > 0 : \mathbf{P}[|Y(J)| \geqslant t] \leqslant \delta\}. \tag{3.10}$$

Our goal is to give an estimate from above to $P[|f| < 2a]$ in terms of ε and δ, which will tend to zero when δ is positive and fixed and $\varepsilon \to 0$.

Set $W(J) := \sum_{j \in J} w_j^2$. This is the variance of $Y(J)$ conditioned on J. Note that

$$\mathbf{P}[|Y(J)| \geqslant a \mid J] = \mathbf{P}[Y(J)^2 > a^2 \mid J] \leqslant \mathbf{E}[Y(J)^2 \mid J]/a^2 = W(J)/a^2.$$

Therefore,

$$\begin{aligned} \delta = \mathbf{P}[|Y(J)| > a] &= \sum_{X \subset [n]} \mathbf{P}[|Y(J)| > a \mid J = X]\mathbf{P}[J = X] \\ &\leqslant \sum_{X \subset [n]} \min\{1, a^{-2}W(X)\}\mathbf{P}[J = X] = \mathbf{E}\left[\min\{1, a^{-2}W(J)\}\right], \end{aligned} \tag{3.11}$$

and we conclude that

$$\mathbf{P}[W(J) \geqslant \delta a^2/2] \geqslant \delta/2. \tag{3.12}$$

Now let $z_1, z_2, \ldots, z_n$ be independent variables that are uniform in [0, 1], and are independent from $(x_1, \ldots, x_n)$. Let m be the largest integer such that $m\varepsilon \leqslant 1$. Let $I_1, \ldots, I_m$ be disjoint open intervals in [0, 1], each of length ε. Let $I_0 := [0, 1] - \cup_{k=1}^m I_k$. Let J_k $(k = 0, 1, \ldots, m)$ be the set of $i \in [n]$ with $z_i \in I_k$. Then each J_k with $k > 0$ has the same distribution as J above. Let $\mathscr{A}_k$ be the event that $W(J_k) \geqslant \delta a^2/2$. Then from (3.12) with J_k in place of J we find that $\mathbf{P}[\mathscr{A}_k] \geqslant \delta/2$ for $k = 1, 2, \ldots, m$.

We claim that for $k \neq k'$ the events $\mathscr{A}_k$ and $\mathscr{A}_{k'}$ are negatively correlated. This can be established by proving by induction on n that the events $W(J_k) \geqslant s_1$ and $W(J_{k'}) \geqslant s_2$ are negatively correlated for each $s_1, s_2 \in \mathbf{R}$ (which is intuitively obvious, since the intervals I_k and $I_{k'}$ are disjoint). Let K be the number of $k > 0$ such that the event $\mathscr{A}_k$ occurs. Then

$$\mathbf{E}[K] = \sum_{k=1}^m \mathbf{P}[\mathscr{A}_k] \geqslant m\delta/2,$$

and

$$\mathbf{E}[K^2] - \mathbf{E}[K]^2 = \sum_{k, k'} \mathbf{P}[\mathscr{A}_k \cap \mathscr{A}_{k'}] - \left(\sum_k \mathbf{P}[\mathscr{A}_k]\right)^2 \leqslant \mathbf{E}[K],$$

because the events $\mathcal{A}_k, \mathcal{A}_{k'}$ are negatively correlated when $k \neq k'$. Therefore

$$\begin{aligned} \mathbf{P}[\mathrm{K} < m\delta/4] &\leqslant \mathbf{P}\left[(\mathrm{K} - \mathbf{E}\mathrm{K})^2 > (m\delta/4)^2\right] \\ &\leqslant 4(m\delta)^{-2}\mathbf{E}\left[(\mathrm{K} - \mathbf{E}\mathrm{K})^2\right] = 4(m\delta)^{-2}\left(\mathbf{E}[\mathrm{K}^2] - (\mathbf{E}\mathrm{K})^2\right) \\ &\leqslant 4(m\delta)^{-2}\mathbf{E}[\mathrm{K}] \leqslant 4m^{-1}\delta^{-2}. \end{aligned} \tag{3.13}$$

Let L be the set of $k \in [m]$ such that $|\mathrm{Y}(\mathrm{J}_k)| > a\sqrt{\delta}/10$. By (3.8), applied to $\mathrm{Y}(\mathrm{J}_k)$ in place of f,

$$\mathbf{P}[k \in \mathrm{L} \mid \mathcal{A}_k] \geqslant 8/100.$$

Moreover, conditioned on all the J_k, the events $\{k \in \mathrm{L}\}$ are independent. Consequently, a calculation similar to (3.13) gives

$$\mathbf{P}[|\mathrm{L}| < m\delta/100 \mid \mathrm{K} \geqslant m\delta/4] \leqslant \mathrm{O}(1)\, m^{-1}\delta^{-2}.$$

When we use this and (3.13) together, we get

$$\mathbf{P}[|\mathrm{L}| < m\delta/100] \leqslant \mathrm{O}(1)\, m^{-1}\delta^{-2}. \tag{3.14}$$

If we condition on L, on all $\mathrm{Y}(\mathrm{J}_k)$ for $k \notin \mathrm{L}$ and on all $|\mathrm{Y}(\mathrm{J}_k)|$ for $k \in \mathrm{L}$, then what remains to determine f are only the signs of $\mathrm{Y}(\mathrm{J}_k)$ with $k \in \mathrm{L}$. Moreover, these signs are independent, and are + or − with probability 1/2. Hence we may apply Lemma 3.5 with $b := a\sqrt{\delta}/10$, $d := |\mathrm{L}|$, $s := s_0 - \sum_{k \notin \mathrm{L}} \mathrm{Y}(\mathrm{J}_k)$, $g = \sum_{k \in \mathrm{L}} \mathrm{Y}(\mathrm{J}_k)$, and take $v = (v_k)$ to be the sequence $(|\mathrm{Y}(\mathrm{J}_k)| : k \in \mathrm{L})$. The conclusion is that for $t \geqslant 1$

$$\mathbf{P}\left[|f - s_0| \leqslant ta\sqrt{\delta}/10 \;\middle|\; |\mathrm{L}| \geqslant m\delta/100\right] \leqslant \mathrm{O}(1)\, t/\sqrt{m\delta}.$$

Together with (3.14), (and choosing $t = 20/\sqrt{\delta}$) this gives

$$\mathbf{P}\left[|f - s_0| \leqslant 2a\right] \leqslant \mathrm{O}(1)\left(\varepsilon\delta^{-2} + \sqrt{\varepsilon}\delta^{-1}\right). \tag{3.15}$$

We now come to analyze the effect of noise. Because $2\mathrm{Y}(\mathrm{J})$ has the same distribution as $f - \mathrm{N}_\varepsilon f$, for every $a > 0$

$$\mathbf{P}[\mathcal{M} \triangle \mathrm{N}_\varepsilon \mathcal{M}] \leqslant \mathbf{P}[|f - s_0| \leqslant 2a] + \mathbf{P}[|\mathrm{Y}(\mathrm{J})| \geqslant a].$$

Choose $\delta := \varepsilon^{1/4}$ and, as before, use (3.10) to define a. Then $\mathbf{P}[|f - s_0| \leqslant 2a] \leqslant \mathrm{O}(1)\varepsilon^{1/4}$ and $\mathbf{P}[|\mathrm{Y}(\mathrm{J})| \geqslant a] \leqslant \varepsilon^{1/4}$. Consequently,

$$\mathbf{P}[\mathcal{M} \triangle \mathrm{N}_\varepsilon \mathcal{M}] \leqslant \mathrm{O}(1)\varepsilon^{1/4}, \tag{3.16}$$

and the theorem immediately follows. □

Question **3.1.** — What is the best exponent possible on the right hand side of (3.16)?

Late Remark **3.6.** — Yuval Peres and Elchanan Mossel found a simple proof showing that, as expected, the correct exponent is 1/2.

Remark **3.7.** — It follows from Theorems 1.8 and 1.3 that $\inf\{\mathbf{H}(\mathcal{M}) : \mathcal{M} \in \mathfrak{M}\} > 0$. (A direct proof will follow.) We conjecture that $\mathbf{H}(\mathcal{M})$ is minimized among $\mathcal{M}_{w,0} \subset \Omega_n$ in $\mathfrak{M}$ when all the weights are equal. It is a consequence of Theorems 1.8 and 1.9 that

$$\lim_{k\to\infty} \sup_{\mathcal{M}\in\mathfrak{M}} \sum_{|S|>k} \hat{\chi}_{\mathcal{M}}(S)^2 = 0.$$

We actually expect that among weighted majority events in Ω_n, the one with equal weights is the least stable, and for every $k > 1$ maximizes $\sum_{|S|>k} \hat{\chi}_{\mathcal{M}}(S)^2$.

For the proof of 1.7, the following will be needed.

Lemma **3.8.** — *Let $\mathcal{A}_m, \mathcal{B}_m \subset \Omega_{n_m}$ be two sequences of events. Suppose that the sequence $\{\mathcal{A}_m\}$ is noise-sensitive, while the sequence $\{\mathcal{B}_m\}$ is noise-stable. Then*

$$\lim_m \mathbf{P}[\mathcal{A}_m \cap \mathcal{B}_m] - \mathbf{P}[\mathcal{A}_m]\mathbf{P}[\mathcal{B}_m] = 0.$$

Proof. — This can be proven directly, but since $\mathbf{P}[\mathcal{A}_m \cap \mathcal{B}_m] = \mathbf{E}[\chi_{\mathcal{A}_m}\chi_{\mathcal{B}_m}]$, the lemma is immediate from 1.9. □

Let the **influence vector** of an event $\mathcal{A} \subset \Omega_n$ be the vector $\mathrm{I}^{\mathcal{A}} := (\mathrm{I}_1(\mathcal{A}), \ldots, \mathrm{I}_n(\mathcal{A})) \in \mathbf{R}^n$.

Proof of **1.7.** — The "only if" direction follows from Theorem 1.8 and Lemma 3.8.

For the other direction, we need to show that monotone, noise-insensitive events $\mathcal{A} \subset \Omega_n$ have a non-vanishing correlation with some weighted majority event $\mathcal{M}_w$, $w \in [0, 1]^n$. Talagrand's Theorem 1.1 [33] gives a lower bound on the correlation of monotone events. This theorem asserts, in particular, that for two monotone events, if the inner product of their influence vectors is bounded away from zero, then the correlation between them is also bounded away from zero (1).

We know from Theorem 1.3 that for noise-insensitive events, $\|\mathrm{I}^{\mathcal{A}}\|_2$ is bounded away from zero. It remains to show that for every $v \in [0, 1]^n$ with $\|v\|_2 = 1$, we can

(1) For uniformly stable events, it seems that also the converse is true: if the correlation is bounded away from zero, then so is the inner product of their influence vectors. For monotone uniformly stable events, this follows from the two-event version of Theorem 2.6.

find a weighted majority function $\mathcal{M} = \mathcal{M}_w$, $w \in [0, 1]^n$, such that the inner product $\langle \mathbf{I}^{\mathcal{M}}, v \rangle$ is bounded away from zero. We will prove that this holds when one chooses $w := v$.

Given any $w \in \mathbf{R}^n$, $w \neq 0$, let $\mathbf{I}^w \in \mathbf{R}^n$ denote the influence vector of $\mathcal{M}_w$, $\mathbf{I}^w_j := \mathbf{I}_j(\mathcal{M}_w)$.

Proposition **3.9.** — *There is an absolute constant $c > 0$ such that $\langle w, \mathbf{I}^w \rangle \geqslant c$ for every $n = 1, 2, \ldots$ and every $w \in \mathbf{R}^n$ with nonnegative coordinates and $\|w\|_2 = 1$.*

Proof. — Set $f(x) = \sum_{j=1}^n (2x_j - 1)w_j$ for $x \in \Omega_n$. Then $\widehat{f}(\{j\}) = w_j$ for $j \in [n]$ and $\widehat{f}(S) = 0$ for $S \subset [n]$, $|S| \neq 1$. On the other hand, $\mathbf{I}_j(\mathcal{M}_w) = \widehat{\mathbf{M}}_w(\{j\})$, where $\mathbf{M}_w = \operatorname{sign}(f)$. Therefore,

$$\langle w, \mathbf{I}^w \rangle = \langle \widehat{f}, \widehat{\mathbf{M}}_w \rangle = \langle f, \mathbf{M}_w \rangle = \mathbf{E}\,|f(x)|,$$

which is bounded from below, by (3.8). This completes the proof of the proposition, and the proof of Theorem 1.7. □

Remark **3.10.** — We now show that one cannot remove the log in Theorem 1.6. Fix some $k, n \in \mathbf{Z}$ with $n \geqslant k > 0$. Let $w_j = 1/\sqrt{j \log n}$ for $j = 1, \ldots, n$, and let $u_j = 1/\sqrt{k}$ for $j \leqslant k$ and $u_j = 0$ for $j > k$. Set $f_w(x) = \sum_{j=1}^n (2x_j - 1)w_j$ and $f_u(x) = \sum_{j=1}^n (2x_j - 1)u_j$, where $x \in \Omega_n$. Then the event $f_w \geqslant 0$ is noise stable, by 1.8. We show that $\mathbf{P}[f_w \geqslant 0 \mid f_u \geqslant 0] \to 1/2$ as $n \to \infty$, no matter how $k = k(n)$ is chosen. Indeed, given any $x \in \Omega$, let $s(x) := \sqrt{k} f_u(x) = \sum_{j \leqslant k} (2x_j - 1)$. If $s(x) < 0$, let $\bar{x}$ be obtained from x by replacing $-s(x)$ of the 0 entries in x by 1's, where the set of entries replaced is chosen randomly and uniformly among all possibilities, and if $s(x) \geqslant 0$, set $\bar{x} = x$. Then $\mathbf{P}[f_w(x) \geqslant 0 \mid f_u(x) \geqslant 0] = \mathbf{P}[f_w(\bar{x}) \geqslant 0]$. Therefore, by Lemma 3.5 applied to w, it is enough to show that $f_w(\bar{x}) - f_w(x) \to 0$ in probability as $n \to \infty$. This follows from

$$\mathbf{E}|f_w(\bar{x}) - f_w(x)| = (2/k)\mathbf{E}[\max\{0, -s(x)\}] \sum_{j=1}^{k} w_j = O(1)/\sqrt{\log n}.$$

4. An application to percolation

Let R be an $(m + 1) \times m$ rectangle in the square grid $\mathbf{Z}^2$, and let Ω be the set of all functions from E, the set of edges of R, to $\{0, 1\}$. We identify Ω with Ω_n; where $n = n_m = |E| = 2m^2 - 1$. A point $x \in \Omega$ is called a **configuration**, and can be identified with the subgraph consisting of all vertices of R and all edges e with $x(e) = 1$. A connected component of this graph is called a **percolation cluster**.

Let $\mathscr{C} = \mathscr{C}_m \subset \Omega$ be the event that there is a left-right crossing of R; that is, $\mathscr{C}$ is the set of all configurations that contain a path joining the left and right boundaries of R. An easy and well known application of duality shows that $\mathbf{P}[\mathscr{C}] = 1/2$.

Kesten [24] gives an estimate from above for the probability that an edge near the middle of R is pivotal for $\mathscr{C}$. Similar estimates for edges near the boundary can probably be extracted from Kesten's paper. These give an inequality of the form $I_j(\mathscr{C}_m) \leqslant m^{-1-c}$, $c > 0$, for each j. Then Theorem 2.5 implies 1.2. However, we prefer to present another proof, based on Theorem 1.6.

The only percolation background needed to understand the proof is that in our situation the probability that a vertex in R is connected in the configuration to some vertex at Euclidean distance r is at most $Cr^{-1/\rho}$, for some constants $C, \rho > 0$. This follows from the celebrated Russo-Seymour-Welsh Theorem [29, 30] (see also [19]).

Proof of **1.2.** — Let E_r be the set of edges in the right half of R, with edges exactly centered included. Let $K \subset E_r$. We now estimate $\mathbf{E}(\chi_{\mathscr{C}} M_K)$.

Consider the following algorithmic method of randomly selecting a configuration. Let ω^K and $\widehat{\omega}^K$ be two independent elements of $\Omega_{|K|}$ and $\Omega_{n-|K|}$, respectively. Let V_1 be the set of vertices on the left boundary of R, and set VISITED $= \emptyset$. As long as there is some edge $[v, u] \notin$ VISITED joining a vertex $v \in V_1$ to a vertex $u \notin V_1$, choose some such edge $e = [v, u]$, and do the following. Append e to VISITED. If $e \in K$, let $y(e)$ be the first bit in the sequence ω^K that has not been previously used by the algorithm, while if $e \notin K$ let $y(e)$ be the first bit in the sequence $\widehat{\omega}^K$ that has not been previously used by the algorithm. If $y(e) = 1$, then adjoin to V_1 the vertex u.

This procedure defines y for all $e \in$ VISITED. Let $z \in \Omega$ be random, uniform, and independent of y, and let $x = y$ on VISITED while $x = z$ on E − VISITED. This defines a configuration $x \in \Omega$.

The following is obvious:

Lemma **4.1.** — *The configuration x given by the above algorithm is uniformly distributed in Ω. The event $x \in \mathscr{C}$ is equal to the event that at the end of the algorithm V_1 intersects the right boundary and is independent from z (can be determined by y).* □

Let us estimate the probability that $K \cap$ VISITED is large. An edge $e \in K$ is in VISITED iff there is in x a path joining a vertex of e to the left boundary of R. Since $K \subset E_r$, it follows from the above stated consequence of the Russo-Seymour-Welsh Theorem that the probability for the latter event is bounded by $Cm^{-1/\rho}$, for some constants $C, \rho > 0$. Consequently,

$$\mathbf{E}|K \cap \text{VISITED}| \leqslant C|K|m^{-1/\rho},$$

which implies

$$\mathbf{P}[\mathscr{A}_1] \leqslant Cm^{-1/(3\rho)},$$

where $\mathcal{A}_1$ is the event

$$\mathcal{A}_1 := \left\{ x \in \Omega : |\mathrm{K} \cap \mathrm{VISITED}| \geqslant |\mathrm{K}| m^{-2/(3\rho)} \right\}.$$

Let $\mathcal{A}_2$ be the event that there is an integer j in the range $1 \leqslant j \leqslant |\mathrm{K}| m^{-2/(3\rho)}$ such that

$$\left| \frac{j}{2} - \sum_{i=1}^{j} \omega_i^{\mathrm{K}} \right| \geqslant \sqrt{|\mathrm{K}| m^{-2/(3\rho)}} \log m.$$

It is easy to see that the $\mathbf{P}[\mathcal{A}_2]$ decays super-polynomially in m; in particular,

$$\mathbf{P}[\mathcal{A}_2] \leqslant \mathrm{O}(m^{-1/\rho}).$$

As $\mathbf{P}[\mathcal{A}_1 \cup \mathcal{A}_2] \leqslant \mathrm{O}(1) m^{-1/(3\rho)}$, we have

$$\mathbf{E}(\chi_{\mathcal{A}_1 \cup \mathcal{A}_2} \chi_{\mathscr{C}} \mathrm{M}_{\mathrm{K}}) \leqslant \mathrm{O}(1)\, m^{-1/(3\rho)}. \tag{4.1}$$

Now suppose that the algorithm produced a y such that $\mathcal{A}_1 \cup \mathcal{A}_2$ does not hold. Then it follows that

$$\left| \frac{|\mathrm{VISITED} \cap \mathrm{K}|}{2} - \sum_{e \in \mathrm{VISITED} \cap \mathrm{K}} y(e) \right| \leqslant \mathrm{O}(1) \sqrt{|\mathrm{K}| m^{-2/(3\rho)}} \log m.$$

This implies that

$$\mathbf{E}[\mathrm{M}_{\mathrm{K}}(x) \mid y] \leqslant \mathrm{O}(1)\, m^{-1/(3\rho)} \log m, \qquad \forall y \notin \mathcal{A}_1 \cup \mathcal{A}_2.$$

Since $x \in \mathscr{C}$ can be determined from y, we get

$$\mathbf{E}\Big((1 - \chi_{\mathcal{A}_1 \cup \mathcal{A}_2}) \chi_{\mathscr{C}} \mathrm{M}_{\mathrm{K}} \Big) \leqslant \mathrm{O}(1)\, m^{-1/(3\rho)} \log m.$$

In view of (4.1) this implies

$$\mathbf{E}(\chi_{\mathscr{C}} \mathrm{M}_{\mathrm{K}}) \leqslant \mathrm{O}(1)\, m^{-1/(3\rho)} \log m,$$

and Cor. 3.2 gives

$$\mathrm{I}_{\mathrm{K}}(\mathscr{C}) \leqslant \mathrm{O}(1) \sqrt{|\mathrm{K}|} m^{-1/(3\rho)} (\log m)^{3/2} \tag{4.2}$$

for every $\mathrm{K} \subset \mathrm{E}_r$, since $\mathscr{C}$ is monotone. By symmetry, this would also hold for $\mathrm{K} \subset \mathrm{E} - \mathrm{E}_r$, and therefore for every $\mathrm{K} \subset \mathrm{E}$. Consequently, by the proof of Theorem 1.6

$$\mathbf{H}(\mathscr{C}) \leqslant \mathrm{O}(1)\, m^{-2/(3\rho)} (\log m)^4. \tag{4.3}$$

An appeal to Theorem 2.5 completes the proof. □

Remark **4.2.** -- Since $I(\mathscr{C}) = \sum_e I_e(\mathscr{C})$ is also the expected number of pivotal edges for $\mathscr{C}$, (4.2) shows that the expected number of pivotal edges is bounded by

$$O(1)\, m^{1-1/(3\rho)} (\log m)^{3/2}.$$

Although this is better than the general bound of $O(1)\, m$ that follows from Theorem 3.3, a somewhat better bound can be extracted from Kesten's [24].

Corollary **4.3.** -- *There is a constant $c > 0$ with the following property. If $\varepsilon = c/\log m$, then for large m, with probability at least $1/4$, $\left| \{x, N_\varepsilon(x)\} \cap \mathscr{C} \right| = 1$. That is, if each edge is switched with probability $c/\log m$, independently, then the crossing is likely to be created or destroyed.*

The corollary follows from (4.3) and Theorem 1.5. The details are left to the reader.

5. Some conjectures and problems concerning percolation

5.1. *Other sensitivity conjectures*

Consider the crossing event $\mathscr{C}_m$ for an $(m+1) \times m$ rectangle in the square grid $\mathbf{Z}^d$. By Theorem 1.2 and Section 2, from knowing which edges are open for all but a small random set of edges, we have almost no information whether crossing occurs. This suggests that for some deterministic subsets of the rectangle $R = R_m$, knowing the configuration restricted to that configuration typically gives almost no information whether crossing occurs. It follows from the Russo-Seymour-Welsh Theorem [29, 30] that E_r, the set of edges in the right half of the rectangle, is not such a subset. Yet we believe that all the horizontal edges (or all the vertical edges) is such a subset. That is, let $x, y \in \Omega$ be two independent uniform-random configurations. Let $z(e) = x(e)$ for horizontal edges e, and $z(e) = y(e)$ for vertical edges. Let $p(\omega) = \mathbf{P}[z \in \mathscr{C} \,|\, x = \omega]$.

Conjecture **5.1.** - - For any $\varepsilon > 0$, for all sufficiently large m,

$$\mathbf{P}\left\{\omega \in \Omega : |p(\omega) - 1/2| > \varepsilon\right\} < \varepsilon.$$

Here is a variant of this conjecture for **Voronoi percolation**. Fix a square in $\mathbf{R}^2$. Voronoi percolation is performed in two steps. First pick n points in the square uniformly and independently. Second each cell in the Voronoi tessellation determined by the chosen points is declared open with probability $1/2$, and closed otherwise, independently of the other cells, (see Benjamini and Schramm [5] for the exact definitions and a study of Voronoi percolation). By duality, the probability of open left-right open crossing is $1/2$. In the spirit of Theorem 1.2, we conjecture that typically, knowing the Voronoi tessellation (but not knowing which cells are open) gives almost no information whether an open left-right crossing exits.

5.2. *Stronger sensitivity conjectures*

As before, let $\mathscr{C}_m$ denote the crossing event for an $(m+1) \times m$ rectangle in the square grid $\mathbf{Z}^2$.

Conjecture **5.2.** — There exists a $\beta > 0$ so that $\lim_{m\to\infty} \phi(\mathscr{C}_m, m^{-\beta}) = 0$.

It is known [25, 24], respectively, that for some reals $0 < b_1 < b_2 < 1$,

$$m^{b_1} \leqslant \mathrm{I}(\mathscr{C}_m) \leqslant m^{b_2},$$

and it is conjectured (see, e.g., [14], p. 91) that $\mathrm{I}(\mathscr{C}_m)$ behaves like $m^{3/4}$.

Problem **5.3.** -- Is it true that $\lim_{m\to\infty} \phi(\mathscr{C}_m, \varepsilon_m) = 0$ when $\varepsilon_m = o(m^{-3/4})$?

Recall that our proof of noise sensitivity for $\mathscr{C}_m$ used (indirectly) upper bounds on $\mathrm{I}(\mathscr{C}_m)$. On the other hand, there is a simple heuristic argument that directly relates noise sensitivity to the *lower bounds* on $\mathrm{I}(\mathscr{C}_m)$ and the distribution of the number of pivotal edges. Namely, we cannot expect the crossing event $\mathscr{C}_m$ to be stable under noise which, with a very high probability, will flip pivotal edges. This argument tends to support Conjecture 5.2.

5.3. *Dynamical percolation*

Dynamical percolation was introduced by Häggström, Peres and Steif [20]. Consider the following process. Let $\{\mathrm{X}_e\}$ be independent Poisson point processes in $\mathbf{R}$ indexed by the edges $e \in \mathrm{E}_\mathrm{R}$ of the $(m+1) \times m$ rectangle $\mathrm{R} = \mathrm{R}_m$ in $\mathbf{Z}^2$. Let $x_0 : \mathrm{E}_\mathrm{R} \to \{0, 1\}$ be random-uniform. For each $t > 0$ set $x_t(e) := x_0(e)$ if the number of points in $(0, t] \cap \mathrm{X}_e$ is even, and $x_t(e) := 1 - x_0(e)$ if the number is odd. This gives a continuous time stationary Markov chain x_t in $\Omega = \{0, 1\}^{\mathrm{E}_\mathrm{R}}$. Write $\widetilde{\mathbf{P}}$ for the probability measure governing this process. For each fixed t, the random variable x_t can be thought of as ordinary (Bernoulli (1/2)) percolation in $\mathbf{Z}^2$.

An interesting problem raised by [20] is weather there are (exceptional, random) times t in which there is an infinite percolation cluster in x_t. The result described below might be relevant.

As before, let $\mathscr{C}_m$ denote the set of configurations in Ω that have an open left-right crossing of R_m. For all t, $\widetilde{\mathbf{P}}[x_t \in \mathscr{C}_m] = 1/2$. Let S_m be the set of switching times; that is, S_m is the boundary of $\{t \geqslant 0 : x_t \in \mathscr{C}_m\}$. As a corollary of Theorem 1.2, we have,

Corollary **5.1.** — $\left|\mathrm{S}_m \cap [0, 1]\right| \to \infty$ *in probability.*

Proof. — Suppose $s > t \geqslant 0$. Observe that the distribution of the pair (x_t, x_s) is the same as the distribution of the pair $\left(x_0, N_\varepsilon(x_0)\right)$, where ε is a function of $s-t$ and $\varepsilon > 0$ when $s > t$. (Actually, $\varepsilon/(s-t) \to 1$ as $s - t \to 0$.)

Let k be some positive integer, and set $\varepsilon = \varepsilon(1/k)$. Let $t_j := j/k$. Let $\mathscr{W}$ be the set of $\omega \in \Omega$ such that $\left|\mathbf{P}[N_\varepsilon(\omega) \in \mathscr{C}] - 1/2\right| > 1/4$. Then $\mathbf{P}[\mathscr{W}] \to 0$ as $m \to \infty$, by Theorem 1.2. Let $\mathscr{Z}(a, b)$ be the event that $S \cap [a, b] = \emptyset$. Observe that for $\omega \notin \mathscr{W}$, we have

$$\tilde{\mathbf{P}}\left[\mathscr{Z}(t_j, t_{j+1}) \mid x_{t_j} = \omega\right] \leqslant 3/4,$$

because $\mathscr{Z}(t_j, t_{j+1})$ is disjoint from the event $\left|\{x_{t_j}, x_{t_{j+1}}\} \cap \mathscr{C}\right| = 1$. Hence we can make the following estimate,

$$\begin{aligned}
\tilde{\mathbf{P}}[\mathscr{Z}(0, t_{j+1})] &= \tilde{\mathbf{P}}[\mathscr{Z}(0, t_j) \cap \mathscr{Z}(t_j, t_{j+1})] \\
&= \sum_{\omega \in \Omega} \tilde{\mathbf{P}}\left[\mathscr{Z}(0, t_j) \cap \mathscr{Z}(t_j, t_{j+1}) \mid x_{t_j} = \omega\right] \mathbf{P}\{\omega\} \\
&= \sum_{\omega \in \Omega} \tilde{\mathbf{P}}\left[\mathscr{Z}(0, t_j) \mid x_{t_j} = \omega\right] \tilde{\mathbf{P}}\left[\mathscr{Z}(t_j, t_{j+1}) \mid x_{t_j} = \omega\right] \mathbf{P}\{\omega\} \\
&\qquad \text{(by the Markov property for } x_t\text{)} \\
&\leqslant \mathbf{P}[\mathscr{W}] + \sum_{\omega \in \Omega - \mathscr{W}} \tilde{\mathbf{P}}\left[\mathscr{Z}(0, t_j) \mid x_{t_j} = \omega\right] \tilde{\mathbf{P}}\left[\mathscr{Z}(t_j, t_{j+1}) \mid x_{t_j} = \omega\right] \mathbf{P}\{\omega\} \\
&\leqslant \mathbf{P}[\mathscr{W}] + (3/4) \sum_{\omega \in \Omega - \mathscr{W}} \tilde{\mathbf{P}}\left[\mathscr{Z}(0, t_j) \mid x_{t_j} = \omega\right] \mathbf{P}\{\omega\} \\
&\leqslant \mathbf{P}[\mathscr{W}] + (3/4) \sum_{\omega \in \Omega} \tilde{\mathbf{P}}\left[\mathscr{Z}(0, t_j) \mid x_{t_j} = \omega\right] \mathbf{P}\{\omega\} \\
&= \mathbf{P}[\mathscr{W}] + (3/4) \tilde{\mathbf{P}}\left[\mathscr{Z}(0, t_j)\right].
\end{aligned}$$

Using this inequality and induction gives $\tilde{\mathbf{P}}\left[\mathscr{Z}(0, t_j)\right] \leqslant 4\mathbf{P}[\mathscr{W}] + (3/4)^j$. By stationarity, for every $t \geqslant 0$, the same estimate for the probability of $\mathscr{Z}(t, t + j/k)$ holds. Since k may be chosen arbitrarily large, and $\mathbf{P}[\mathscr{W}] \to 0$ as $m \to \infty$, the corollary easily follows. □

5.4. *Limits and conformal invariance*

The motivating questions behind this work were the conjecture regarding the existence of the limit and the conformal invariance conjecture for two-dimensional percolation.

These conjectures say, roughly, that the crossing probabilities inside a domain between two boundary arcs have a limit as the mesh of the grid goes to zero, and the

limit is invariant under conformal transformations of the domain and the boundary arcs. For more details, see Langlands, Pouliot and Saint-Aubin [26].

Consider a triple $\mathscr{G} = \langle G, A, B \rangle$, where $G = (V, E)$ is a finite planar graph with m edges, and $A, B \subset V$. Let $p_{\mathscr{G}}$ be the probability that there is an open crossing from A to B in a uniform-random configuration $x \in \Omega = \{0, 1\}^E$.

Let $\mathscr{H} = \langle H, A', B' \rangle$ be a triple obtained from G by the following operation: for every edge e of G delete e with probability $(1 - t)/2$ contract e with probability $(1 - t)/2$ and leave e unchanged with probability t, independently of the other edges. $\mathscr{H}$ is a random variable which takes values in planar graphs with two distinguished vertex sets.

Of course, $E(p_{\mathscr{H}}) = p_{\mathscr{G}}$, and noise sensitivity, when it applies, asserts that the value of $p_{\mathscr{H}}$ is concentrated around the mean. Noise sensitivity enables one to relate the crossing probabilities of percolation on different graphs and we had hoped that it will be relevant to conjecture regarding the existence of the limit conjecture. At present, however, such applications are beyond our reach as we do not have a good understanding of planar graphs which are obtained by random deletions and contractions of the form described above when $\mathscr{G}$ is a rectangle in the square grid.

To be more specific, suppose that we take G to be the $m \times cm$ rectangle in $\mathbf{Z}^2$ ($c > 0$ some fixed constant) and let A and B be its left and right boundaries. It follows from Theorem 1.5 and (4.3) that $p_{\mathscr{H}} - p_{\mathscr{G}} \to 0$ in probability, provided that $t \cdot \log m \to \infty$. (Conjecture 5.2 would give it even when $t \cdot m^{\beta} \to \infty$ for some $\beta > 0$.) It is conjectured that the crossing probability tends to a limit as m tends to infinity and an approach to this conjecture would be to relate the distribution of such random planar graphs starting from similar rectangles of different sizes. (The values of t should depend on the size but be large enough that noise sensitivity applies).

In a different direction, the random planar graphs $\mathscr{H}$ obtained when you start with the $(m + 1) \times m$ grid and apply certain random contractions and deletions may be related to models of random planar graphs in mathematical physics [1].

5.5. *Fourier-Walsh coefficients of percolation*

It is a natural question to try to understand the Fourier-Walsh coefficients of boolean functions given by percolation problems. Consider (again) the event $\mathscr{C} = \mathscr{C}_m$ of a left-right crossing of an $(m + 1) \times m$ rectangle $R = R_m$ of the square grid, $\mathbf{Z}^2$. Let $f_m := \chi_{\mathscr{C}_m}$. The Fourier coefficients of f_m are indexed by subsets of E_R, the edges in R_m. The values $\hat{f}^2$ can be regarded as a measure on the space of subgraphs of R_m.

Problem **5.4.** — Describe this measure!

It follows from Theorem 1.5 and our estimates for $\mathbf{H}(\mathscr{C}_m)$, that all but a negligible part of the L^2 weight of the Fourier coefficients $\hat{f}(S)$, where S is non-empty, is for

$|S| > c \log m$. Conjecture 5.2 is equivalent to the assertion that, in fact, this is true for $|S| > m^{\beta}$ for some $\beta > 0$. Conjecture 5.1 is equivalent to the statement that for all but a negligible part of these Fourier coefficients, the number of vertical edges in S tends to infinity with m.

5.6. *Other models of statistical mechanics*

It would be of interest to extend the results of this paper as well as earlier results on influence ([23, 18]) to other models of statistical mechanics, such as the Ising and Potts models. Many of the results on influence and on noise sensitivity should be extendible to measures on Ω_n for which the coordinate variables are **positively associated**, namely, measures for which every two monotone real functions are positively correlated.

6. Some further examples

We will discuss now four examples, the first two were considered by Ben-Or and Linial [4].

6.1. *Tribes*

Consider n boolean variables divided into t tribes $T_1, T_2..., T_t$ of size s each, and let f be the boolean function which take the value 1 if for some j, $1 \leqslant j \leqslant t$, all variables of T_j equal 1. If $s = \log n - \log\log n + \log\log 2$, then $\mathbf{P}[f = 1] \approx \frac{1}{2}$. Also note that $I_k(f) \sim \log n/n$ for every k. It is easy to show directly that f will be immune to ε-noise when $\varepsilon = o(1/\log n)$ and will be devastated by ε-noise if $\varepsilon \log n \to \infty$.

Thus, $J(f) \sim \log n/n$.

6.2. *Recursive majority on the ternary tree*

Consider $n = 3^t$ boolean variables which form the leaves of a rooted ternary tree of height t. A boolean function f is defined as follows: Given values for the variable on the leaves compute for each other vertex its value as the majority of the values of its sons and set the value of f to be the value of the root.

Ben-Or and Linial showed that $I_k(f) \sim n^{-\log 2/\log 3}$ for every k and thus $\alpha(f) \to 1 - \log 2/\log 3$ as $t \to \infty$. It is easy to see that also $\beta(f) \to 1 - \log 2/\log 3$. This follows at once from the following observation: for $t = 1$, if we switch the value of each leaf with probability p independently, then for small p the probability that the outcome will be switched is $(3/2)p + o(p)$.

Conjecture **6.1.** — There is an absolute constant $\beta_0 < 1/2$ (find it!) such that for every monotone Boolean function f, $\beta(f) \leqslant \beta_0$.

Late Remark **6.1.** — It was pointed out by Mossel and Peres by considering certain recursive majorities on larger trees that this conjecture is false.

6.3. *Number of runs*

We considered mainly monotone events. Here is an interesting noise stable non-monotone event. Given a string of n bits $x_1, x_2, ..., x_n$, let $R(x_1, x_2, ...x_n)$ be the number of runs. Thus, R is one plus the number of pairs of consecutive variables with different values.

The event that $R(x_1, x_2, ...x_n)$ is larger than its median is noise stable. Indeed, write $y_i = x_i \oplus x_{i+1}$, $i = 1, ..., n-1$, and note that the y_i's are independent, and R is just the majority event on the $y_i's$. (Here $\oplus$ is addition mod 2; that is, xor.)

6.4. *Majority of triangles*

We considered only the case where p is a constant. When p tends to zero with n, new phenomena occur. Consider, for example, random graphs on n vertices with edge probability $p = n^{-a}$, $a > 0$ and the event that the number of triangles in the graph is larger than its median. This is a noise stable event but its correlation with majority (or any weighted majority) tends to 0 as n tends to infinity.

7. Relations with complexity theory

There is an interesting connection between the complexity of boolean functions and the notions studied in this paper.

7.1. *AC0 and influences*

An important complexity class AC0 of Boolean functions are those which can be expressed by Boolean circuits of polynomial size (in the number of variables) and bounded depth. Boppana [9] proved that if f is expressed by a depth-c circuit of size N then

$$I(f) \leqslant C_1 \log^{c-1} N. \tag{7.1}$$

Earlier, Linial, Mansour and Nisan [27] proved that the Fourier coefficients of functions which can be expressed by Boolean circuits of polynomial (or quasi-polynomial) size and bounded depth in AC0 decays exponentially above poly-logarithmic "frequencies". Both these results rely on the fundamental Håstad Switching Lemma, see [21, 2].

Recall that a **monotone circuit** is one where all the gates are monotone increasing in the inputs; i.e., there are no "not" gates. The Håstad lemma for monotone boolean circuits is easier and was proved already by Boppana [8].

We conjecture that a reverse relation to 7.1 also holds.

Conjecture **7.1** *(Reverse Håstad).* — For every $\varepsilon > 0$ there is a $K = K(\varepsilon) > 0$ satisfying the following. For every monotone $\mathcal{A} \subset \Omega_n$, there is a $\mathcal{B} \subset \Omega_n$ such that $\mathbf{P}[\mathcal{A} \triangle \mathcal{B}] < \varepsilon$ and $\mathcal{B}$ can be expressed as a Boolean circuit such that

$$(\log N)^{c-1} < KI(\mathcal{A}),$$

where c and N are the depth and size of the circuit, respectively.

Monotone Boolean functions with bounded influence were characterized by Friedgut [16, 17]. The results of [11] are also relevant to this conjecture.

Ha Van Vu raised the question if there is a spectral way to distinguish between bounded depth circuits of polynomial size and bounded depth circuits of quasi-polynomial size. In particular, he was looking for a way to show that the graph property "having a clique of size $\log n$" for graphs with n vertices, cannot be expressed by a bounded depth circuit of polynomial size. (Here the set of variables correspond to the $\binom{n}{2}$ possible edges.)

Conjecture **7.2.** — Let $\varepsilon > 0$ be a fixed real number. Let $\mathcal{A}$ be a monotone property expressed by a depth-2 circuit of size M and let $f = \chi_{\mathcal{A}}$. Then there is a set $\mathcal{S}$ of polynomial size in M (where the polynomial depends on c and ε) so that

$$\sum\{\hat{f}^2(S) : S \notin \mathcal{S}\} \leqslant \varepsilon.$$

This conjecture may also apply to TC0, see below. It would be of great interest to characterize Boolean functions for which most of the weight of the Fourier coefficients is concentrated on a set of polynomial size in n.

7.2. *TC0 and noise sensitivity*

Noise sensitivity seems related to another class of boolean functions - threshold circuits of bounded depths see [36, 22]. In a threshold circuit each gate is a weighted majority function. For the study of spectral properties of signs of low degree polynomials see Bruck [12] and Bruck and Smolensky [13].

Conjecture **7.3.** — Let f be a boolean function given by a monotone threshold circuit of depth c and size M. Then

$$J(f) = O(1)(\log M)^{c-1}. \tag{7.2}$$

Thus, for $1/\varepsilon \leqslant O(1)(\log M)^{c-1}$ we expect that $\mathrm{var}(f, \varepsilon)$ is bounded away from zero. Also here it is a tempting conjecture that a reverse relation holds.

We conjecture further that all functions f that can be expressed by a depth-c monotone threshold circuit where all the threshold gates are balanced are uniformly stable. (And in particular, $J(f) = O(1)$.) Possibly, functions in this class of functions approximate arbitrary well arbitrary uniform stable monotone Boolean functions.

Conjecture 7.3 implies theorems of Yau [36] and Håstad and Goldmann [22]. They proved that the and/or tree (or equivalently the example of ternary tree of Section 6) does not belong to monotone TC0; i.e., it cannot be expressed as a monotone bounded depth circuit of polynomial size.

The results of Yau and Håstad are still open for the non-monotone case. This would follow if relation 7.2 holds even for every monotone boolean function f given by a (general) threshold circuit of depth c and size M.

8. Random walks

For nonempty $\mathscr{A} \subset \Omega_n$, consider a random walk defined as follows: start with a point chosen at random uniformly from $\mathscr{A}$, and at each step, stay where you are with probability $1/2$, and with probability $1/(2n)$ move to any one of the neighboring vertices. Let $\mathbf{P}^t_{\mathscr{A}}$ be the measure on Ω_n given by the location of the walk after t steps, and set

$$W(\mathscr{A}, \varepsilon) := \inf\{t : \|\mathbf{P}^t_{\mathscr{A}} - \mathbf{P}\| < \varepsilon\}.$$

Here $\|\mathbf{P}^t_{\mathscr{A}} - \mathbf{P}\|$ is the measure (L^1) norm of the difference between $\mathbf{P}^t_{\mathscr{A}}$ and the uniform measure.

Theorem **8.1.** — *Suppose that $\mathscr{A}_m \subset \Omega_{n_m}$ is a sequence of events satisfying* $\inf_m \mathbf{P}[\mathscr{A}_m] > 0$.

1. $\{\mathscr{A}_m\}$ *are asymptotically noise sensitive iff* $\lim_m W(\mathscr{A}_m, \varepsilon)/n_m = 0$ *for every fixed* $\varepsilon > 0$.
2. *If* $\beta(\mathscr{A}_m) \to \beta$, *then* $W(\mathscr{A}_m, \varepsilon) \leqslant n^{1-\beta-o(1)}$.

Proof. — Set $f_t(x) := 2^{n_m}\mathbf{P}^t_{\mathscr{A}_m}[\{x\}]$. Note that

$$f_{t+1} = (1/2) f_t + (2n_m)^{-1} \sum_{j=1}^{n_m} \sigma_j f_t.$$

Consequently,

$$\hat{f}_t(S) = \left(\frac{2n_m - |S|}{2n_m}\right)^t \hat{f}_0(S) = \mathbf{P}[\mathscr{A}_m]^{-1} \left(\frac{2n_m - |S|}{2n_m}\right)^t \hat{\chi}_{\mathscr{A}_m}(S).$$

This gives for every $k = 1, 2, \ldots$

$$\|\mathbf{P}^t_{\mathscr{A}_m} - \mathbf{P}\|^2 \leqslant \|f_t - 1\|_2^2 = \mathbf{P}[\mathscr{A}_m]^{-1} \sum_{0 \neq s \subset \Omega_{n_m}} \left(\frac{2n_m - |s|}{2n_m}\right)^{2t} \widehat{\chi}_{\mathscr{A}_m}(s)^2$$

$$\leqslant \mathbf{P}[\mathscr{A}_m]^{-1} \left(\frac{2n_m - k}{2n_m}\right)^{2t} + \sum_{0<|s|<k} \widehat{\chi}_{\mathscr{A}_m}(s)^2$$

$$\leqslant \mathbf{P}[\mathscr{A}_m]^{-1} \exp(-tk/n_m) + \sum_{0<|s|<k} \widehat{\chi}_{\mathscr{A}_m}(s)^2.$$

The theorem follows. □

9. Changing a fixed size set of bits

The noise operator N_ε changes every input variable independently of the others, and the expected number of bits changed is εn, where n is the number of variables. Understanding the effect of different types of noise may be of interest. We consider a variant where a fixed number of bits are changed. In other words, for $x \in \Omega_n$ and $q \in [n]$, let $\widetilde{\mathrm{N}}_q(x) = x \oplus s$, where s is chosen randomly uniformly among $s \in \Omega_n$ with cardinality q, independent from x. Here $\oplus$ is addition mod 2; that is, xor.

The analysis of the noise $\widetilde{\mathrm{N}}_q$ is similar to that of N_ε, but a little care is needed. Consider the following example. Let $\mathscr{P} \subset \Omega_n$ consist of those $x \in \Omega_n$ such that $|x|$ is odd. This event $\mathscr{P}$ is called **parity**. Observe that for each fixed q, the conditioned probability $\mathbf{P}[\widetilde{\mathrm{N}}_q(x) \in \mathscr{P} | x = y]$ is either zero or 1. In other words, knowing x allows a perfect prediction for $\widetilde{\mathrm{N}}_q(x) \in \mathscr{P}$. Note that $\widehat{\chi}_{\mathscr{P}}(\mathrm{S})$ is nonzero only when $\mathrm{S} \in \{\emptyset, [n]\}$. This means that the vanishing of the weight of the lower Fourier coefficients does not imply sensitivity to $\widetilde{\mathrm{N}}_q$, as in Theorem 1.9.

For $f: \Omega_n \to \mathbf{R}$ and $q \in [n]$ set

$$\widetilde{\mathrm{var}}(f, q) = \mathrm{var}_y\Big(\mathbf{E}[f(\widetilde{\mathrm{N}}_q(x)) \mid x = y]\Big) = \mathbf{E}_y\mathbf{E}\Big(f(\widetilde{\mathrm{N}}_q(x)) \mid x = y\Big)^2 - (\mathbf{E}f)^2.$$

We say that a sequence of events $\mathscr{A}_m \subset \Omega_{n_m}$ is **asymptotically noise sensitive with respect to** $\widetilde{\mathrm{N}}$ if for every $\varepsilon \in (0, 1)$ and every sequence $\{q_m\}$ with $\varepsilon n_m \leqslant q_m \leqslant (1 - \varepsilon) n_m$, we have

$$\lim_m \widetilde{\mathrm{var}}(\mathscr{A}_m, q_m) = 0.$$

Note that this is equivalent to the straightforward analog of the definition for asymptotic noise sensitivity to our current setting.

Theorem **9.1.** — *Let $\mathcal{B}_m \subset \Omega_{n_m}$ be a sequence of events, and set $g_m = \chi_{\mathcal{B}_m}$.*

1. *This sequence is asymptotically noise sensitive with respect to $\tilde{\mathrm{N}}$ iff for every finite k*

$$\lim_m \sum \left\{ \hat{g}_m(\mathrm{S})^2 : \mathrm{S} \subset [n],\ 1 \leqslant |\mathrm{S}| \leqslant k \text{ or } |\mathrm{S}| \geqslant n-k \right\} = 0.$$

2. *A sufficient condition for asymptotic noise sensitivity is $\mathbf{H}(\mathcal{B}_m) \to 0$.*

Proof. — For $f: \Omega_n \to \mathbf{R}$ set

$$\tilde{\mathrm{T}}_q f(y) = \mathbf{E} f(\tilde{\mathrm{N}}_q(y)).$$

We now compute the Fourier coefficients of $\tilde{\mathrm{T}}_q f$. Take $r \in \Omega_n$.

$$\begin{aligned}
\mathbf{E}(\tilde{\mathrm{T}}_q f \cdot u_r) &= 2^{-n} \sum_x \tilde{\mathrm{T}}_q f(x)(-1)^{|r \cap x|} \\
&= 2^{-n} \binom{n}{q}^{-1} \sum_x \sum_{|s|=q} f(x \oplus s)(-1)^{|r \cap x|} \\
&= 2^{-n} \binom{n}{q}^{-1} \sum_y \sum_{|s|=q} f(y)(-1)^{|r \cap y|}(-1)^{|r \cap s|} \\
&= \hat{f}(r) \binom{n}{q}^{-1} \sum_j (-1)^j \binom{|r|}{j} \binom{n-|r|}{q-j}.
\end{aligned}$$

Consequently,

$$\tilde{\mathrm{T}}_q f = \sum_{r \in \Omega_n} c(n, q, |r|) \hat{f}(r) u_r, \tag{9.1}$$

where

$$c(n, q, k) = \binom{n}{q}^{-1} \sum_j (-1)^j \binom{k}{j} \binom{n-k}{q-j}.$$

Since $c(n, q, 0) = 1$, this gives,

$$\widetilde{\mathrm{var}}(f, q) = \|\mathrm{T}_q f\|_2^2 - \hat{f}(\emptyset)^2 = \sum_{\emptyset \neq \mathrm{S} \subset [n]} c(n, q, |\mathrm{S}|)^2 \hat{f}(\mathrm{S})^2. \tag{9.2}$$

Consequently, for 9.1.1 it is enough to understand the behavior of the coefficients $c(n, q, k)$. For this, consider the sequence

$$a_j = \binom{k}{j} \binom{n-k}{q-j} \binom{n}{q}^{-1}.$$

The sequence has a unique maximum, which occurs when j is an integer j' close to qk/n. Consequently, $c(n, q, k) \leqslant 2a_{j'}$. Now let $n, k, q \to \infty$, and assume that $\varepsilon n \leqslant q \leqslant (1-\varepsilon)n$ and $n - k \to \infty$, where $\varepsilon > 0$ is fixed. Then

$$\lim c(n, q, k) = 0.$$

This gives one direction in 9.1.1.

Also note that when $q < n/(3k)$, $|c(n, q, k)|$ is approximately a_0. This gives $\liminf |c(n, q, k)| > 0$ when k is fixed, $n \to \infty$ and q is about $n/3k$. Since $c(n, q, k) = \pm\, c(n, q, n-k)$, we get the other direction of 1.

Now assume that $\mathbf{H}(\mathscr{C}_m) \to 0$. From Theorem 1.9 we know that

$$\lim_m \sum \left\{ \widehat{g}_m(\mathrm{S})^2 : \mathrm{S} \subset [n],\ 1 \leqslant |\mathrm{S}| \leqslant k \right\} = 0$$

for every fixed k. Equation (2.14) gives,

$$\mathbf{H}(\mathscr{C}_m) \geqslant \frac{n_m}{2} \sum \left\{ \widehat{g}_m(\mathrm{S})^2 \widehat{g}_m(\mathrm{S}')^2 : \mathrm{S}, \mathrm{S}' \subset [n],\ |\mathrm{S}|, |\mathrm{S}'| \geqslant 3n_m/4 \right\}$$

$$\geqslant \left(\sum_{|\mathrm{S}| \geqslant 3n_m/4} \widehat{g}_m(\mathrm{S})^2 \right)^2 .$$

Consequently,

$$\lim_m \sum \left\{ \widehat{g}_m(\mathrm{S})^2 : \mathrm{S} \subset [n],\ |\mathrm{S}| \geqslant n - k \right\} = 0$$

and the proof is complete. □

REFERENCES

[1] J. Ambjorn, B. Durhuus and T. Jonsson, *Quantum Geometry*, Cambridge University Press, Cambridge, 1997.

[2] N. Alon and J. Spencer, *The Probabilistic Method*, Wiley, New York (1992).

[3] W. Beckner, Inequalities in Fourier analysis, *Annals of Math.* **102** (1975), 159-182.

[4] M. Ben-or and N. Linial, Collective coin flipping, in *Randomness and Computation* (S. Micali, ed.), Academic Press, New York (1990), pp. 91-115. Earlier version: Collective coin flipping, robust voting games, and minima of Banzhaf value, Proc. 26th IEEE Symp. on the Foundation of Computer Science (1985), 408-416.

[5] I. Benjamini and O. Schramm, Conformal invariance of Voronoi percolation, *Commun. Math. Phys.*, **197** (1998), 75-107.

[6] I. Benjamini, G. Kalai and O. Schramm, Noise sensitivity, concentration of measure and first passage percolation, in preparation.

[7] A. Bonami, Etude des coefficients Fourier des fonctions de $L^p(G)$, *Ann. Inst. Fourier*, **20** (1970), 335-402.

[8] R. Boppana, Threshold functions and bounded depth monotone circuits, *Proceedings of 16th Annual ACM Symposium on Theory of Computing* (1984), 475-479.

[9] R. Boppana, The average sensitivity of bounded depth circuits, *Inform. Process. Lett.* **63** (1997) 257-261.

[10] J. Bourgain, J. Kahn, G. Kalai, Y. Katznelson and N. Linial, The influence of variables in product spaces, *Isr. J. Math.* **77** (1992), 55-64.

[11] J. Bourgain and G. Kalai, Influences of variables and threshold intervals under group symmetries, *Geom. Funct. Anal.*, **7** (1997), 438-461.

[12] J. Bruck, Harmonic analysis of polynomial threshold functions. SIAM J. Discrete Math. **3** (1990), 168-177.

[13] J. Bruck and R. Smolensky, Polynomial threshold functions, AC^0 functions, and spectral norms. SIAM J. Comput. **21** (1992), 33-42.

[14] A. Bunde and S. Havlin (ed.s'), *Fractals and Disordered Systems*, Springer 1991.

[15] J. T. Chayes, L. Chayes, D. S. Fisher and T. Spencer, Finite-size scaling and correlation length for disordered systems, *Phys. Rev. Lett.* **57** (1986), 2999-3002.

[16] E. Friedgut, Boolean functions with low average sensitivity, *Combinatorica* **18** (1998), 27-36.

[17] E. Friedgut, Necessary and sufficient conditions for sharp thresholds of graphs properties and the k-sat problem, *Jour. Amer. Math. Soc.* **12** (1999), 1017-1054.

[18] E. Friedgut and G. Kalai, Every monotone graph property has a sharp threshold, *Proc. Amer. Math. Soc.* **124** (1996), 2993-3002.

[19] G. Grimmett, *Percolation*, Springer-Verlag, Berlin (1989).

[20] O. Haggstrom, Y. Peres and J. E. Steif, Dynamical percolation, *Ann. IHP* **33** (1997), 497-528.

[21] J. Håstad, Almost optimal lower bounds for small depth circuits, in *Randomness and Computation*, **5**, ed. S. Micali, (1989), 143-170.

[22] J. Håstad and M. Goldmann, On the power of small-depth threshold circuits, *Computational Complexity*, **1** (1991), 113-129.

[23] J. Kahn, G. Kalai and N. Linial, The influence of variables on boolean functions, *Proc. 29-th Ann. Symp. on Foundations of Comp. Sci.*, (1988), 68-80.

[24] H. Kesten, Scaling relations for 2*D*-percolation, *Comm. Math. Phys.* **109** (1987), 109-156.

[25] H. Kesten and Y. Zhang, Strict inequalites for some critical exponents in 2D-percolation. *J. Statist. Phys.* **46** (1987), 1031-1055.

[26] R. P. Langlands, P. Pouliot and Y. Saint-aubin, Conformal invariance in two-dimensional percolation, *Bull. Amer. Math. Soc. (N.S.)* **30** (1994), 1-61.

[27] N. Linial, Y. Mansour and N. Nisan, Constant depth circuits, Fourier transform, and learnability, *J. Assoc. Comput. Mach.* **40** (1993), 607-620.

[28] V. V. Petrov, *Limit theorems of probability theory*, Oxford University Press, (1995).

[29] L. Russo, A note on percolation, *ZW.* **43** (1978), 39-48.

[30] P. Seymour and D. Welsh, Percolation probabilities on the square lattice. Advances in Graph Theory. *Ann. Discrete Math.* **3** (1978), 227-245.

[31] M. Talagrand, On Russo's approximate zero-one law, *Ann. of Prob.* **22** (1994), 1576-1587.

[32] M. Talagrand, Concentration of measure and isoperimetric inequalities in product spaces, *Publ. I.H.E.S.*, **81** (1995), 73-205.

[33] M. Talagrand, How much are increasing sets positively correlated? *Combinatorica* **16** (1996), 243-258.

[34] B. Tsirelson, Fourier-Walsh coefficients for a coalescing flow (discrete time), preprint, math.PR/9903068.

[35] B. Tsirelson, The Five noises, preprint.

[36] A. Yao, Circuits and local computation, *Proceedings of 21st Annual ACM Symposium on Theory of Computing*, (1989), 186-196.

I. B.
The Weizmann Institute of Science,
Rehovot 76100, Israel
itai@wisdom.weizmann.ac.il
http://www.wisdom.weizmann.ac.il/~itai/

G. K.
The Hebrew University, Givat Ram,
Jerusalem 91904, Israel
kalai@math.huji.ac.il
http://www.ma.huji.ac.il/~kalai/

O. S.
The Wiezmann Institute of Science,
Rehovot 76100, Israel
schramm@wisdom.weizmann.ac.il
http://www.wisdom.weizmann.ac.il/~schramm/

Manuscrit reçu le 4 janvier 1999.

Annals of Mathematics, **171** (2010), 619–672

Quantitative noise sensitivity and exceptional times for percolation

By Oded Schramm and Jeffrey E. Steif

Abstract

One goal of this paper is to prove that dynamical critical site percolation on the planar triangular lattice has exceptional times at which percolation occurs. In doing so, new *quantitative* noise sensitivity results for percolation are obtained. The latter is based on a novel method for controlling the "level k" Fourier coefficients via the construction of a randomized algorithm which looks at random bits, outputs the value of a particular function but looks at any fixed input bit with low probability. We also obtain upper and lower bounds on the Hausdorff dimension of the set of percolating times. We then study the problem of exceptional times for certain "k-arm" events on wedges and cones. As a corollary of this analysis, we prove, among other things, that there are no times at which there are two infinite "white" clusters, obtain an upper bound on the Hausdorff dimension of the set of times at which there are both an infinite white cluster and an infinite black cluster and prove that for dynamical critical bond percolation on the square grid there are no exceptional times at which three disjoint infinite clusters are present.

1. Introduction
2. Noise sensitivity of algorithmically dilute functions
   2.1. Noise sensitivity background
   2.2. Proof of Theorem 1.8
3. Percolation background and notation
4. Noise sensitivity for percolation
   4.1. The simply connected case
   4.2. Annulus case
5. Exceptional times

Research supported by the Swedish Research Council and the Göran Gustafsson Foundation (KVA).

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_13, © Springer Science+Business Media, LLC 2011

6. Hausdorff dimension of exceptional times
7. Exceptional times for k-arm events
8. Upper bounds for k-arm times
9. The square lattice
10. Some open questions
Appendix A. Quasi-multiplicativity
References

1. Introduction

Consider bond percolation on an infinite connected locally finite graph G, where for some $p \in [0,1]$, each edge (bond) of G is, independently of all others, open with probability p and closed with probability $1-p$. Write π_p for this product measure. The main questions in percolation theory (see [12]) deal with the possible existence of infinite connected components (clusters) in the random subgraph of G consisting of all sites and all open edges. Write $\mathscr{C}$ for the event that there exists such an infinite cluster. By Kolmogorov's 0-1 law, the probability of $\mathscr{C}$ is, for fixed G and p, either 0 or 1. Since $\pi_p(\mathscr{C})$ is nondecreasing in p, there exists a critical probability $p_c = p_c(G) \in [0,1]$ such that

$$\pi_p(\mathscr{C}) = \begin{cases} 0 & \text{for } p < p_c \\ 1 & \text{for } p > p_c. \end{cases}$$

At $p = p_c$ we can have either $\pi_p(\mathscr{C}) = 0$ or $\pi_p(\mathscr{C}) = 1$, depending on G.

Häggström, Peres and Steif [13] initiated the study of dynamical percolation. (The notion of dynamical percolation was invented independently by I. Benjamini. While the present paper was motivated by [13], the question studied here had previously been asked by Benjamini, as we recently became aware.) In this model, with p fixed, the edges of G switch back and forth according to independent 2 state continuous time Markov chains where closed switches to open at rate p and open switches to closed at rate $1-p$. Clearly π_p is a stationary distribution for this Markov process. The general question studied in [13] was whether, when we start with distribution π_p, there could exist atypical times at which the percolation structure looks markedly different from that at a fixed time.

Write Ψ_p for the underlying probability measure of this Markov process, and write $\mathscr{C}_t$ for the event that there is an infinite cluster of open edges at time t.

Two results in [13] which are relevant to us are

PROPOSITION 1.1. *For any graph G,*

$$\begin{cases} \Psi_p(\mathscr{C}_t \text{ occurs for every } t) = 1 & \text{if } p > p_c(G), \\ \Psi_p\big((\neg\mathscr{C}_t) \text{ occurs for every } t\big) = 1 & \text{if } p < p_c(G). \end{cases}$$

THEOREM 1.2. *For $d \geq 19$, the integer lattice $\mathbb{Z}^d$ satisfies*

$$\Psi_{p_c}\big((\neg\mathscr{C}_t) \text{ occurs for every } t\big) = 1.$$

One important aspect of the proof of the latter result is that it uses the fact, proved in [14], that for $d \geq 19$,

$$\pi_p(0 \text{ is in an infinite open cluster}) = O(|p - p_c|). \tag{1.1}$$

It is proved in [21] that (1.1) does not hold for $d = 2$. Therefore, the question of whether Theorem 1.2 is true for $d = 2$ becomes interesting. At this point, we mention that site percolation is the analogous model where the vertices (rather than the edges) are open or closed independently, each with probability p, and dynamical percolation is defined in a completely analogous manner. Our main result says that Theorem 1.2 does not hold for site percolation on the planar triangular grid. The triangular grid is the graph whose vertex set is the subset of $\mathbb{C} = \mathbb{R}^2$ consisting of the points

$$\mathbb{Z} + \exp(2\pi i/3)\,\mathbb{Z} = \big\{(k + \ell/2, \sqrt{3}\,\ell/2) : k, \ell \in \mathbb{Z}\big\}$$

and two such points have an edge between them if and only if their distance is 1. Explicitly stated, our main result is

THEOREM 1.3. *Almost surely, the set of times $t \in [0, 1]$ such that dynamical critical site percolation on the triangular lattice has an infinite open cluster is nonempty.*

There are no other transitive graphs for which it is known that dynamical critical percolation has such exceptional times. (In [13], it was argued that the event discussed in Theorem 1.3 is measurable. A similar comment applies to our other results below. Thus, measurability issues will not concern us here.)

We are convinced that Theorem 1.3 is true for bond percolation on the square lattice. However, our proof uses the existence and exact values of certain so-called *critical exponents*, which are only known to hold for site percolation on the triangular lattice. These are believed to be the same for bond percolation on the square lattice, but even their existence has not yet been established in that case. However, the methods of this paper seem to come quite close to a proof for the square grid as well: it seems that there are several ways in which this can perhaps be achieved without determining these critical exponents. These issues will be further discussed in Section 9.

It is interesting to note that by [13, Cor. 4.2], a.s. at every time t the set of vertices that are contained in some infinite cluster has zero density.

On a heuristic level, for Theorem 1.3 to hold, it is necessary that the configuration "changes fast" in order to have "many chances" to percolate so that we will in fact have a percolating time. Mathematically, "changing fast" can be interpreted as

having small correlations over short time intervals, which then suggests the use of the second moment method which we indeed will use. In other words, one needs to know that the configuration at a given time tells us almost nothing about how it will look a short time later. The notion of *noise sensitivity* introduced in [2] is the relevant tool which describes this phenomenon. We now briefly explain this.

Given an integer m, a subset A of $\{0,1\}^m$ and an $\varepsilon > 0$, define

$$N(A,\varepsilon) := \mathrm{var}\Big[\mathbf{P}\big[(Y_1,\ldots,Y_m) \in A \;\big|\; X_1,\ldots,X_m\big]\Big]$$

where $\{X_i\}_{1\le i\le m}$ are i.i.d. with $\mathbf{P}\big[X_i = 1\big] = 1/2 = \mathbf{P}\big[X_i = 0\big]$ and conditional on the $\{X_i\}$'s, $\{Y_i\}_{1\le i\le m}$ are independent with $Y_i = X_i$ with probability $1-\varepsilon$ and $Y_i = 1 - X_i$ with probability ε.

Definition 1.4. Let $\{n_m\}_{m\ge 1}$ be an increasing sequence in $\mathbb{N}$ going to ∞ and let A_m be a subset of $\{0,1\}^{n_m}$ for each m. We say that the sequence $\{A_m\}_{m\ge 1}$ is *noise sensitive* if for every $\varepsilon > 0$,

$$\lim_{m\to\infty} N(A_m,\varepsilon) = 0. \tag{1.2}$$

This says that for large m knowing the values of $X_1,\ldots,X_{n_m}$ gives us almost no information concerning whether $(Y_1,\ldots,Y_{n_m}) \in A_m$. This is not the exact definition of noise sensitivity given in [2] but is easily shown to be equivalent; see page 14 in that paper. It is also shown in [2] that if (1.2) holds for some $\varepsilon \in (0,1/2)$, then it holds for all such ε and in addition that $N(A,\varepsilon)$ is decreasing in ε on $[0,1/2]$.

Let n_m be the number of edges in an $(m+1)\times m$ box in $\mathbb{Z}^2$ and let A_m be the event of a left to right crossing in such a box. By duality, $\mathbf{P}\big[A_m\big] = 1/2$ for every m (see [12]). In [2], the following result is proved.

THEOREM 1.5. *The sequence $\{A_m\}_{m\ge 1}$ is noise sensitive.*

A by-product of the tools needed to prove Theorem 1.3 will imply the following more quantitative version of Theorem 1.5, which was conjectured in [2].

THEOREM 1.6. *There exists $\gamma > 0$ so that*

$$\lim_{m\to\infty} N(A_m, m^{-\gamma}) = 0.$$

We have the same result for the triangular lattice but with a better γ, since critical exponents are known in this case.

THEOREM 1.7. *For critical site percolation on the triangular lattice, let A'_m be the event of the existence of a left-right crossing in a domain D approximating a square of sidelength m. Then for all $\gamma < 1/8$,*

$$\lim_{m\to\infty} N(A'_m, m^{-\gamma}) = 0.$$

In proving our quantitative noise sensitivity results (Theorems 1.6 and 1.7 as well as those later on necessary for obtaining Theorem 1.3), one of two key steps will be Theorem 1.8, which gives estimates of certain quantities involving Fourier coefficients of a function based on the properties of an algorithm calculating the function; the other key step will be the construction of an appropriate algorithm. Precise definitions of undefined terms will be given in Section 2, where the connection with noise sensitivity will also be recalled.

THEOREM 1.8. *Let $n \in \mathbb{N}$ and set $\Omega = \Omega_n := \{0,1\}^n$. Let $f : \Omega \to \mathbb{R}$ be a function. Suppose that there is a randomized algorithm A for determining the value of f which examines some of the input bits of f one by one, where the choice of the next bit examined may depend on the bits examined so far. Let $J \subseteq [n] := \{1,2,\dots,n\}$ be the (random) set of bits examined by the algorithm. Set $\delta = \delta_A := \sup\{\mathbf{P}[i \in J] : i \in [n]\}$. Then, for every $k = 1,2,\dots$, the Fourier coefficients of f satisfy*

$$\sum_{S\subseteq[n],\,|S|=k} \hat{f}(S)^2 \le \delta\, k\, \|f\|^2, \tag{1.3}$$

where $\|f\|$ denotes the L^2 norm of f with respect to the uniform probability measure on Ω.

This result might have some applications to theoretical computer science. We will call δ_A the *revealment* of the algorithm A. The restriction of x to J (the set of bits examined by the algorithm) is a *witness* for the function f, in the sense that it determines $f(x)$. As explained in Section 2.2, Theorem 1.8 extends to some other types of witnesses.

In the case $k = 1$, the inequality (1.3) cannot be improved by more than a factor of $O(1/\log n)$: there is an example showing this with $\delta \le n^{-1/3}\log(n)$, which appears in [4, §4]. The paper [4] investigates how small the revealment can be for a balanced boolean function on $\{0,1\}^n$. When the function is monotone, it is shown that the revealment cannot be much smaller than $n^{-1/3}$ and in general it cannot be much smaller than $n^{-1/2}$. Examples are given there which come within logarithmic factors of meeting these bounds.

We do not know if (1.3) is close to being optimal for $k \gg 1$. One is tempted to speculate that the inequality can be improved to $\sum_{|S|\le k} \hat{f}(S)^2 \le O(1)\, k\, \delta\, \|f\|^2$. We do not know any counterexample to this inequality. However, the AND function $f(x) = \prod_{j=1}^n x_j$ gives an example where

$$O(1) \sum_{|S|=k} \hat{f}(S)^2 \ge \sqrt{k}\, \delta\, \|f\|^2$$

for k satisfying $|k - n/2| = O(n^{1/2})$. (It is easy to check that the best revealment possible for this f is exactly $(2-2^{1-n})/n$.)

Once Theorem 1.3 is established, it is natural to ask: how large is the set of "exceptional" times at which percolation occurs? In this direction, we have the following result.

THEOREM 1.9. *The Hausdorff dimension of the set of times at which dynamical critical site percolation on the triangular lattice has an infinite cluster is an almost sure constant which lies in* $[\frac{1}{6}, \frac{31}{36}]$.

We conjecture that $\frac{31}{36}$ is the correct answer. In a different direction, once we know that there are exceptional times at which percolation occurs, it is natural to ask how many clusters can exist at these exceptional times. The following provides the answer.

THEOREM 1.10. *On the triangular lattice, a.s. there are no times at which dynamical critical site percolation has two or more infinite open clusters.*

For the square grid, we can only prove

THEOREM 1.11. *On* $\mathbb{Z}^2$, *a.s. there are no times at which dynamical critical bond percolation has three or more infinite open clusters.*

In some of the figures, we will represent open sites by white hexagons on the dual grid, and closed sites by black hexagons. Thus, percolation clusters correspond to connected components of the union of the white hexagons. These will also be called white clusters. Likewise, we may also consider black clusters, which are connected components of black hexagons.

Asking whether two infinite white clusters can coexist at some time is very different from asking whether two infinite clusters of *different* colors can coexist at some time. We conjecture that there are in fact exceptional times at which there are both a white and a black infinite cluster and that the Hausdorff dimension of such times is $2/3$. We can however prove the following.

THEOREM 1.12. *On the triangular lattice, a.s. the Hausdorff dimension of the set of times at which there are both an infinite white cluster and an infinite black cluster is at most* $2/3$.

We also have the following two results concerning the upper half-plane.

THEOREM 1.13. *On the triangular lattice intersected with the upper half-plane, a.s. the Hausdorff dimension of the set of times at which there is an infinite cluster is at most* $5/9$.

THEOREM 1.14. *On the triangular lattice intersected with the upper half-plane, a.s. the set of times at which there are both an infinite white cluster and an infinite black cluster is empty.*

Theorems 1.10, 1.12, 1.13 and 1.14 will follow immediately from generalizations presented in the last part of the paper, which are concerned with studying dynamical percolation on two other two-dimensional objects, namely wedges and cones. For every $\theta \in (0, \infty)$, we let W_θ denote the wedge of angle θ and C_θ denote the cone of angle θ. For C_θ, we will require that θ is a multiple of $\pi/3$. The precise definitions of these will be given in Section 3. First, we mention that for all θ, the critical value for site percolation on W_θ and on C_θ is $1/2$, as for site percolation on the triangular grid and bond percolation on $\mathbb{Z}^2$.

The following results provide upper and lower bounds on the critical angle for which there are exceptional times for certain k-arm type events as well as provide estimates for the Hausdorff dimension of the set of exception times for a given angle. In these results, if an upper bound on the Hausdorff dimension is negative, this means that the set in question is empty.

We will only do the case where the arms are alternating in color (and hence for the case of cones, there will be one or an even number of arms). We do this partially because it is easier than the general case and because it is all that is needed in order to make statements concerning the number of infinite clusters.

By a k-arm event, we mean an event of the form "there are k disjoint infinite paths having a specified color sequence"; for a wedge, the color sequence is well-defined while for a cone, it is well-defined up to cyclic permutations.

THEOREM 1.15. *Fix the wedge W_θ and for integer $k \geq 1$, let $A^k_{W_\theta}$ be the event that there are k infinite disjoint paths in W_θ whose colors alternate. Then a.s. the Hausdorff dimension, $H^k_{W_\theta}$, of the set of exceptional times at which $A^k_{W_\theta}$ occurs satisfies*

$$1 - \frac{4k(k+1)\pi}{3\theta} \leq H^k_{W_\theta} \leq 1 - \frac{2k(k+1)\pi}{9\theta}.$$

In particular, for any $k \geq 1$, there are exceptional times for the event $A^k_{W_\theta}$ for $\theta > \frac{4k(k+1)\pi}{3}$ and there are no exceptional times for $\theta < \frac{2k(k+1)\pi}{9}$.

THEOREM 1.16. *Fix the cone C_θ with θ a multiple of $\pi/3$ and let, for $k = 1$ or $k > 1$ even, $A^k_{C_\theta}$ be the event that there are k infinite disjoint paths in C_θ whose colors alternate (if $k > 1$). Then a.s. the Hausdorff dimension, $H^k_{C_\theta}$, of the set of exceptional times at which $A^k_{C_\theta}$ occurs satisfies*

$$1 - \frac{5\pi}{3\theta} \leq H^1_{C_\theta} \leq 1 - \frac{5\pi}{18\theta}$$

and for $k \geq 2$

$$1 - \frac{4(k^2-1)\pi}{3\theta} \leq H^k_{C_\theta} \leq 1 - \frac{2(k^2-1)\pi}{9\theta}.$$

In particular, for $k = 1$, there are exceptional times for the event $A^1_{C_\theta}$ for $\theta > \frac{5\pi}{3}$ and there are no exceptional times for $\theta < \frac{5\pi}{18}$, while for $k \geq 2$, there are exceptional times for the event $A^k_{C_\theta}$ for $\theta > \frac{4(k^2-1)\pi}{3}$ and there are no exceptional times for $\theta < \frac{2(k^2-1)\pi}{9}$.

Theorem 1.16 is presumably true for other values of θ provided that a proper definition of C_θ is given.

Remark. One should note that the upper bounds on the Hausdorff dimension given in Theorems 1.9,1.12,1.13, 1.15 and 1.16 are all of the form $1-(4/3)\xi$ where ξ is the critical exponent for the given event.

There is an abstract theory of Lévy processes on groups [16], [9], which gives a criterion for a Lévy process (such as ω_t) to hit a set A (such as the set of configurations which contain an infinite component). Basically, to show that A is hit, one needs to prove that there exists a probability measure μ on A which has $\|\mu\|_* < \infty$ for an appropriate Hilbert norm $\|\cdot\|_*$, based on the Fourier transform. It seems that we could use this framework in the present paper, but that would not essentially simplify the core issues we deal with. Moreover, it seems that our hands-on approach facilitates some generalizations, which the Lévy process theory does not cover, which brings us to our next remark.

The fact that the time between flips has an exponential distribution is not really essential here, and the results apply in greater generality. Let $\omega_t(v)$ denote the indicator function for the event that at time t the site v is white. Basically, all of the results concerning existence of exceptional times and lower bounds on Hausdorff dimension go through (with essentially the same proofs) in the more general setting where we assume that

(i) The processes $t \mapsto \omega_t(v)$ are independent (possibly with different distributions depending on v) as v runs over all sites.

(ii) $\mathbf{P}[\omega_t(v) = 1] = 1/2$ for all t and v.

(iii) There is $c > 0$ so that

$$\left|\mathbf{E}\left[(-1)^{\omega_t(v)}(-1)^{\omega_s(v)}\right]\right| \leq 1 - c\,|t-s|$$

for all v and all t and s satisfying $|t-s| < c$.

(iv) For each v, the process $\omega_t(v)$ has right continuous paths a.s.

(Condition (iv) is just a technical condition to insure that the events that we consider are measurable.)

For results concerning upper bounds on Hausdorff dimension, the proofs go through when we assume (i), (ii), (iv) and

(v) There is a $c > 0$ so that

$$\mathbf{E}[\text{number of flips of } \omega_t(v) \text{ during } (t_1, t_2)] \leq c(t_2 - t_1)$$

for all v and all $t_1, t_2 \in \mathbb{R}$ satisfying $t_1 < t_2 < t_1 + c$.

However, for simplicity, we stick to the original setup.

We mention a few other papers where analogous questions to those studied in [13] have been studied for other models. First, the results in [13] were extended and refined in [26]. Next, analogous questions for the Boolean model where the points undergo independent Brownian motions were studied in [5]. Analogous questions for the lattice case for certain interacting particle systems (where updates are not done in an independent fashion) are studied in [6]. Finally in [3], it is shown that there are exceptional two-dimensional slices for the Boolean model in four dimensions.

The rest of the paper is organized as follows. In Section 2, we will first provide background on the Fourier-Walsh expansion of a function defined on $\{0,1\}^n$ as well as connections with noise sensitivity and then continue on to give the proof of Theorem 1.8 as well as a generalization to the case where the algorithm is not required to always determine the value of the function f. (This will accommodate readers who are only interested in Theorem 1.8.) In Section 3, we will give necessary background concerning percolation including a discussion of critical site percolation on the triangular lattice as well as a brief discussion of interfaces and critical exponents. In Section 4, we will construct two algorithms determining certain events involving critical site percolation on the triangular lattice and analyze them to obtain upper bounds on the probability that a vertex is looked at during the algorithm. (For readers who only want to read Theorem 1.8 and see how to apply it, they can just glance through §3 and then read §4.)

Section 4 also gives a very detailed discussion of interfaces and completes the proofs of Theorems 1.6 and 1.7 by application of Theorem 1.8. In Section 5, we give the proof of Theorem 1.3, and in Section 6 we give the proof of Theorem 1.9. (Although the upper bound of $31/36$ given in Theorem 1.9 is a special case of Theorem 1.16, we choose to give a different direct proof of this without reference to the work done in §8.) In Section 7, we prove the lower bounds on the Hausdorff dimension stated in Theorems 1.15 and 1.16. In Section 8, we prove the upper bounds on the Hausdorff dimension stated in Theorems 1.15 and 1.16. This will be based on a general formula which gives an upper bound on the Hausdorff dimension of various random sets (or proves they are empty) in terms of *influences* (Theorem 8.1). We conclude the section by showing that Theorems 1.10, 1.12, 1.13 and 1.14 immediately follow from Theorems 1.15 and 1.16. After this, in Section 9, we prove Theorem 1.11 and explain several plausible ways in which the

proof of Theorem 1.3 might be extended to bond percolation on the square lattice. In Section 10, we present some open questions.

Finally, the appendix proves some results about (nondynamical) critical percolation that are needed for Theorems 1.10–1.16. The main result is that if $r < r' < r''$, then the probability of having j crossings in a prescribed color sequence between distances r and r'' from 0 is equal to the product of the corresponding probabilities between radii r and r' and between radii r' and r'', times an error term (depending on j) that is bounded away from 0 and infinity. Another consequence is that one gets good control on the positions of the crossings at the inner and outer radii, as was already demonstrated by Kesten [19, Lemma 2]. The proofs in the appendix also establish the corresponding statements for critical bond percolation in $\mathbb{Z}^2$.

2. Noise sensitivity of algorithmically dilute functions

In this section, we give some background and then prove Theorem 1.8.

2.1. *Noise sensitivity background.* For a function f from $\Omega = \Omega_n := \{0,1\}^n$ to $\mathbb{R}$, the Fourier-Walsh expansion of f is given by $f = \sum_{S\subseteq[n]} \hat{f}(S)\chi_S$, where, $\chi_S(T) = (-1)^{|S\cap T|}$ and $\hat{f}(S) = \int f\chi_S$. Here and in the following, $\int$ refers to integration with respect to uniform measure and we identify any vector $x \in \Omega_n$ with the subset $\{j \in [n] : x_j = 1\}$ of $[n] = \{1, 2, \ldots, n\}$. Consequently, $|x|$ denotes the cardinality of that set; that is, $|x| = \|x\|_1$ for $x \in \Omega_n$. The $\{\chi_S\}_{S\subseteq[n]}$ are an orthonormal basis for the 2^n-dimensional vector space of functions from Ω_n to $\mathbb{R}$. In particular,

$$\|f\|^2 = \sum_{S\subseteq[n]} \hat{f}(S)^2.$$

We now generalize the definition of $N(A,\varepsilon)$ (given in the introduction) to any function $f : \Omega \to \mathbb{R}$ by defining

$$N(f,\varepsilon) := \mathrm{var}\Big[\mathbf{E}\big[f(Y_1,\ldots,Y_m)|X_1,\ldots,X_m\big]\Big].$$

It is easy to see that (see page 14 in [2])

$$N(f,\varepsilon) = \sum_{\varnothing\neq S\subseteq[n]} \hat{f}(S)^2(1-2\varepsilon)^{2|S|}. \tag{2.1}$$

This explains the importance of the Fourier-Walsh expansion in the study of noise sensitivity.

2.2. *Proof of Theorem 1.8.* Before giving the proof, we discuss some heuristics. One may first believe that an estimate such as (1.3) would be valid because when the algorithm terminates, the value of f is completely determined, and hence perhaps all the nonzero Fourier coefficients $\hat{f}(S) \neq 0$, $S \neq \varnothing$, must satisfy

$S \cap J \neq \varnothing$. However, this is easily shown not to be the case. At the t-th step of the algorithm, after t bits have been determined, we may consider a new function f_t, which is f with those determined bits substituted. If at the $(t+1)$-th step, the i-th bit ω_i of the input $\omega \in \Omega$ is examined, then in the passage from f_t to f_{t+1}, there is a collapsing of Fourier coefficients: $\hat{f}_{t+1}(S) = \hat{f}_t(S) + (-1)^{\omega_i} \hat{f}_t(S \cup \{i\})$ and $\hat{f}_{t+1}(S \cup \{i\}) = 0$ for every $S \subseteq [n] \setminus \{i\}$. Thus, the coefficient $\hat{f}_{t+1}(S)$ may vanish when some bit $i \in S$ is examined by time $t+1$ or when some $i \notin S$ is chosen at time $t+1$ and it happens that $\hat{f}_t(S) + (-1)^{\omega_i} \hat{f}_t(S \cup \{i\}) = 0$. The latter, which we call "collapsing from above", may seem like a highly nongeneric situation. However, we cannot rule it out because we are primarily interested in very nongeneric functions, namely, functions with values in $\{0, 1\}$. The proof below uses a simple decomposition argument to handle the possibility of collapsing from above.

In the following, we let $\widetilde{\Omega}$ denote the probability space that includes the randomness in the input bits of f and the randomness used to run the algorithm and we let $\mathbf{E}$ denote the corresponding expectation. Without loss of generality, elements of $\widetilde{\Omega}$ can be represented as $\widetilde{\omega} = (\omega, \tau)$ where ω are the random bits and τ represents the randomness necessary to run the algorithm.

Proof. Fix $k \geq 1$. Let

$$g(\omega) := \sum_{|S|=k} \hat{f}(S)\, \chi_S(\omega), \qquad \omega \in \Omega.$$

The left-hand side of (1.3) is equal to $\|g\|^2$. Let $J \subset [n]$ be the random set of all bits examined by the algorithm. Let $\mathcal{A}$ denote the minimal σ-field for which J is measurable and every ω_i, $i \in J$, is measurable; this can be viewed as the relevant information gathered by the algorithm. For any function $h : \Omega \to \mathbb{R}$, let $h_J : \Omega \to \mathbb{R}$ denote the random function obtained by substituting the values of the bits in J. More precisely, if $\widetilde{\omega} = (\omega, \tau)$ and $\omega' \in \Omega$, then $h_J(\widetilde{\omega})(\omega')$ is $h(\omega'')$ where ω'' is ω on $J(\widetilde{\omega})$ and is ω' on $[n] \setminus J(\widetilde{\omega})$. In this way, h_J is a random variable on $\widetilde{\Omega}$ taking values in the set of mappings from Ω to $\mathbb{R}$ and it is immediate that this random variable is $\mathcal{A}$-measurable. When the algorithm terminates, the unexamined bits in Ω are unbiased and hence $\mathbf{E}\big[h \,\big|\, \mathcal{A}\big] = \int h_J (= \hat{h_J}(\varnothing))$ where $\int$ is defined, as usual, to be integration with respect to uniform measure on Ω. It follows that $\mathbf{E}[h] = \mathbf{E}[\int h_J]$.

More generally, if $u : \mathbb{R} \to \mathbb{R}$, then $(u \circ h)_J = u \circ h_J$ and hence, as above, $\mathbf{E}[u(h)] = \mathbf{E}\Big[\int u(h_J)\Big]$. In particular, for all h,

$$\|h\|^2 = \mathbf{E}\big[h^2\big] = \mathbf{E}\Big[\int h_J^2\Big] = \mathbf{E}\big[\|h_J\|^2\big]. \tag{2.2}$$

Since the algorithm determines f, it is $\mathcal{A}$ measurable, and we have

$$\|g\|^2 = \mathbf{E}[g\, f] = \mathbf{E}\Big[\mathbf{E}\big[g\, f \;\big|\; \mathcal{A}\big]\Big] = \mathbf{E}\Big[f\, \mathbf{E}\big[g \;\big|\; \mathcal{A}\big]\Big].$$

Since $\mathbf{E}\big[g \;\big|\; \mathcal{A}\big] = \hat{g}_J(\varnothing)$, Cauchy-Schwarz therefore gives

$$\|g\|^2 \le \sqrt{\mathbf{E}[\hat{g}_J(\varnothing)^2]}\, \|f\|\,. \tag{2.3}$$

We may write,

$$\mathbf{E}[\hat{g}_J(\varnothing)^2] = \mathbf{E}\big[\|g_J\|^2\big] - \mathbf{E}\Big[\sum_{|S|>0} \hat{g}_J(S)^2\Big].$$

This and (2.2) with $h = g$ imply that

$$\begin{aligned}\mathbf{E}[\hat{g}_J(\varnothing)^2] &\le \|g\|^2 - \mathbf{E}\Big[\sum_{|S|=k} \hat{g}_J(S)^2\Big] \\ &= \sum_{S\subseteq[n]} \hat{g}(S)^2 - \mathbf{E}\Big[\sum_{|S|=k} \hat{g}_J(S)^2\Big] = \sum_{|S|=k} \mathbf{E}\big[\hat{g}(S)^2 - \hat{g}_J(S)^2\big].\end{aligned}$$

It is easily seen that for any function h, $h_J = \sum_S \hat{h}(S)\,(\chi_S)_J$. We apply this with $h = g$. Since $\hat{g}(S') = 0$ if $|S'| > k$, it follows that for all $S \subset [n]$ satisfying $|S| = k$

$$\hat{g}_J(S) = \begin{cases} \hat{g}(S), & S \cap J = \varnothing, \\ 0, & S \cap J \neq \varnothing\,. \end{cases}$$

The above estimate for $\mathbf{E}[\hat{g}_J(\varnothing)^2]$ therefore gives

$$\mathbf{E}[\hat{g}_J(\varnothing)^2] \le \sum_{|S|=k} \hat{g}(S)^2\, \mathbf{P}\big[S \cap J \neq \varnothing\big] \le \sum_{|S|=k} \hat{g}(S)^2 \sum_{i\in S} \mathbf{P}[i \in J] \le \|g\|^2\, k\, \delta\,.$$

Substituting this estimate in (2.3) and squaring the resulting inequality complete the proof of the theorem. □

The theorem may be easily generalized to situations where the algorithm does not always determine the value of f precisely; that is, $f_J(x)$ still depends on $x \in \Omega$.

Set

$$\mathrm{var}_\Omega(f_J) := \int (f_J)^2 - \Big(\int f_J\Big)^2 = \sum_{S\neq\varnothing} \hat{f}_J(S)^2\,,$$

where the integrations are with respect to the uniform probability measure on Ω. Note that $\mathbf{E}\big[\mathrm{var}_\Omega(f_J)\big]$ is an indicator for how precisely the algorithm can be used to approximate f; when $\mathrm{var}_\Omega(f_J)$ is small, with high conditional probability, $|\hat{f}_J(\varnothing) - f|$ is not too large.

When $\mathrm{var}_\Omega(f_J) \neq 0$, we have to replace the calculation in the proof of Theorem 1.8 by the following

(2.4)

$$\|g\|^2 = \mathbf{E}\big[g_J\ f_J\big] = \mathbf{E}\big[g_J\ \hat{f}_J(\varnothing)\big] + \mathbf{E}\big[g_J\ (f_J - \hat{f}_J(\varnothing))\big]$$
$$\leq \Big(\mathbf{E}[\hat{g}_J(\varnothing)^2]\,\mathbf{E}[\hat{f}_J(\varnothing)^2]\Big)^{1/2} + \Big(\mathbf{E}[g_J^2]\,\mathbf{E}[\mathrm{var}_\Omega(f_J)]\Big)^{1/2}.$$

Since $\mathbf{E}\big[\hat{f}_J(\varnothing)^2 + \mathrm{var}_\Omega(f_J)\big] = \mathbf{E}[f^2]$, we have $\mathbf{E}[\hat{f}_J(\varnothing)^2] \leq \|f\|^2$. Thus, the above gives

$$\|g\|^2 \leq \sqrt{\mathbf{E}\big[\hat{g}_J(\varnothing)^2\big]}\,\|f\| + \|g\|\,\sqrt{\mathbf{E}[\mathrm{var}_\Omega(f_J)]}\,.$$

Using the same estimate for $\mathbf{E}\big[\hat{g}_J(\varnothing)^2\big]$ as in the proof of Theorem 1.8, we obtain

$$\|g\|^2 \leq \|g\|\,\|f\|\sqrt{k\,\delta} + \|g\|\,\sqrt{\mathbf{E}[\mathrm{var}_\Omega(f_J)]}\,.$$

Consequently, squaring both sides gives the following generalization of (1.3):

$$\sum_{|S|=k} \hat{f}(S)^2 \leq \Big(\|f\|\sqrt{k\,\delta} + \sqrt{\mathbf{E}[\mathrm{var}_\Omega(f_J)]}\Big)^2 \tag{2.5}$$
$$\leq 2\,k\,\delta\,\|f\|^2 + 2\,\mathbf{E}[\mathrm{var}_\Omega(f_J)]\,.$$

Theorem 1.8 holds more generally than stated. If W is a random subset of $[n]$, we say that W is a *witness* for $f : \Omega \to \mathbb{R}$ if the value of f is determined by its restriction to W. We say that W is δ-*dilute* if $\max_{i\in[n]} \mathbf{P}\big[i \in W(\omega)\big] \leq \delta$. The related notion of short witnesses is of central importance in Talagrand's epic isoperimetric saga [32]. The proof of Theorem 1.8 holds in the more general setting where the random set J is a witness, with the property that for all subsets $A \subseteq [n]$, conditioned on $J = A$ (assuming this has positive probability) and conditioned on ω restricted to A, the $\{\omega_i\}_{i\notin A}$ are uniform i.i.d. bits. As pointed out to us by Asaf Nachmias, Theorem 1.8 does not hold for arbitrary witnesses, even if we allow for a multiplicative constant in the right-hand side of (1.3): if you take "Recursive Ternary Majority" on $n = 3^h$ bits, there is a (symmetric) witness having only 2^h elements, yielding a δ which is $(2/3)^h$; however, the sum of the squares of the level 1 Fourier coefficients is $(3/4)^h$.

3. **Percolation background and notation**

Duality plays a central role in the theory of percolation in two dimensions. A dual-open path on the triangular grid is defined as a path in the grid whose vertices are all closed. For the square grid, a dual-open path is defined as a path in the dual of the square grid that does not intersect any open edge in the primal grid. The

basic observation is that for site percolation on the triangular grid at $p = 1/2$ the distribution of the collection of dual-open paths is the same as the distribution of the collection of primal open paths. (Sometimes, we use the term "primal open path", for an open path, to make the distinction with the dual-open path clearer.) For critical bond percolation on the square grid at $p = 1/2$, the distribution of the dual-open paths is the image of the distribution of the open paths under translation by $(1/2, 1/2)$. This simple duality is one of the important ingredients in the proof by Kesten that $p_c = 1/2$ for these two percolation models [18, p. 53] and the earlier proof by Harris (see [15]) that there is a.s. no infinite cluster at $p = 1/2$.

For $0 \le r < R < \infty$, let $A(r, R)$ denote the event that there is an open crossing of the annulus $r \le |z| \le R$, namely, an open path connecting a vertex inside the disk $|z| \le r$ to a vertex in $|z| \ge R$. Let $\alpha(r, R)$ denote the probability of $A(r, R)$ at percolation parameter $p = p_c = 1/2$. Abbreviate $\alpha(0, R)$ by $\alpha(R)$. For convenience, we adopt the convention $\alpha(r, R) = 1$ whenever $r \ge R$. The function $\alpha(r, R)$ is essentially multiplicative, in the following sense: there is a constant $C > 0$ such that for every $0 \le r_1 \le r_2 \le r_3 < \infty$,

$$C^{-1}\,\alpha(r_1, r_3) \le \alpha(r_1, r_2)\,\alpha(r_2, r_3) \le C\,\alpha(r_1, r_3)\,. \tag{3.1}$$

In fact, this holds for critical bond percolation on the square grid as well as for critical site percolation on the triangular grid. The (standard) proof of (3.1) is based on the Harris-FKG inequality and the celebrated Russo-Seymour-Welsh (RSW) theorem (see [12], [18]). A proof of a generalization of (3.1) is given in the appendix. Another consequence of RSW that we will use is the existence of a constant $c > 0$ such that for every $r > 0$,

$$c \le \alpha(r, 2r)\,. \tag{3.2}$$

The Stochastic Löwner Evolution (SLE) introduced in [28] is a one parameter family of random curves indexed by a real positive parameter κ. It was conjectured in [28] that the scaling limit of outer boundaries of critical site percolation clusters on the triangular grid as well as bond percolation clusters on $\mathbb{Z}^2$ are (chordal) SLE_κ with $\kappa = 6$. Smirnov [30], [29] proved the corresponding statement for critical site percolation on the triangular lattice. (See also [7].) We now explain some of this more precisely. We first perform independent site percolation on the upper half of the triangular lattice but declare the sites $\{(k, 0) : k > 0\}$ to all be open and $\{(k, 0) : k \le 0\}$ to all be closed. In the hexagonal grid dual to the triangular grid, there will then be a unique path in the upper half-plane from $(\frac{1}{2}, 0)$ to ∞ which has white hexagons containing open sites on the right and black hexagons containing closed sites on the left. (See Figure 3.1.) Smirnov's result is that the limit (in an appropriate topology) as the mesh size of the lattice goes to 0 of the law of this

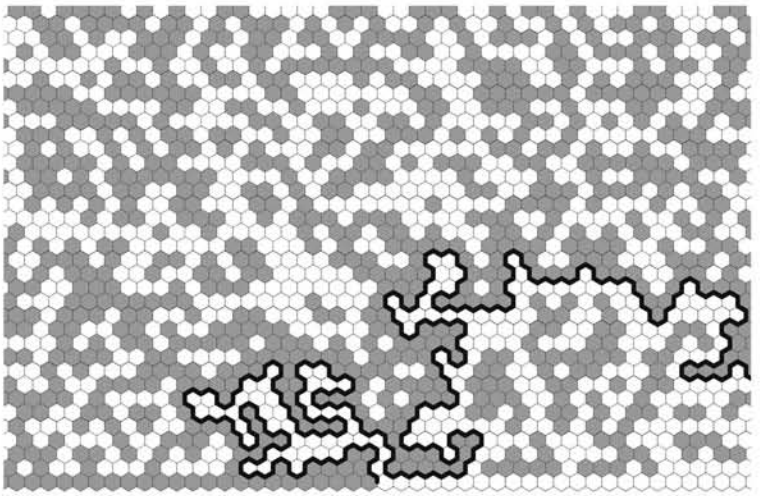


Figure 3.1. The percolation interface.

path is chordal SLE_6. This path described above, which has open sites on its right and closed sites on its left, is an example of what is called an *interface*.

The conformal invariance and the SLE description of critical percolation on the triangular lattice allowed researchers to prove a number of conjectures by physicists concerning so-called *critical exponents*. For example, for critical site percolation on the triangular grid, it was established in [23] that

$$\alpha(R) = R^{-5/48+o(1)} \text{ as } R \to \infty. \tag{3.3}$$

In fact, the same proof actually gives for $R \geq r \geq 1$,

$$\alpha(r, R) = (R/r)^{-5/48+o(1)} \text{ as } R/r \to \infty. \tag{3.4}$$

For $1 \leq r \leq R$, the 2-arm function $\alpha_2(r, R)$ denotes the probability that there are both an open path from $|z| \leq r$ to $|z| \geq R$ and also a dual-open path from $|z| \leq r$ to $|z| \geq R$. We abbreviate $\alpha_2(1, R)$ by $\alpha_2(R)$.

Next, let M be a half-plane in $\mathbb{R}^2$ and let v be some vertex in M satisfying $\text{dist}(v, \partial M) \leq 2$. Denote by $\alpha^+(R)$ the probability that there is an open path in M from v to distance at least R away from v. This quantity depends on R, v, and M, but the dependence on v and M will usually be suppressed. More generally, for $0 \leq r < R$, let $\alpha^+(r, R)$ denote the probability that there is an open path in M from some vertex u satisfying $|u - v| \leq r$ to some vertex w satisfying $|w - v| \geq R$.

As with $\alpha(r, R)$, we adopt the convention $\alpha^+(r, R) = 1$ whenever $r \geq R$.

It is known that the functions α^+ and α_2 also satisfy (3.1), with possibly different constants. (When considering α^+, this applies to any fixed choice of M and v.) These inequalities are valid for site percolation on the triangular grid as

well as bond percolation on the square grid. A proof can be found in the appendix. For site percolation on the triangular grid, the corresponding exponents were established in [31]:

$$\alpha^+(R) = R^{-1/3+o(1)}, \qquad \alpha_2(R) = R^{-1/4+o(1)}, \tag{3.5}$$

as $R \to \infty$. In fact, the same proofs actually give for $R \geq r \geq 1$

$$\alpha^+(r, R) = (R/r)^{-1/3+o(1)}, \qquad \alpha_2(r, R) = (R/r)^{-1/4+o(1)}, \tag{3.6}$$

as $R/r \to \infty$.

For bond percolation on the square grid, such exact estimates are unavailable, because there is currently no proof that the interface converges to SLE_6. In the case of the square grid, the estimate $\alpha(r, R) \leq C\,(R/r)^{-\varepsilon}$, where $C, \varepsilon > 0$ are constants, follows easily from the RSW theorem (see [12], [18]). The RSW proofs can give an actual value for ε, but it is rather small. We can also use the obvious estimate $\alpha^+(r, R) \leq \alpha(r, R)$ to obtain a similar bound for α^+.

We now give the precise definitions for wedges and cones. For this purpose, we first recall the definition of the infinitely branched cover of $\mathbb{R}^0$ over 0. Let $X = \{(z, \theta) : z \in \mathbb{C} \setminus \{0\},\ \theta \in \mathbb{R}, e^{i\theta}|z| = z\}$, and set $\psi(z, \theta) = z$. On the surface X we define the metric d_X as the pullback of the euclidean metric of $\mathbb{R}^2$ under ψ, namely, $d_X(x, y)$ is the infimum of the length of $\psi \circ \gamma$ for any continuous path $\gamma \subset X$ connecting x and y. Let C_∞ denote the completion of (X, d_X). Since $\mathbb{R}^2$ is complete, it is easy to see that $C_\infty \setminus X$ consists of a single point, which we denote by 0. We extend the map ψ by setting $\psi(0) = 0$. Let V be the set of points in C_∞ that are mapped to vertices of the triangular grid under ψ. The triangular grid on C_∞ has vertices V and an edge between any two vertices at distance 1 apart. Now the wedge $W_\theta \subset C_\infty$ is defined by $W_\theta := \{0\} \cup \{(z, \theta') \in X : \theta' \in [0, \theta)\}$. The triangular grid on W_θ is just the intersection of the triangular grid on C_∞ with W_θ.

On C_∞ we may define the rotation R_θ by $R_\theta(0) = 0$ and

$$R_\theta(z, \theta') = (e^{i\theta} z, \theta + \theta').$$

This is clearly an isometry of C_∞. The cone C_θ is defined as the quotient C_∞ / R_θ; that is, the set of equivalence classes of points in C_∞, where two points are considered equivalent if one is mapped to the other by a power of R_θ. Now suppose that $\theta = n\,\pi/3$ where $n \in \mathbb{N}_+$. Then R_θ restricts to an isomorphism of the triangular grid on C_∞. In this case we define the triangular grid on C_θ as the quotient of the grid on C_∞ under R_θ. In other words, the vertices are equivalence classes of vertices in C_∞ and an edge appears between two equivalence classes if there is an edge connecting representatives of these classes. Note that $C_{2\pi}$ is just the euclidean plane with the triangular lattice.

We end this section by describing the so-called full and half-plane exponents for k-arm events derived in [31].

For integer $k \geq 1$, let $A^k(r, R)$ be the event that there are k disjoint crossings of the annulus $\{z \in \mathbb{R}^2 : r \leq |z| \leq R\}$ with a specified color sequence (up to rotations), where we require that both colors appear in the color sequence. For $k \geq 2$, and $r \geq 10k$, it was proved in [31] that

$$\alpha_k(r, R) := \mathbf{P}\big[A^k(r, R)\big] = (R/r)^{\frac{1-k^2}{12}+o(1)}, \tag{3.7}$$

as $R \to \infty$ while r is fixed. (The result for $\alpha_2(R)$ in (3.5) above is a special case of this.) Next, when $A^k_+(r, R)$ is the event that there are k disjoint paths in the upper half-plane from $|z| \leq r$ to $|z| \geq R$ with any specified color sequence, then for $k \geq 1$, and $r \geq 10k$, it was proved in [31] that

$$\alpha^+_k(r, R) := \mathbf{P}\big[A^k_+(r, R)\big] = (R/r)^{\frac{-k(k+1)}{6}+o(1)}, \tag{3.8}$$

as $R \to \infty$ while r is fixed. (The result for $\alpha_+(R)$ in (3.5) is a special case of this.) Just as we said that the proofs of (3.5) actually yield (3.6), it is also the case that the proofs of (3.7) and (3.8) also yield versions when $R/r \to \infty$ while $r \geq 10\,k$ is not necessarily fixed.

4. **Noise sensitivity for percolation**

4.1. *The simply connected case.* To apply Theorem 1.8 to percolation, we will need to describe algorithms achieving small revealment. One result of that nature is

THEOREM 4.1. *Let $Q = Q_R$ be the indicator function for the event that critical site percolation on the standard triangular grid contains a left to right crossing in some grid-approximating domain D to a large square of side length R. (For example, we could take D to be the union of the hexagons in the dual grid that are contained in the square.) Then there is a randomized algorithm A determining Q such that $\delta_A \leq R^{-1/4+o(1)}$ as $R \to \infty$.*

For critical bond percolation on the square grid, there is such an algorithm satisfying $\delta \leq C\ R^{-a}$ for some constants $a, C > 0$.

Remark. Theorem 4.1 says that there is an algorithm for the relevant event which exposes on average at most $R^{7/4+o(1)}$ bits. Since the probability of points not too close to the boundary being pivotal is about $R^{-5/4+o(1)}$ (this is the 4-arm event) and for a monotone function f, $\hat{f}(\{i\})$ is the probability that x_i is pivotal, the case $k = 1$ in Theorem 1.8 implies that the revealment is at least $R^{-1/2+o(1)}$. As pointed out in Peres, Schramm, Sheffield and Wilson [25], this can also be obtained using an inequality of O'Donnell and Servedio.

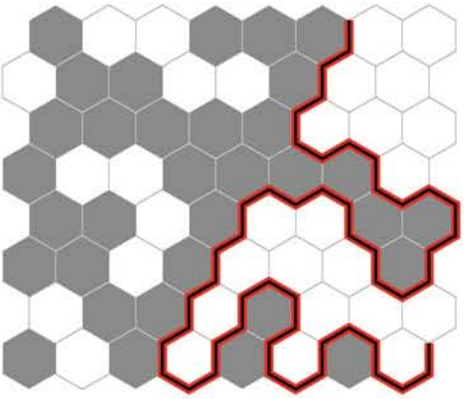


Figure 4.1. Following the interface from the corner.

Theorems 1.8 and 4.1 immediately give

COROLLARY 4.2. *For every $\varepsilon > 0$ there is a constant $C = C(\varepsilon)$ such that*

$$\sum_{|S|=k} \hat{Q}_R(S)^2 \leq C\, k\, R^{-1/4+\varepsilon}$$

holds for every $k = 1, 2, \ldots$ and for every $R > 0$. □

The basic idea of the proof of Theorem 4.1 is rather simple. First we consider the interface started at the lower right corner and stopped when it hits the upper or left edges. (See Figure 4.1.) This interface is sufficient to determine Q. If we traverse the interface, revealing just the bits necessary for its determination, then with high probability most bits will not be examined. However, this does not yield an algorithm with small revealment because the hexagons near the lower right corner are very likely to be examined. To rectify this problem, we instead start the interface at a different (random) location p_0 on the right side of D. This determines the existence of a crossing from the right side above p_0 to the left side. Then another interface started at p_0 will determine if there is a crossing that starts below p_0.

Let us now be a bit more precise regarding the notion of an interface. In the following, we use an equivalent dual version of the site percolation model on the triangular grid. The dual graph is the hexagonal grid, and we color the hexagon white if the site contained in it is open and black if the site in it is closed. Of course, there is no essential difference between these two representations. The advantage of this dual framework is that the figures are clearer and the notion of the interface is slightly more natural.

Note that bond percolation on the square grid also has a similar coloring representation. One such scheme is to color the squares of sidelength $1/2$ centered at the sites of $\mathbb{Z}^2$ white, color the squares of sidelength $1/2$ that are concentric

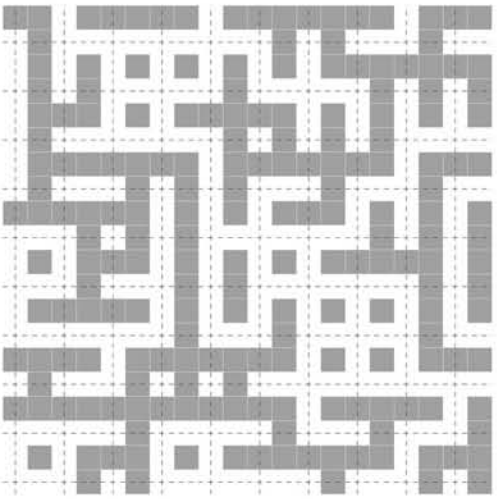


Figure 4.2. A color scheme for bond percolation on $\mathbb{Z}^2$.

with the square faces of $\mathbb{Z}^2$ black, and color each square of sidelength $1/2$ whose center is the midpoint of an edge of $\mathbb{Z}^2$ white or black if that edge is open or closed, respectively. See Figure 4.2. This scheme has the important property that the boundary between the union of the white clusters is a 1-manifold; that is, at every vertex of the grid $(1/2)\,\mathbb{Z}^2 + (1/4, 1/4)$ there are two edges that are on the common boundary.

We now assume that D is a bounded, simply connected, domain which is the interior of a union of hexagons in the hexagonal grid. Suppose that p_0 is a point on ∂D that is on the boundary of a single hexagon in D, and that ζ is a closed arc in $\partial D \setminus \{p_0\}$. The interface in $\bar{D}$ from p_0 to ζ is a random path β contained in the 1-skeleton of the hexagonal grid starting at p_0 and ending at a point in ζ, as indicated in Figure 4.3. We can precisely define β as the unique oriented simple path from p_0 to ζ that is contained in the union of the boundaries of the hexagons contained in $\bar{D}$ and satisfies (1) $\beta \cap \zeta$ consists of the terminal point of β, (2) whenever β traverses an arc along the boundary of a black hexagon $H \subset \bar{D}$, the arc is traversed counterclockwise around ∂H, and (3) whenever β traverses an arc along the boundary of a white hexagon $H \subset \bar{D}$, the arc is traversed clockwise around ∂H. It is easy to verify that this uniquely defines β, as follows. First, the initial arc of β is determined by the color of the hexagon in $\bar{D}$ containing p_0. When β first meets a hexagon contained in $\bar{D}$, its turn is clearly specified. (If the hexagon is black, then β must make a $\pi/3$ turn to the right, and if the hexagon is white, then β must make a $\pi/3$ turn to the left.) Now consider the situation in which β first meets a hexagon that is not contained in $\bar{D}$. Let β' be the arc of β from p_0 up to that point. Then β at that point makes the turn into the component of $\bar{D} \setminus \beta'$ that contains ζ, as in the figure.

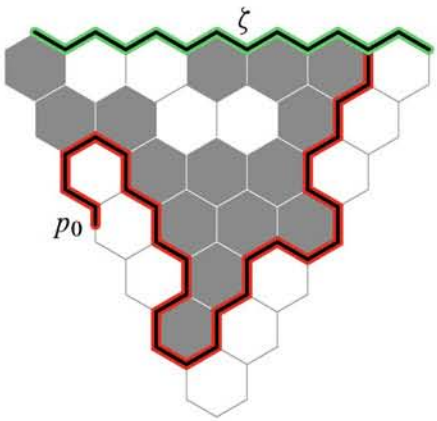


Figure 4.3. An interface started at p_0 and headed towards ζ.

Another way to describe this interface is that we color the counterclockwise arc of ∂D from p_0 to ζ white and the clockwise arc from p_0 to ζ black, and β then is the common boundary component between white and black in $\overline{D}$ starting at p_0. We will call this type of interface a *chordal interface*, to differentiate it from the interface needed later when discussing the annulus crossing event. (The chordal interface was proved by Smirnov [29] to converge to chordal SLE(6).)

Proof of Theorem 4.1. We mostly concentrate on the case of site percolation on the triangular grid. The details in the case of bond percolation on $\mathbb{Z}^2$ are essentially the same.

The algorithm proceeds as follows. (See Figure 4.4.) There are four distinguished boundary arcs of D, which we call "left", "right", "up" and "down". Pick uniformly at random an edge e_0 on the right-hand boundary of D, and let p_0 be its midpoint. Let ζ be the union of the top and left boundary segments of D. Explore the interface β from p_0 to ζ, examining the bits associated to sites in hexagons touching that interface, only as needed to continue with the determination of the interface. Note that the knowledge of β suffices to determine if there is an open crossing from the right boundary of D above p_0 to the left boundary of D: there is such a crossing if and only if β terminates on the left boundary of D.

Now let ζ' be the union of the bottom and left boundaries of D, and let β' be the interface from p_0 to ζ' that corresponds to the configuration ω' obtained by flipping all the colors of the hexagons in $\overline{D}$ (alternatively, β' is an interface that has black on the right and white on the left). Then β' determines the existence of an open crossing from the right boundary below p_0 to the left boundary. Consequently, after the algorithm examines β and β', the correct value of Q is determined. We now need to bound the revealment of this algorithm.

LEMMA 4.3. *Let ∂_W denote the counterclockwise arc of $\partial D \setminus \zeta$ from p_0 to ζ (the white arc), and let ∂_B denote the clockwise arc of $\partial D \setminus \zeta$ from p_0 to ζ. Let*

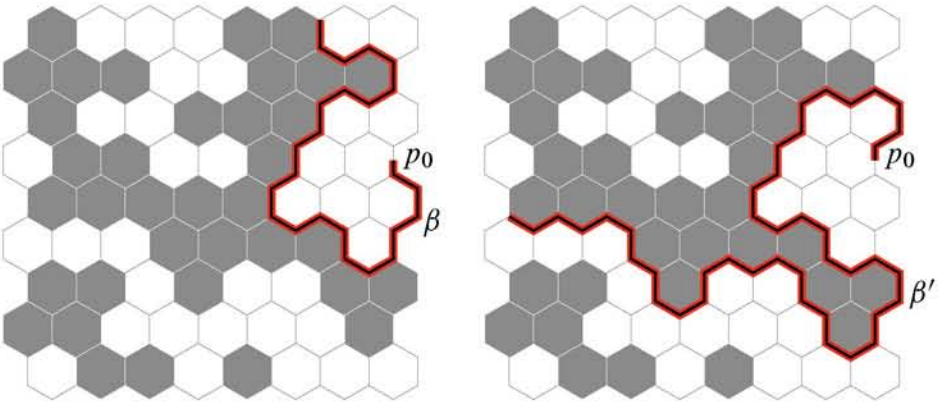


Figure 4.4. The interfaces β and β'.

H be a grid hexagon in $\bar{D}$. Set $r_1 := \min\{\mathrm{dist}(H, \partial_W), \mathrm{dist}(H, \partial_B)\}$ and $r_2 := \max\{\mathrm{dist}(H, \partial_W), \mathrm{dist}(H, \partial_B)\}$. Then

$$\mathbf{P}\big[\partial H \cap \beta \neq \varnothing \;\big|\; p_0\big] \leq \alpha_2(r_1)\, \alpha^+\big(2\, r_1 + O(1), r_2 - r_1 - O(1)\big)\,.$$

Proof. Suppose that $\partial H \cap \beta \neq \varnothing$. Then there is a path in black hexagons from ∂H to ∂_B and there is a path in white hexagons from ∂H to ∂_W, because the chains of hexagons along the two sides of β provide such paths. Suppose, for example, that $r_1 = \mathrm{dist}(H, \partial_W)$. Let w be a closest point to H in ∂_W. Let M be a half-space that contains D, which satisfies $\mathrm{dist}(\partial M, w) \leq C$ where $C = O(1)$. (Here, we are using the fact that D approximates a convex domain.) Let w' be a point closest to H on ∂M. Then $\mathrm{dist}(w', H) \leq r_1 + C$. Set $R_1 = 2\, r_1 + 2\, C + 2\,\mathrm{diam}(H)$ and $R_2 = r_2 - r_1 - C - \mathrm{diam}(H)$, and assume for now that $R_2 > R_1$. Consider next the annulus centered at w' with inner radius R_1 and outer radius R_2. Now, M intersected with this annulus contains a black crossing between the two boundary circles of this annulus (because there is a black crossing from H to ∂_B), and there are black and white crossings between H and the circle of radius r_1 around the center of H. These events are independent given p_0, which implies the lemma in the case $r_1 = \mathrm{dist}(H, \partial_W)$, $R_2 > R_1$. If $R_2 \leq R_1$ and $r_1 = \mathrm{dist}(H, \partial_W)$, we only need to consider the crossings between H and the circle of radius r_1 around its center. The case $r_1 = \mathrm{dist}(H, \partial_B)$ is treated similarly. □

Proof of Theorem 4.1, continued. Fix a hexagon $H \subset \bar{D}$. Note that the value of r_1 in the lemma does not depend on p_0, since $r_1 = \mathrm{dist}(H, \partial D \setminus \zeta)$. Let z_0 be the closest point to H on the right boundary of the square which D approximates. Observe that $r_2 \geq |p_0 - z_0| - O(1)$. This implies that for every $r \in [1, R]$, $\mathbf{P}\big[r/2 \leq r_2 < r\big] \leq O(r/R)$. Using the monotonicity of α^+, Lemma 4.3 therefore gives

$$\mathbf{P}\big[\partial H \cap \beta \neq \varnothing\big]$$
$$\leq O(1) \max_{0\leq r_1\leq R}\Big(\alpha_2(r_1) \sum_{j=0}^{\lceil \log_2 R\rceil} 2^{-j}\alpha^+\big(2r_1 + O(1), 2^{-j}R - r_1 - O(1)\big)\Big).$$

The same estimate also applies to β', by symmetry. Consequently, (3.6) gives

$$\mathbf{P}\big[H \text{ examined by algorithm}\big]$$
$$\leq R^{o(1)} \max_{0\leq r_1\leq R}\Big({r_1}^{-1/4} \sum_{j=0}^{\lceil \log_2 R\rceil} 2^{-j}(2^{-j}R/r_1)^{-1/3}\Big)$$
$$\leq R^{o(1)} \max_{0\leq r_1\leq R}\big({r_1}^{1/12}R^{-1/3}\big) = R^{-1/4+o(1)},$$

as $R\to\infty$. This proves the theorem in the case of the triangular grid.

The proof for the square grid is essentially the same. Since in that case, we cannot use the values of the critical exponents, we just use the bounds $\alpha^+(r,r') \leq C\,(r/r')^{\varepsilon_0}$ and $\alpha_2(r_1)\leq C\,r_1^{-\varepsilon_0}$, which are valid for some constants $C,\varepsilon_0>0$. The theorem follows. □

We now give the proof of Theorem 1.7. We will not prove Theorem 1.6, but rather simply say that it is proved in a similar way.

Proof of Theorem 1.7. Fix $\gamma < 1/8$. Let f be the indicator function $f(\omega) = 1_{A'_m}(\omega)$. By (2.1), we have

$$N(A'_m, m^{-\gamma}) = \sum_{\varnothing\neq S\subseteq[n_m]} \hat{f}(S)^2(1-2m^{-\gamma})^{2|S|}$$

where n_m is the number of sites in D. By Corollary 4.2, with $\varepsilon>0$ chosen so that $2\gamma+\varepsilon<1/4$, this is at most

$$C\sum_{k=1}^{n_m}(1-2m^{-\gamma})^{2k}\,k\,m^{-1/4+\varepsilon}$$
$$\leq C\,m^{-1/4+\varepsilon}\sum_{k=1}^{\infty} k\,(1-2\,m^{-\gamma})^{\frac{m^\gamma}{2}\frac{4k}{m^\gamma}} \leq C\,m^{-1/4+\varepsilon}\sum_{k=1}^{\infty} k\,e^{-4km^{-\gamma}}.$$

It is easy to check that

$$\sum_{k=1}^{\infty} k\,e^{-4km^{-\gamma}} \leq O(m^{2\gamma}),$$

and so the result follows immediately. □

4.2. *Annulus case.* For the proof of Theorem 1.3, we will need the following variant of Theorem 4.1 regarding the percolation crossings of an annulus.

THEOREM 4.4. *Let $2 \le r < R$. Let f_r^R be the indicator function for the event that there is a crossing of the annulus $\{z \in \mathbb{R}^2 : r \le z \le R\}$ from the inner circle to the outer circle by a cluster of white hexagons. Then there is a randomized algorithm determining f_r^R with*

$$\delta \le r^{o(1)}\, \alpha(r, R)\, \alpha_2(r)\,. \tag{4.1}$$

We stress that the $r^{o(1)}$ factor depends on r only and not on R. It is possible to replace the factor $r^{o(1)}$ by $O(1)$, but in order to do this it seems that one would first need to appeal to the analogue of (3.1) for α_2, which is proved in the appendix. In order to have a more direct proof of our main theorem, we prefer, at this point, not to rely on the appendix. By (3.4) and (3.5), the right-hand side in (4.1) is equal to $R^{-5/48+o(1)}\, r^{-7/48}$, but its writing in (4.1) is more suggestive and more useful.

Since $\|f_r^R\|^2 = \alpha(r, R)$, Theorems 1.8 and 4.4 give

COROLLARY 4.5.

$$\sum_{|S|=k} \hat{f_r^R}(S)^2 \le k\, r^{o(1)}\, \alpha(r, R)^2\, \alpha_2(r)$$

holds when $1 \le r < R < \infty$ and $k > 0$. □

Before we prove Theorem 4.4, we have to discuss the kind of interface that is used in the algorithm, as it is slightly different from the interface used to determine a possible crossing of a square.

Fix $R > 0$ large. Let $\bar{D} = \bar{D}_R$ be the union of the hexagons of the hexagonal grid that intersects the disk $|z| \le R$. Let $V^* = V_R^*$ denote the set of vertices of the hexagonal grid that are in $\bar{D}$. Let p_0 be some point in $\partial\bar{D} \setminus V^*$. Let H_0 denote the hexagon containing the origin, and let q_0 be some point in $\partial H_0 \setminus V^*$. We define the *radial interface* $\beta = \beta(R, p_0, q_0, \omega)$, inductively as a simple path from p_0 to q_0. (See Figure 4.5.) The construction is segment by segment, and the concatenation of the first m segments will be denoted β_m. If the (unique) hexagon in $\bar{D}$ containing p_0 is white (respectively, black), then the first segment β_1 of β traverses the boundary of that hexagon (counter-) clockwise, until the first encounter with a point in V^*. Suppose inductively β_m has been constructed, that it is a simple path, that $p_m \in V^*$ and p_0 are the two endpoints of β_m, and that there is a path α in the hexagonal grid in $\bar{D}$ from p_m to q_0 whose only intersection with β_m is p_m. The first step of such a path α is along an edge $\hat{e}$ starting at p_m. If there is just one possible $\hat{e}$ among all such α, then β_{m+1} also uses that edge $\hat{e}$. Clearly, there are at most two possible $\hat{e}$, since the edge terminating at p_m and used by β_m cannot be used. If there are two possible $\hat{e}$, then β_{m+1} chooses between them according to the color of the hexagon containing them both; i.e., the hexagon just encountered by β_m. If that hexagon is white (respectively, black), then the edge

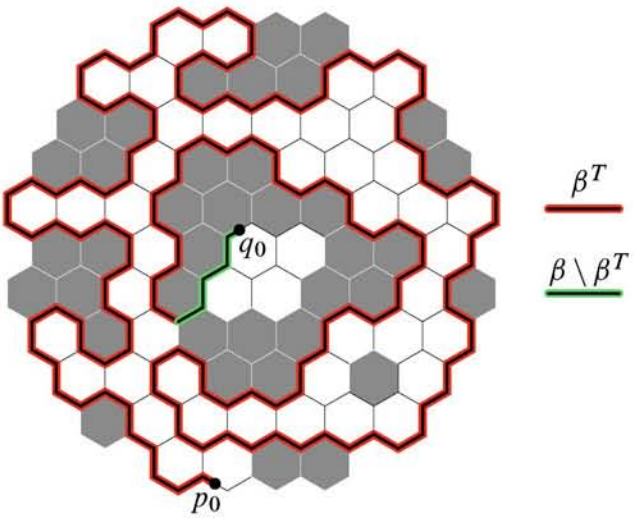


Figure 4.5. The radial interface β.

chosen is the one that traverses H (counter-) clockwise. If the edge chosen contains q_0, then the path stops at q_0 and the construction terminates. Otherwise, β_{m+1} is defined as the union of β_m and the chosen continuation edge. This completes the definition of β.

It is not hard to verify that for every simple path $\hat{\beta}$ in the hexagonal grid from p_0 to q_0 that stays in $\bar{D}$, the probability that $\beta = \hat{\beta}$ is precisely 2^{-n} if n is the number of hexagons in $\bar{D}$ that intersect $\hat{\beta}$. However, we will not use this fact.

Let $r \in [0, R]$. We now define a truncated version of β, which will suffice, as we will see, to determine f_r^R. We say that β completed a (counter-) clockwise loop at some dual vertex $v \in V^*$ if v is visited by β and there is a hexagon H containing v and another point $u \in \partial H$, which was visited by β prior to v. The oriented arc of β from u to v together with the line segment $[v, u] \subset H$ form a (counter-) clockwise loop surrounding 0. Let β^T denote the initial segment of β up to the first time in which β completed a counterclockwise loop around 0 or until it hits q_0, if there is no such loop.

LEMMA 4.6. *The truncated radial interface β^T meets the disk $|z| \le r$ if and only if $f_r^R = 1$.*

Proof. Let $[u, v]$ be an edge in β^T, with u occurring before v along β. We claim that if the hexagon H immediately to the right of $[u, v]$ is contained in $\bar{D}$, then it is white. Indeed, suppose the contrary. Let w be the first vertex along β in which ∂H is visited, and let β^w be the initial segment of β from p_0 to w. Observe that the counterclockwise arc from w to v is a feasible continuation of β^w, since β contains a path from v to q_0 and there is no other point but w in $\partial H \cap \beta^w$. Since

we are assuming that $H \subset \bar{D}$ is black, it follows that the immediate continuation of β^w was along ∂H in the counterclockwise direction until some $w' \in V^*$ is hit.

Consider the directed cycle obtained by joining the line segment $[u, w']$ to the arc of β from w' to u. This directed cycle surrounds v, because w is connected in β^w to p_0, which is certainly in the unbounded component of this cycle, and the line segment $[v, w]$ intersects the cycle precisely once, crossing the line segment $[u, w']$ inside H. Moreover, if we consider the orientation in which these two line segments cross, we conclude that the cycle surrounds v counterclockwise. Because the arc of β from v to q_0 does not intersect the cycle, we conclude that the cycle also surrounds 0 counterclockwise. This contradicts the definition of the truncated path β^T, since we are assuming $v \in \beta^T$. This verifies our claim, that to the right of edges in β^T there are only white hexagons and hexagons that are not contained in $\bar{D}$.

Note that if e and e' are two consecutive segments along β, then the hexagon to the right of e is either the same as the one to the right of e', or these hexagons are adjacent. We therefore conclude that every white hexagon visited by β^T is connected by a chain of white hexagons to ∂D. Therefore, if β^T hits the set $|z| \le r$, then clearly $f_r^R = 1$.

Now suppose that β^T does not hit $|z| \le r$. This implies that β^T has terminated by completing a counterclockwise loop around the set $|z| \le r$. Consider a hexagon H on the inner boundary of this loop. Because the orientation of the loop is counterclockwise, the first time in which H is visited, the path chosen to traverse ∂H is counterclockwise. This implies that the hexagon is black. Thus, there is a loop of black hexagons in $\bar{D}$ that surrounds the set $|z| \le r$. This implies that $f_r^R = 0$. □

We can now specify the algorithm promised by Theorem 4.4. The algorithm starts by selecting the point p_0 uniformly along $\partial \bar{D}$, and selecting q_0 arbitrarily in $\partial H_0 \setminus V^*$. It then proceeds to inspect the colors of the hexagons necessary to develop the truncated interface β^T, until it terminates or hits the set $|z| \le r$. At that point, the correct value of f_r^R is determined by Lemma 4.6.

In order to bound the revealment of this algorithm, it will be convenient to introduce a different interface, which in the end will turn out to be equivalent to β.

Let $\hat{D}$ denote the branched double cover of $\bar{D}$ about 0, and let $\phi : \hat{D} \to \bar{D}$ denote the projection map. Concretely, define $\hat{D}$ as the preimage of $\bar{D}$ under the map $\phi(z) = z^2$. Let $\hat{p}_0$ be one of the preimages of p_0 under ϕ, and let $\hat{q}_0$ be one of the preimages of q_0. Let $\hat{H}_0$ be the closure of one of the connected components of $\phi^{-1}(H_0) \setminus [\hat{q}_0, -\hat{q}_0]$. Let $\mathfrak{H}$ denote the set of hexagons H that are contained in $\bar{D}$. Let $\hat{\mathfrak{H}}$ denote the set of connected components of preimages $\phi^{-1}(H)$, $H \in \mathfrak{H}$, except that the single preimage of H_0 is replaced by the two sets $\hat{H}_0$ and $-\hat{H}_0$. Note that if $\hat{H} \in \hat{\mathfrak{H}}$, then $-\hat{H} \in \hat{\mathfrak{H}}$ and $\phi(\hat{H}) = \phi(-\hat{H}) \in \mathfrak{H}$. Let $\hat{\mathfrak{H}}' \subset \hat{\mathfrak{H}}$ be a maximal collection of elements of $\hat{\mathfrak{H}}$ with the property that $\hat{\mathfrak{H}}' \cap \{-\hat{H} : \hat{H} \in \hat{\mathfrak{H}}'\} = \varnothing$.

Now color at random each of the elements of $\widehat{\mathfrak{H}}'$ white or black independently, with probability $1/2$. For every $\widehat{H} \in \widehat{\mathfrak{H}}'$, let $-\widehat{H}$ have the opposite color to the color of $\widehat{H}$.

Now let $\hat{\beta}$ denote the chordal interface in $\widehat{D}$ from $\hat{p}_0$ to $-\hat{p}_0$, with white cells on the right and black cells on the left, as defined in the simply connected setting in Section 4.1. That is, we consider the exterior of the counterclockwise arc from $\hat{p}_0$ to $-\hat{p}_0$ along $\partial\widehat{D}$ as white, the exterior of the complementary arc as black, and take $\hat{\beta}$ as the interface between white and black passing through $\hat{p}_0$ and through $-\hat{p}_0$. Finally, let $\beta^\dagger := \phi(\hat{\beta}) \setminus \text{interior}(H_0)$.

LEMMA 4.7. *Given p_0 and q_0, the laws of $\beta^\dagger$ and of β are the same.*

(In this statement, we consider β as a set, and forget about the fact that it has the structure of an oriented path.)

Proof. The map $z \mapsto -z$ preserves $\hat{\beta}$, by the symmetry of the interface. Consequently, near every point $p \in \beta^\dagger \setminus \{p_0, q_0\}$, $\beta^\dagger$ looks like a piecewise linear path. Moreover, $\beta^\dagger$ is connected and contains p_0. Since a compact simple path has two endpoints, we conclude that $q_0 \in \beta^\dagger$ as well. We now consider building $\beta^\dagger$ by adding one segment at a time. When it hits a previously unvisited hexagon H which is contained in $\overline{D}$, it is equally likely (given its past) to turn right or left. (This is because both preimages of H are unvisited by both preimages of the past of $\beta^\dagger$.) When it hits a previously visited hexagon (or a hexagon that is not contained in $\overline{D}$), it turns in such a way that it will eventually be able to reach q_0 without crossing itself, and this uniquely specifies this turn. Consequently, the lemma follows. □

Remark 4.8. The radial interface converges to radial SLE(6) as the mesh tends to zero.

Proof of Theorem 4.4. Given all of our preparations, the proof is rather easy. We have shown that the above algorithm provides the correct answer. It therefore remains to estimate its revealment. Consider some hexagon $H \subset \overline{D}$. We want to prove that the right-hand side of (4.1) is an upper bound for the probability that H is examined. Let $a := \text{dist}(0, H)$, $b := \text{dist}(H, \partial\overline{D})$ and $c := \text{dist}(p_0, H)$. Let S_1 be the disk of radius $(a \wedge b)/2$ concentric with H, and let $\hat{S}_1$ be one of the two connected components of $\phi^{-1}(S_1)$. We have to bound the probability that the algorithm inspects H. Clearly, we may assume $a \geq r - O(1)$. For H to be inspected, β^T has to get to the circle $|z| = R \wedge (2a)$. This probability is $\alpha(2a, R)$. Given that this has happened, how can we estimate the probability that β is adjacent to H? At this point, we use the equivalence of β and $\beta^\dagger$. The information that $\beta^\dagger$ reached the circle $|z| = R \wedge (2a)$ has no impact on the distribution of the colors of the cells in $\widehat{\mathfrak{H}}$ whose images under ϕ intersect S_1. (Here, we assume that a is

not too small, so that the corresponding sets of cells are disjoint. Certainly $a > 10$ suffices. If a is smaller, then the estimate we are now striving for is trivial.) Since there is no hexagon intersecting both $\hat{S}_1$ and $-\hat{S}_1$, it follows that the conditional distribution of the colors of the cells meeting $\hat{S}_1$ is uniform i.i.d. Consequently, the conditional probability that $\beta^\dagger$ hits H is bounded by $\alpha_2((a \wedge b)/2)$. Thus,

$$\mathbf{P}\big[H \text{ visited}\big] \le O(1)\,\alpha(2a, R)\,\alpha_2((a \wedge b)/2)\,.$$

In the case $b \ge a \ge 2r$, we may use independence on disjoint sets to conclude that

$$\begin{aligned}\mathbf{P}\big[H \text{ visited}\big] &\le O(1)\,\alpha_2(r)\,\alpha_2(2r, a/2)\,\alpha(2a, R) \le O(1)\,\alpha_2(r)\,\alpha(2r, a/2)\,\alpha(2a, R)\\ &\overset{(3.2)}{\le} O(1)\,\alpha_2(r)\,\alpha(r, 2r)\,\alpha(2r, a/2)\,\alpha(a/2, 2a)\,\alpha(2a, R)\\ &\overset{(3.1)}{\le} O(1)\,\alpha_2(r)\,\alpha(r, R)\,.\end{aligned}$$

On the other hand, if $b \ge a$ and $a \le 2r$, then we use our assumption $a \ge r - O(1)$ and (3.5) to get $\alpha_2(a/2) \le r^{-1/4+o(1)} \le \alpha_2(r)\, r^{o(1)}$, which is also sufficient.

In the case $b < a$, a similar argument (and similar to the proof of Lemma 4.3) shows that

$$\mathbf{P}\big[H \text{ visited } \big|\, c\big] \le O(1)\,\alpha_2(b/2)\,\alpha^+\big(2b + O(1), c - b - O(1)\big).$$

Next, picking a constant $q \in (1/4, 1/3)$, we then have by the above and (3.6)

$$\mathbf{P}\big[H \text{ visited } \big|\, c\big] \le O(1)\,\alpha_2(b/2)\,(c/b)^{-q}.$$

As in the proof of Theorem 4.1, we have $\mathbf{P}\big[2^j \le c < 2^{j+1}\big] \le O(1)\,2^j/R$. It easily follows that

$$\mathbf{P}\big[H \text{ visited}\big] \le O(1)\,\alpha_2(b/2)\,(R/b)^{-q}.$$

If $b/2 > r$, then we may estimate

$$\alpha_2(b/2) \le \alpha_2(r)\,\alpha_2(2r, b/2) \le \alpha_2(r)\,\alpha(2r, b/2)$$

and

$$(R/b)^{-q} \overset{(3.6)}{\le} O(1)\,\alpha(b/2, R)$$

and we get from (3.1) and the above

$$\mathbf{P}\big[H \text{ visited}\big] \le O(1)\,\alpha_2(r)\,\alpha(r, R)\,.$$

If $b/2 \le r$, we use instead

$$\alpha_2(b/2)\,(R/b)^{-q} \le O(1)\,\alpha_2(b/2)\,\alpha_2(b/2, r)\,\alpha(r, R) \overset{(3.6)}{\le} r^{o(1)}\,\alpha_2(r)\,\alpha(r, R)\,.$$

This completes the proof. □

Remark 4.9. It is easy to see that if we assume the analogue of (3.1) for α_2 proved in the appendix, then the $r^{o(1)}$ term in (4.1) can be replaced by $O(1)$.

5. **Exceptional times**

In this section we prove Theorem 1.3. We point out that the absolute key necessary step is to get a good bound on the correlation for an event occurring at two different but close-by times. Once this is done, the rest is fairly standard. Proposition A16 in Lawler [22] indicates this general type of argument.

Proof of Theorem 1.3. By Kolmogorov's 0-1 law, it suffices to prove that with positive probability there are times in $[0, 1]$ when the origin is in an infinite cluster. Fix $R > 2$ large and let $V_{t,R}$ be the event that at time t there is an open path from the origin to distance R away. We then let

$$X = X_R := \int_0^1 1_{V_{t,R}}\, dt$$

be the Lebesgue measure of the set of times in $[0, 1]$ at which $V_{t,R}$ occurs. The first moment of X is given by

$$\mathbf{E}\big[X\big] = \int_0^1 \mathbf{P}\big[V_{t,R}\big]\, dt = \mathbf{P}\big[V_{0,R}\big] = \alpha(R)\,.$$

The second moment is

$$\mathbf{E}\big[X^2\big] = \mathbf{E}\Big[\int_0^1 \int_0^1 1_{V_{s,R}}\, 1_{V_{s',R}}\, ds\, ds'\Big] = \int_0^1 \int_0^1 \mathbf{P}\big[V_{s,R} \cap V_{s',R}\big]\, ds\, ds'. \tag{5.1}$$

For each site v we let

$$\chi_v^s := \begin{cases} -1 & v \text{ is open at time } s \\ 1 & \text{otherwise,} \end{cases}$$

and for a finite set of sites S set

$$\chi_S^s := \prod_{v \in S} \chi_v^s\,.$$

Fix $s, s' \in [0, 1]$, and set $t := |s - s'|$. Recall that the state of a site v flips between closed and open with rate $1/2$. Equivalently, we may think of the state as being re-randomized with rate 1. Consequently, $\mathbf{P}\big[\chi_v^{s'} = \chi_v^s \;\big|\; \omega_s\big] = e^{-t} + (1 - e^{-t})/2 = (1 + e^{-t})/2$, and hence,

$$\mathbf{E}\big[\chi_v^s\, \chi_v^{s'}\big] = \exp(-t)\,, \qquad \mathbf{E}\big[\chi_S^s\, \chi_S^{s'}\big] = \prod_{v \in S} \mathbf{E}\big[\chi_v^s\, \chi_v^{s'}\big] = \exp(-t\,|S|)\,.$$

Moreover, if $S \neq S'$, then $\mathbf{E}\big[\chi_S^s\, \chi_{S'}^{s'}\big] = 0$. Consequently, if f is a function depending on the states of finitely many lattice points and has the expansion $f(\omega) =$

$\sum_S \hat{f}(S)\,\chi_S(\omega)$, then

$$\mathbf{E}\big[f(\omega_s)\,f(\omega_{s'})\big] = \sum_S \hat{f}(S)^2 \exp(-t\,|S|)\,. \tag{5.2}$$

Let $f_r^R(\omega)$ be as in Theorem 4.4. Fix some $t \in (0,1]$ and let $r \in [2,R)$. Clearly, $0 \le f_0^R(\omega) \le f_0^r(\omega)\,f_{2r}^R(\omega)$ for every ω. Consequently,

$$\begin{aligned}\mathbf{P}\big[V_{s,R} \cap V_{s',R}\big] &= \mathbf{E}\big[f_0^R(\omega_s)\,f_0^R(\omega_{s'})\big] \le \mathbf{E}\big[f_0^r(\omega_s)\,f_{2r}^R(\omega_s)\,f_0^r(\omega_{s'})\,f_{2r}^R(\omega_{s'})\big] \\ &= \mathbf{E}\big[f_0^r(\omega_s)\,f_0^r(\omega_{s'})\big]\mathbf{E}\big[f_{2r}^R(\omega_s)\,f_{2r}^R(\omega_{s'})\big] \le \mathbf{E}\big[f_0^r(\omega_s)\big]\mathbf{E}\big[f_{2r}^R(\omega_s)\,f_{2r}^R(\omega_{s'})\big].\end{aligned}$$

(To obtain the second equality, we have used the independence on disjoint sets of sites.) Thus,

$$\mathbf{P}\big[V_{s,R} \cap V_{s',R}\big] \le \alpha(r)\,\mathbf{E}\big[f_{2r}^R(\omega_s)\,f_{2r}^R(\omega_{s'})\big] = \alpha(r)\sum_S e^{-t|S|}\,\hat{f}_{2r}^R(S)^2.$$

The latter sum restricted to S with $|S| = k$ for fixed $k \neq 0$ is estimated using Corollary 4.5, while for $k = 0$, we use $\hat{f}_{2r}^R(\varnothing) = \alpha(2r,R)$. This yields

$$\mathbf{P}\big[V_{s,R} \cap V_{s',R}\big] \le \alpha(r)\Big(\alpha(2\,r,R)^2 + r^{o(1)}\sum_{k=1}^{\infty} e^{-kt}\,k\,\alpha(2\,r,R)^2\,\alpha_2(r)\Big).$$

It is easy to check that $\sum_{k=1}^{\infty} k\,e^{-kt} \le O(t^{-2})$. This and the inequalities (3.1) and (3.2) allow us to write this estimate as

$$\mathbf{P}\big[V_{s,R} \cap V_{s',R}\big] \le O(1)\,\alpha(R)^2\,\alpha(r)^{-1}\big(1 + r^{o(1)}\,t^{-2}\,\alpha_2(r)\big)\,. \tag{5.3}$$

We proved the above claim for all $t \in (0,1]$ and $r \in [0,R)$ but now we observe that (5.3) is also trivially true when $r \ge R$ as well. We now choose $r = 2\,t^{-8} = 2\,|s-s'|^{-8}$. Applying this in (5.3) with (3.3) and (3.5) gives

$$\mathbf{P}\big[V_{s,R} \cap V_{s',R}\big] \le O(1)\,\alpha(R)^2\,|s-s'|^{-5/6+o(1)}. \tag{5.4}$$

Hence,

$$\int_0^1\int_0^1 \mathbf{P}\big[V_{s,R} \cap V_{s',R}\big]\,ds\,ds' \le O(1)\,\alpha(R)^2. \tag{5.5}$$

The Cauchy-Schwarz inequality tells us that

$$\mathbf{P}\big[X > 0\big] \ge \frac{\mathbf{E}\big[X\big]^2}{\mathbf{E}\big[X^2\big]}\,.$$

Consequently, the above inequality, the fact that $\mathbf{E}\big[X\big] = \alpha(R)$, the expression (5.1) for $\mathbf{E}\big[X^2\big]$ and (5.5) show that $\inf_{R>0}\mathbf{P}\big[X_R > 0\big] > 0$. Let $T_R := \{t \in [0,1] : V_{t,R} \text{ holds}\}$. Fatou's lemma tells us that with positive probability $T_R \neq \varnothing$ for

infinitely many $R \in \mathbb{N}$. Since $T_R \supset T_{R'}$ when $R' > R$, this implies that

$$\mathbf{P}\big[\cap_{R>0} \{T_R \neq \varnothing\}\big] > 0.$$

Our goal is to show that $\mathbf{P}\big[\bigcap_R T_R \neq \varnothing\big] > 0$. Since the T_R's are not closed sets, $\cap_{R>0} \{T_R \neq \varnothing\}$ does not immediately imply $\bigcap_R T_R \neq \varnothing$. (The reason that T_R is not necessarily closed is that the set of times at which an edge is open is not a closed set since we have a right continuous process.) This technicality is taken care of by the following lemma from [13].

LEMMA 5.1 ([13]). *Let $0 < p < 1$ and let G be any graph where $\pi_p(\mathscr{C}) = 0$. Let $\{\omega_t\}$ represent our dynamical percolation process in that $\omega_t(v)$ is the state of vertex v at time t. Consider the process $\{\bar{\omega}_t\}$ obtained from $\{\omega_t\}$ by setting, for every vertex v, the set $\{t : \bar{\omega}_t(v) = 1\}$ to be the closure of the set $\{t : \omega_t(v) = 1\}$. Then Ψ_p-a.s., for every vertex v*

$$\{t \in [0,\infty) : v \text{ percolates in } \bar{\omega}_t\} = \{t \in [0,\infty) : v \text{ percolates in } \omega_t\}.$$

In particular, a.s. this set of times is closed.

Returning to our proof, let $\overline{T_R}$ be the closure of T_R. It is easily checked that

$$\bigcap_{R>0} \overline{T_R} = \{t \in [0,1] : 0 \text{ percolates in } \bar{\omega}_t\},$$

where $\{\bar{\omega}_t\}$ is defined as in Lemma 5.1. By compactness, if the T_R's are all nonempty, it follows that $\bigcap_R \overline{T_R}$ is nonempty. This implies that there is some time at which $\bar{\omega}_t$ percolates and hence by Lemma 5.1, some time at which the original process ω_t percolates. □

For future reference, we note that Lemma 5.1 implies that a.s.

$$\bigcap_{R>0} \overline{T_R} = \bigcap_{R>0} T_R. \tag{5.6}$$

6. Hausdorff dimension of exceptional times

In this section, we prove Theorem 1.9. This is separated into two theorems, Theorem 6.1 and Theorem 6.3, where lower and upper bounds are given. We point out however that the lower bound is simply a refinement of the argument for proving that there exist exceptional times. First note the fact that the Hausdorff dimension is an almost sure constant follows immediately from ergodicity.

THEOREM 6.1. *As such, the Hausdorff dimension of the set of exceptional times is at least* $1/6$.

Proof. Fix $\gamma < 1/6$. It suffices by ergodicity and countable additivity to show that with positive probability, the set of exceptional times in $[0,1]$ at which the

origin percolates has Hausdorff dimension at least γ. For each integer R, let, as before, $V_{t,R}$ be the event that at time t there is a path from the origin to distance R away and define a random measure σ_R on $[0,1]$ by

$$\sigma_R(S) = \frac{1}{\alpha(R)} \int_S 1_{V_{t,R}} dt$$

for each Borel set $S \subset [0,1]$.

The results in the previous section immediately give that $\mathbf{E}[\|\sigma_R\|] = 1$ and $\mathbf{E}[\|\sigma_R\|^2] \le O(1)$ where $\|\sigma_R\|$ denotes the total variation of the measure σ_R.

Cauchy-Schwarz gives

$$\mathbf{E}\big[\|\sigma_R\|^2\big]^{1/2} \mathbf{P}\big[\|\sigma_R\| > 1/2\big]^{1/2} \ge \mathbf{E}\big[\|\sigma_R\| 1_{\|\sigma_R\|>1/2}\big] \ge \mathbf{E}[\|\sigma_R\|] - 1/2 = 1/2.$$

Consequently, $\mathbf{P}\big[\|\sigma_R\| > 1/2\big] \ge C_1$ for some constant $C_1 > 0$. Given a measure m on $[0,1]$ and $\gamma > 0$, let

$$\mathscr{E}_\gamma(m) = \iint |t-s|^{-\gamma}\, dm(t)\, dm(s).$$

Note that

$$\mathbf{E}\big[\mathscr{E}_\gamma(\sigma_R)\big] = \mathbf{E}\Big[\int_0^1 \int_0^1 \frac{d\sigma_R(t)\, d\sigma_R(s)}{|t-s|^\gamma}\Big] = \int_0^1 \int_0^1 \frac{\mathbf{P}\big[V_{t,R} \cap V_{s,R}\big]}{\alpha(R)^2\, |t-s|^\gamma}\, dt\, ds\,.$$

Therefore, by (5.4) and $\gamma < 1/6$,

$$C_2 := \sup_R \mathbf{E}\big[\mathscr{E}_\gamma(\sigma_R)\big] < \infty\,.$$

By Markov's inequality, for all R and for all T,

$$\mathbf{P}\big[\mathscr{E}_\gamma(\sigma_R) \ge C_2 T\big] \le 1/T.$$

Choose T so that $1/T < C_1/2$. Letting

$$U_R = \{\|\sigma_R\| > 1/2\} \cap \{\mathscr{E}_\gamma(\sigma_R) \le C_2 T\},$$

by the choice of T, we have that

$$\mathbf{P}\big[U_R\big] \ge C_1/2.$$

By Fatou's lemma,

$$\mathbf{P}\big[\limsup_{R\to\infty} U_R\big] \ge C_1/2.$$

We now show that on the event $\limsup_R U_R$, the Hausdorff dimension of the set of percolating times in $[0,1]$ is at least γ. Let $\overline{T_R}$ again be the closure of the set of times in $[0,1]$ at which there is a path from the origin to distance R away. Clearly σ_R is supported on $\overline{T_R}$. By (5.6), it suffices to prove that $\bigcap_{R>0} \overline{T_R}$ has Hausdorff dimension at least γ on the event $\limsup_R U_R$. This is achieved in the following (deterministic) lemma, which completes the proof. □

LEMMA 6.2. *Let $D_1 \supseteq D_2 \supseteq D_3 \ldots$ be a decreasing sequence of compact subsets of* $[0,1]$, *and let $\mu_1, \mu_2, \ldots$ be a sequence of positive measures with μ_n supported on D_n. Suppose that there is a constant C such that for infinitely many values of n,*

$$\|\mu_n\| > 1/C, \quad \text{and} \quad \mathscr{E}_\gamma(\mu_n) \le C. \tag{6.1}$$

Then the Hausdorff dimension of $\bigcap_n D_n$ is at least γ.

Proof. Choose a sequence of integers $\{n_k\}$ for which (6.1) holds. Note that $\|\mu_{n_k}\|^2 \le \mathscr{E}_\gamma(\mu_{n_k}) \le C$. By compactness, choose a further subsequence $\{n'_k\}$ of $\{n_k\}$ so that $\mu_{n'_k}$ converges weakly to some positive measure μ_∞. Clearly μ_∞ is supported on $\bigcap_n D_n$ and $\|\mu_\infty\| \ge 1/C$. For all M, we have that

$$\iint |x-y|^{-\gamma} \wedge M \, d\mu_\infty(x)\, d\mu_\infty(y) = \lim_{k\to\infty} \iint |x-y|^{-\gamma} \wedge M \, d\mu_{n'_k}(x)\, d\mu_{n'_k}(y) \le C.$$

Now let $M \to \infty$ and apply the monotone convergence theorem to conclude that

$$\iint |x-y|^{-\gamma} \, d\mu_\infty(x)\, d\mu_\infty(y) \le C.$$

Since $\|\mu_\infty\| > 0$, it now follows from Frostman's theorem (see for example, [17]) that the Hausdorff dimension of $\bigcap_n D_n$ is at least γ. □

THEOREM 6.3. *As such, the Hausdorff dimension of the set of exceptional times is at most* $31/36$.

Proof. Let U_n be the event that there is a time in $[0, 1/n]$ for which the origin percolates. Since the set of vertices which are open for some $t \in [0, 1/n]$ is an i.i.d. process with density $1/2 + (1 - e^{-1/(2n)})/2 \le 1/2 + 1/n$, it is immediate that

$$\mathbf{P}[U_n] \le \pi_{\frac{1}{2}+\frac{1}{n}}(\mathscr{C}_0),$$

where $\mathscr{C}_0$ is the event that the origin percolates. By page 3 of [31], for every $\varepsilon > 0$, there is a C so that

$$\pi_{\frac{1}{2}+\frac{1}{n}}(\mathscr{C}_0) \le C\, n^{\varepsilon - 5/36}. \tag{6.2}$$

Now let

$$N_n = \sum_{j=1}^{n} 1_{U_{j,n}},$$

where $U_{j,n}$ is the event that there is a time in $[(j-1)/n, j/n]$ for which the origin percolates (so that $U_{1,n} = U_n$ above). By the above, we have that $\mathbf{E}[N_n] \le C n^{\frac{31}{36}+\varepsilon}$.

It follows that

$$\lim_n \frac{\mathbf{E}\big[N_n\big]}{n^{\frac{31}{36}+2\varepsilon}} = 0$$

and so from Fatou's lemma, we get

$$\mathbf{E}\Big[\liminf_n \frac{N_n}{n^{\frac{31}{36}+2\varepsilon}}\Big] = 0.$$

Therefore

$$\liminf_n \frac{N_n}{n^{\frac{31}{36}+2\varepsilon}} = 0$$

a.s. This says that a.s. for infinitely many n, the set of exceptional times in $[0,1]$ at which the origin percolates can be covered by $n^{\frac{31}{36}+2\varepsilon}$ intervals of length $1/n$. Hence, the Hausdorff dimension of the set of these exceptional times is at most $\frac{31}{36}+2\varepsilon$ a.s. By countable additivity, we are done. □

Remark 6.4. The upper bound will be proved again by a different argument when we prove Theorem 1.16. The above proof is included here, because it is shorter. One should nonetheless point out that the above argument uses (6.2), while the argument below is more self-contained.

7. **Exceptional times for k-arm events**

In this section, we give the proofs of the lower bounds in Theorems 1.15 and 1.16, but generally omit those details which are the same as in the corresponding proofs of Theorems 1.3 and 6.1.

For $\theta > 0$ and integer $k \geq 1$, let $A^k_{W_\theta}(r,R)$ be the event that we have k disjoint crossings of alternating colors (with black most clockwise) between distances r and R of the origin in W_θ and let $\alpha^k_{W_\theta}(r,R) = \mathbf{P}\big[A^k_{W_\theta}(r,R)\big]$. If r is suppressed, then it is taken to be $10k$.

We will, of course, need the asymptotics of $\alpha^k_{W_\theta}(r,R)$. For this purpose, conformal invariance will be used. Although when $\theta > 2\pi$ the surface W_θ is not planar and conformal invariance is usually stated for planar domains, the proof of conformal invariance certainly holds in this setting. The asymptotic decay as $R/r \to \infty$ of the probability of k disjoint crossings between distances r and R in W_θ from the origin in the percolation scaling limit is determined by conformal invariance. Specifically, the map $z \mapsto z^{\pi/\theta}$ maps W_θ to the upper half-plane, and we may conclude from (3.8) that the decay (for the percolation scaling limit) is of the form $(R/r)^{\frac{-\pi k(k+1)}{6\theta}+o(1)}$, as $R/r \to \infty$ while k stays fixed. Then, one can conclude, as for the other exponents we have discussed, that for $R \geq r \geq 10k$,

$$\alpha^k_{W_\theta}(r,R) = (R/r)^{\frac{-\pi k(k+1)}{6\theta}+o(1)}, \tag{7.1}$$

as $R/r \to \infty$ while k is fixed, by the argument in [31]. We will also use the fact that the quasi-multiplicativity relations (3.1) and (3.2) hold for $\alpha^k_{W_\theta}$ and for α_2. This is proved in the appendix; see Remark A.6.

Proof of the lower bound in Theorem 1.15. We first handle the case $k = 1$ and therefore abbreviate temporarily $\alpha^1_{W_\theta}(r, R)$ by $\alpha_{W_\theta}(r, R)$. (A different approach will be needed for $k \geq 2$.) Let D be the union of the hexagons in W_θ that contain points whose distance from the origin is in $[r, R]$. Let $\partial_R D$ and $\partial_r D$ denote the set of points in ∂D that are at distance $\geq R$ (respectively, $\leq r$) from the origin. Also, we denote by $\partial^0 D$ and $\partial^\theta D$, the components of $\partial D \cap \partial W_\theta$ that are at angle about 0 (respectively, about θ) in radial coordinates on W_θ.

The algorithm we use to determine if there exists a crossing of D is essentially the same as the algorithm determining the existence of a left to right crossing of a square, where $\partial_R D$ plays the role of the right side of the square and where $\partial_r D$ plays the role of the left side of the square. (This is of course crucial; if we reversed things, then the hexagons near the inner circle would be revealed with too high a probability.) It is clear that this algorithm works and so we now need to compute its revealment. We will show that the revealment is

$$O(1)\,\alpha_2(r)\,\alpha_{W_\theta}(r, R)\,. \tag{7.2}$$

Using (7.1) and (3.6), one can show that this is essentially (i.e., up to some $O(1)$ factor) monotone decreasing in r in the relevant range $\theta > 8\pi/3$.

We just look at the first interface arising in the algorithm, the one which terminates when it hits $\partial_r D \cup \partial^\theta D$, since the estimates for the second interface will be essentially the same.

Fix some hexagon $H \subset D$. Let $s = \mathrm{dist}(H, \partial_R D \cup \partial^0 D \cup \{0\})$ with 0 denoting the origin. We also use $|p|$ to denote distance from 0 in W_θ. We distinguish several different cases.

Case 1: $\mathrm{dist}(H, \{0\}) = s$. For H to be visited, we need our 2-arms event holding within distance $s/2$ of H and a crossing of the desired color between distance $2s$ and distance R from the origin. These are independent and we get that H is visited with probability at most $\alpha_2(s/2)\,\alpha_{W_\theta}(2s, R)$. By the analogues of (3.1) and (3.2) for α_2 and α_{W_θ}, this is compatible with our claimed revealment (7.2).

Case 2: $\mathrm{dist}(H, \partial_R D) = s$. As in the proof of Theorem 4.4 with $c := \mathrm{dist}(p_0, H) \wedge (R/2)$, we obtain

$$\mathbf{P}\big[H \text{ visited } \big|\, p_0\big] \leq O(1)\,\alpha^+(2s + O(1), c - s - O(1))\,\alpha_2(s/2).$$

Proceeding as in that proof, we see that this is also compatible with our claimed revealment (7.2).

Case 3: $\operatorname{dist}(H, \partial^0 D) = s$. Let $w \in \partial^0 D$ so that $\operatorname{dist}(H, w) = s$. We separate Case 3 into three subcases.

Case 3(a): $s \geq |w|/2$. Then the triangle inequality gives $\operatorname{dist}(H, 0) \leq 3s$. For H to be visited, we need our 2-arms event holding within distance $s/2$ of H and a path of the desired type between distance $4s$ and distance R from the origin. These are independent and we get that H is visited with probability at most $\alpha_2(s/2)\,\alpha_{W_\theta}(3s, R)$, which is compatible with our claimed revealment (7.2).

Case 3(b): $s \leq |w|/2 \leq R/4$. For H to be visited, we need our 2-arms event holding within distance $s/2$ of H, a white crossing in the half-annulus centered at w with outer radius $|w|$ and inner radius $2s$ (which is identical to a half-annulus in a half-plane; $|w| \leq R/2$ guarantees that the above half-annulus does not intersect $\partial_R D$) and a white crossing between distance $2|w|$ and distance R from the origin. These are independent and we get that H is visited with probability at most

$$\alpha_2(s/2)\,\alpha_+(2s, |w|)\,\alpha_{W_\theta}(2|w|, R).$$

Since up to an $O(1)$ factor, $\alpha_2(s)\,\alpha_+(s, |w|)$ is increasing in s, the product of the first two terms is at most $O(1)\,\alpha_2(|w|)$ and since $|w| \geq 2s \geq r$, the whole product is at most

$$O(1)\,\alpha_2(r)\,\alpha_{W_\theta}(r, R).$$

Case 3(c): $|w| \geq R/2$; $s \leq |w|/2$. For H to be visited, we need our 2-arms event holding within distance $s/2$ of H and if $2s < d(p_0, w)$ it is also necessary that a white crossing occurs between distances $2s$ and $d(p_0, w) \wedge |w|$ from w. (Note that the latter event takes place in the upper half-plane.) These are independent and since $|w| \geq R/2$ we get

$$\mathbf{P}\big[H \text{ visited } \big|\, p_0\big] \leq O(1)\,\alpha^+\big(2s, d(p_0, w) \wedge (R/2)\big)\,\alpha_2(s/2)\,.$$

As in Case 2, this is compatible with (7.2).

This covers all possible cases, and hence establishes that the revealment is as claimed.

We now proceed to discuss the algorithm and the revealment when $k > 1$. It turns out simplest in fact to modify the event $A^k_{W_\theta}(r, R)$ as follows. Partition the outer boundary $\partial_R D$ into k arcs of roughly equal diameter $Y_1, Y_2, \ldots, Y_k$ (ordered counterclockwise) and let $\tilde{A}^k_{W_\theta}(r, R)$ be the event that for every odd (respectively, even) $i \in \{1, 2, \ldots, k\}$ there is a black (respectively, white) crossing in D from $\partial_r D$ to Y_i. Thus, instead of looking at the set of times for which $A^k_{W_\theta}(r_0, R)$ occurs (where $r_0 = 10k$, say), we will look at the set of times at which $\tilde{A}^k_{W_\theta}(r_0, R)$ occurs. Clearly, $\tilde{A}^k_{W_\theta}(r, R) \subset A^k_{W_\theta}(r, R)$, and therefore this is justified. We will also use

the relation

$$\mathbf{P}\big[A^k_{W_\theta}(r,R)\big] \le C^\theta_k \,\mathbf{P}\big[\tilde{A}^k_{W_\theta}(r,R)\big]\,, \tag{7.3}$$

for some constant C^θ_k, depending only on k and θ, which holds by Remark A.7.

If $Y \subset \partial_R D$ is an arc, let $A^1_Y(r,R)$ (respectively, $A^{-1}_Y(r,R)$) be the event that there is a white (respectively, black) crossing from Y to $\partial_r D$ in D. Suppose that for each $i = 1,2,\dots,k$, we have a partition $Y_i = Y_{i,+} \cup Y_{i,-}$ of Y_i into two arcs $Y_{i,+}$ and $Y_{i,-}$. Since $A^{\pm 1}_{Y_i}(r,R) = A^{\pm 1}_{Y_{i,+}}(r,R) \cup A^{\pm 1}_{Y_{i,-}}(r,R)$, we have

$$\tilde{A}^k_{W_\theta}(r,R) = \bigcup_{y\in\{-,+\}^k} \bigcap_{i=1}^{k} A^{(-1)^i}_{Y_{i,y_i}}(r,R)\,. \tag{7.4}$$

The algorithm starts out by picking points $x_i \in Y_i$, randomly, uniformly and independently. Then $Y_{i,+}$ and $Y_{i,-}$ are chosen as the two components of $Y_i \setminus \{x_i\}$. For each of the 2^k possible $y \in \{-,+\}^k$, the algorithm then proceeds to determine if the corresponding component

$$A(y) := \bigcap_{i=1}^{k} A^{(-1)^i}_{Y_{i,y_i}}(r,R)$$

of (7.4) has occurred. For that purpose, interfaces are started at each of the points x_i, and are followed until the event has been determined one way or the other. (Of course, the interface will have either white on the left and black on the right or vice versa, depending on the color of crossing it is meant to detect and whether the corresponding arc $Y_{i,\pm}$ is to the left or right of x_i.) However, the order in which the interfaces are extended is somewhat important. A simple rule that works is that among the hexagons necessary to extend the k interfaces one more step, the algorithm chooses the one that is farthest away from 0. The event $A(y)$ is decided positively only if all k interfaces reach $\partial_r D$.

The revealment of this algorithm is at most $k\,2^k$ times the maximum probability that the interface, started at x_i, visits a hexagon H before the determination of the corresponding $A(y)$ is terminated. Here, the maximum is over all hexagons $H \subset D$ and all $i \in \{1,2,\dots,k\}$. The corresponding bound is attained as in the case $k = 1$, but now $\alpha^1_{W_\theta}$ is replaced by $\alpha^k_{W_\theta}$. Our rule of thumb for selecting which interface to extend guarantees that we never examine a hexagon H unless $\tilde{A}^k_{W_\theta}\big(\mathrm{dist}(0,H) + O(1), R\big)$ occurred. As in the case $k = 1$, when estimating the revealment it is important that $\alpha_2(r,R) \le O(1)\,\alpha^k_{W_\theta}(r,R)$. In the range $\theta > 4\pi\,k\,(k+1)/3$, which is the relevant range for the lower bound in Theorem 1.15, this follows from (7.1).

The remainder of the proof goes through as before. □

Proof of lower bound in Theorem 1.16. Here we simply say that the proof for the lower bounds in Theorem 1.15 can be carried out in a similar way. In fact, for $k \geq 2$, the proof is simpler topologically than the $k = 1$ case for the plane, since we do not need to worry about interfaces making complete circuits around the origin (if this ever happens, the event in question cannot occur and we stop the algorithm). □

8. Upper bounds for k-arm times

The following result, which will be useful for the proofs of the upper bounds in Theorems 1.15 and 1.16, is abstract: the graph structure does not play any role. Let A be an event involving independent Bernoulli $(1/2, 1/2)$ random variables $X_1, X_2, \ldots, X_m$. Recall that the influence of the index i on A, denoted $I_i(A)$, is the probability that X_i is pivotal; namely, that changing the value of X_i changes whether A occurs or not. The sum of the influences is denoted by $I(A) = \sum_i I_i(A)$.

THEOREM 8.1. *Let $\{A_n\}_{n\geq 1}$ be some sequence of events in $\{0,1\}^V$, each depending on only finitely many coordinates. Assume that $\lim_{n\to\infty} \mathbf{P}[A_n] = 0$. Let $\{\omega_t\}$ be the Markov process on $\{0,1\}^V$ where independently 0's go to 1 at rate 1, 1's go to 0 at rate 1 and started according to its stationary distribution $\pi_{\frac{1}{2}}$. Let T be the set of exceptional times t at which $\omega_t \in \bigcap_{n\geq 1} A_n$. If $\liminf_{n\to\infty} I(A_n) < \infty$, then $T = \varnothing$ a.s. Otherwise, the Hausdorff dimension of T is a.s. at most*

$$\liminf_{n\to\infty} \left(1 - \frac{\log \mathbf{P}[A_n]}{\log I(A_n)}\right)^{-1}. \tag{8.1}$$

Proof. Let $T_n := \{t \in [0,1] : \omega_t \in A_n\}$, let ∂T_n be the boundary points of T_n in $(0,1)$ and set $N_n := |\partial T_n|$. We claim that

$$\mathbf{E}\big[N_n\big] = I(A_n)/2. \tag{8.2}$$

To see this, write $N_n = \sum_v N_n^v$ where N_n^v counts the number of elements in ∂T_n at which time the vertex v flipped. We now need to show that, for each vertex v, $\mathbf{E}\big[N_n^v\big]$ is $I_v(A_n)/2$. Given a time interval $[t, t+dt]$, the probability that there is a time point in the interval which contributes to N_n^v is equal to $I_v(A_n)\,dt/2 + o(dt)$ and the probability of $k \geq 2$ such time points is clearly $O(dt^k)$. From this, (8.2) easily follows.

For any $\varepsilon > 0$, let T_n^ε be the ε-neighborhood of T_n intersected with $[0,1]$. Since $T_n^\varepsilon \subseteq T_n \cup \bigcup_{x\in\partial T_n} [x-\varepsilon, x+\varepsilon]$,

$$\mu(T_n^\varepsilon) \leq \mu(T_n) + 2\,N_n\,\varepsilon\,, \tag{8.3}$$

where μ denotes Lebesgue measure. For any set U and $\varepsilon > 0$, let $\mathcal{N}(U,\varepsilon)$ denote the number of ε intervals needed to cover U. From the above, using the fact that

the intervals comprising T_n^ε all have length at least ε, it follows that $\mathcal{N}(T_n^\varepsilon, \varepsilon) \le 2\,\mu(T_n^\varepsilon)\,\varepsilon^{-1}$, and so, using (8.3),

$$\mathcal{N}(T_n, \varepsilon) \le \mathcal{N}(T_n^\varepsilon, \varepsilon) \le 2\,\mu(T_n)\,\varepsilon^{-1} + 4\,N_n\,.$$

Therefore, by Fubini's theorem and (8.2),

$$\mathbf{E}\big[\mathcal{N}(T_n, \varepsilon)\big] \le 2\,\mathbf{P}[A_n]\,\varepsilon^{-1} + 2\,I(A_n)\,. \tag{8.4}$$

We temporarily assume that $\liminf_{n\to\infty} I(A_n) = \infty$. Let $a_n = \mathbf{P}\big[A_n\big]/I(A_n)$, which goes to 0 as $n \to \infty$. By (8.4), we have

$$\mathbf{E}\big[\mathcal{N}(T, a_n)\big] \le \mathbf{E}\big[\mathcal{N}(T_n, a_n)\big] \le 4\,I(A_n)\,. \tag{8.5}$$

By passing to a subsequence if necessary, we assume with no loss of generality that the lim inf in (8.1) is a limit. Let L denote the value of that limit. It is elementary to check that for every $\varepsilon > 0$, for all sufficiently large n,

$$I(A_n) \le \left(\frac{I(A_n)}{\mathbf{P}[A_n]}\right)^{L+\varepsilon}.$$

This together with (8.5) implies that the Hausdorff dimension of T is at most $L + \varepsilon$ a.s. As ε is arbitrary, this completes the proof in the case $I(A_n) \to \infty$.

Since $T_n \ne \varnothing$ implies that $N_n \ge 1$ or $T_n \supseteq (0, 1)$, it follows by (8.2) and Markov's inequality that

$$\mathbf{P}\big[T_n \ne \varnothing\big] \le \mathbf{P}\big[A_n\big] + I(A_n)\,.$$

Thus, $T = \varnothing$ a.s. when $\liminf_n I(A_n) = 0$.

The case $\liminf_n I(A_n) \in (0, \infty)$ requires a different argument. Let $\varepsilon_n = \sqrt{\mathbf{P}[A_n]}$. By (8.4), we have $\liminf_{n\to\infty} \mathbf{E}\big[\mathcal{N}(T_n, \varepsilon_n)\big] < \infty$. Since $\varepsilon_n \to 0$, the cardinality $|T|$ of T is bounded from above by $\liminf_{n\to\infty} \mathcal{N}(T_n, \varepsilon_n)$. Fatou's lemma yields $\mathbf{E}\big[|T|\big] < \infty$ and hence $\mathbf{P}\big[|T| < \infty\big] = 1$. We finally conclude that $\mathbf{P}[T \ne \varnothing] = 0$ by combining [11, Th. 6.7] and [10, (2.9)]. □

Proof of Theorem 1.16. Since the lower bounds have been established in Section 7, it remains to prove the upper bounds. Fix $k = 1$ or $k > 1$ even. Let A_R be the event that there are k disjoint crossings of the annulus $D_R := \{z \in C_\theta : 10k \le |z| \le R\}$, where we require that the colors be alternating if $k \ne 1$. Here, $|z|$ denotes the distance to 0, which is the apex of the cone C_θ. One can prove that

$$\mathbf{P}\big[A_R\big] = \begin{cases} R^{-5\pi/(24\theta)+o(1)} & k = 1, \\ R^{(1-k^2)\pi/(6\theta)+o(1)} & k > 1, \end{cases} \tag{8.6}$$

in the very same way that we have justified (7.1). By Theorem 8.1 (and easy algebraic manipulation), it therefore suffices to prove that

$$I(A_R)/\mathbf{P}[A_R] \le R^{3/4+o(1)}. \tag{8.7}$$

Let H be a hexagon in C_θ, and let $s = s(H)$ be the distance from H to 0. For H to be pivotal it is necessary that there be k disjoint (alternating, if $k > 1$) crossings from distance $10k$ to $s/2$ from the origin (unless $s/2 \le 10k$) and between distances $2s$ and R (unless $2s \ge R$). Likewise, there should be four alternating crossings between H and distance $(s/2) \wedge \mathrm{dist}(H, \partial D_R)$ from H. These events are independent. Using the quasi-multiplicative property of the k-arm crossing events (Remark A.6) and (3.7) with $k = 4$, this gives (when $s < 8R/9$),

$$I_H(A_R) \le O(1)\,\mathbf{P}[A_R]\, s^{-5/4+o(1)} \tag{8.8}$$

where this $O(1)$ factor (as well as those appearing below) may depend on k and θ. Since the number of hexagons in C_θ satisfying $s = s(H) < \rho$ is $O(\rho^2)$, an easy calculation yields

$$\sum_{H: s(H) < 8R/9} I_H(A_R) \le O(1)\,\mathbf{P}[A_R]\, R^{3/4+o(1)}.$$

Now suppose that H is a hexagon satisfying $s(H) \ge 8R/9$. For H to be pivotal, it is necessary that there be k (alternating, if $k > 1$) crossings in C_θ between $\{|z| = 10k\}$ and $\{|z| = R/2\}$, there should be four alternating crossings between H and distance $\mathrm{dist}(H, \partial D_R)/2$ from H, and there should be three alternating crossings between distance 2 $\mathrm{dist}(H, \partial D_R)$ and distance $R/2$ from a point on ∂D_R closest to H. The latter event is governed by the 3-arm half-plane exponent, whose asymptotic behaviour is described by (3.8). Since $s + \mathrm{dist}(H, \partial D_R) = R + O(1)$, we get

$$I_H(A_R) \le O(1)\,\mathbf{P}[A_R]\,(R-s)^{-5/4+o(1)}\,((R-s)/R)^{2+o(1)}.$$

Since for $b \ge 1$ there are $O(bR)$ hexagons at distance $\le b$ from $\{|z| = R\}$, another easy calculation gives

$$\sum_{H: s(H) \ge 8R/9} I_H(A_R) \le O(1)\,\mathbf{P}[A_R]\, R^{3/4+o(1)}.$$

Together, this yields (8.7) and the proof is complete. □

Proof of Theorem 1.15. The lower bound was proved in Section 7, and so only the upper bound needs to be justified. The proof proceeds like the proof of the upper bound in Theorem 1.16, except that the influence estimates are slightly different.

Let $D_R = \{z \in W_\theta : 10k \le |z| \le R\}$, A_R be the k-arm event in W_θ between $\{z : |z| = 10k\}$ and $\{z : |z| = R\}$, and $H \subset D_R$ be a hexagon. Let $s = s(H) = \mathrm{dist}(0, H)$, and let $b = b(H) = \mathrm{dist}(H, \partial D_R)$, where we write ∂D_R for the boundary of D_R in C_∞; i.e., the points on ∂W_θ are included. For H to be pivotal for A_R, it is necessary that the k-arm event holds between distance $10k$ and $s/2$ from 0

(unless $s/2 \leq 10k$) as well as between distances $2s$ and R (unless $2s \geq R$), that the alternating 4-arm event holds between H and distance $b/2$ away from H, and that the alternating 3-arm event must hold between distances $2b$ and $s/4$ away from a point in ∂D_R closest to H (unless $2b \geq s/4$). There are $O(b's')$ hexagons H satisfying $b(H) \leq b'$ and $s(H) \leq s'$. The rest of the proof proceeds like that of Theorem 1.16, and is left to the reader. □

Proof of Theorems 1.10, 1.12, 1.13 *and* 1.14. At any time at which there are two infinite white clusters in the plane, we must also have the 4-arm event occurring (with alternating colors) but by Theorem 1.16 (with $k = 4$ and $\theta = 2\pi$), there are no such times. This proves Theorem 1.10.

At any time at which there are two infinite different colored clusters, we must also have the 2-arm event occurring (with different colors) but by Theorem 1.16 (with $k = 2$ and $\theta = 2\pi$), the set of such times has Hausdorff dimension at most $2/3$. This proves Theorem 1.12. The other two theorems are similarly proved. □

9. **The square lattice**

We start this section by proving Theorem 1.11. Afterwards, possible ways in which our arguments for Theorem 1.3 may be improved to apply to $\mathbb{Z}^2$ as well, will be discussed.

In the proof of Theorem 1.11 we will use the fact that the six alternating arms exponent is larger than 2, or, more precisely, that the probability for six alternating arms between radii r and R is bounded above by $O(1)\,(r/R)^{2+\varepsilon}$ for some $\varepsilon > 0$. This is essentially due to [20, Lemma 5], but a proof is also given in the appendix (Corollary A.8).

Proof of Theorem 1.11. For $0 < r < R$, let $S(r, R)$ be the event that there are three different clusters that connect the circles of radii r and R about 0. By the above mentioned bound on the alternating 6-arm probabilities, We may choose some fixed $\varepsilon > 0$ and some function $\rho = \rho(r) > r$ such that for static critical bond percolation on $\mathbb{Z}^2$, for all r,

$$\mathbf{P}\big[S(r, \rho)\big] \leq \rho^{-2-\varepsilon}. \tag{9.1}$$

Consider some bond e, and let $F(e)$ be the event that e is pivotal for $S(r, \rho)$. Then $\mathbf{P}\big[F(e)\big]$ is just the influence $I_e(S(r, \rho))$. Assume that $\mathbf{P}\big[F(e)\big] \neq 0$. Note that the events $F(e)$ and $\{e \text{ is open}\}$ are independent events. This implies that $\mathbf{P}\big[S(r, \rho) \mid F(e)\big] = 1/2$, which one may write

$$\mathbf{P}\big[F(e) \cap S(r, \rho)\big] = \mathbf{P}\big[F(e) \cap \neg S(r, \rho)\big].$$

Since this applies to every bond e, we conclude that the expected number of pivotals on the event $S(r, \rho)$ is half of the total influence $I(S(r, \rho))$. However the number

of pivotals for $S(r,\rho)$ is bounded by the total number of edges intersecting the disk of radius ρ about the origin, which is certainly $O(\rho^2)$. Thus,

$$I(S(r,\rho)) \le 2\,\mathbf{P}\big[S(r,\rho)\big]\,O(\rho^2) = O(\rho^{-\varepsilon}).$$

Consequently, by Theorem 8.1, for every $r_0 > 0$ a.s. there are no exceptional times in which $\bigcap_{r>r_0} S(r,\rho(r))$ holds. This proves our theorem. □

Remark 9.1. An alternative way to prove the above result is based on using the fact that the 6-arm exponent is strictly larger than 2 together with the fact that the number of different configurations (counting repetitions) that appear in a ball of radius n during the time interval $[0,1]$ has a Poisson distribution with a parameter which is at most $O(1)n^2$.

As we will briefly explain below, the proof of Theorem 1.3 almost works for bond percolation on the square grid. In fact, there are several alternative routes by which the result might perhaps be extended to $\mathbb{Z}^2$:

(1) Establishing

$$\alpha_2(r) \le O(r^{-\varepsilon})\,\alpha(r)^2 \tag{9.2}$$

for $\mathbb{Z}^2$ for some fixed $\varepsilon > 0$,

(2) improving the estimate (1.3),

(3) proving the existence of an algorithm (or a witness which would still permit the use of Theorem 1.8) with smaller revealment,

(4) extending Smirnov's theorem to $\mathbb{Z}^2$.

Note that the weaker version of (9.2) $\alpha_2(r) \le \alpha(r)^2$ follows from either the Harris-FKG inequality or Reimer's inequality [27]. Kesten and Zhang have proved some related strict inequalities between exponents [21], but it seems that their methods are not sufficient to prove (9.2).

We now explain why (9.2) in the $\mathbb{Z}^2$ setting implies exceptional times for $\mathbb{Z}^2$. First we want to have the revealment for the algorithm determining f_r^R bounded by $O(1)\,\alpha_2(r)\,\alpha(r,R)$. One problem seems to be that the bound on the revealment for the triangular grid involves the summand featuring α^+, which is relatively negligible, while on $\mathbb{Z}^2$, we do not know how to prove that the other summand dominates. The fix is to replace the deterministic R by a random $R'_t \in [R, 2R]$. The random variable R'_t will depend on some extra random bits, that we add, and these random bits also evolve in time. We construct the dependence of R' on these bits so that R' can be calculated by an algorithm with very small revealment. This is rather easy to arrange, because we are not limited in the number of bits that we may take. If we consider an edge whose distance from the

origin a is in the range $[R/2, 2R]$, then the probability that the edge is examined given R' is at most $O(1)\,\alpha_2(R'-a)\,1_{R'\geq a-1}$. By (A.1), this is at most $O(1)\,\alpha_2(R)\,\alpha_2(R'-a,R)^{-1}\,1_{R'\geq a-1}$. The probability that $|R'-a|\leq 2^j$ is at most $O(1)\,2^j/R$. We also know that $\alpha_2(r_1,r_2)^{-1}\leq O(1)\,(r_2/r_1)^{1-\varepsilon'}$ for some $\varepsilon'>0$, by Reimer's inequality [27] and (A.5). It follows that the probability that such an edge is examined is $O(1)\,\alpha_2(R)$. The $r^{o(1)}$ factor in Theorem 4.4 is easily replaced by an $O(1)$ factor, if we use Proposition A.1 in the course of the proof. Then we get (5.3) for the square grid, but without the $r^{o(1)}$ factor. We may then choose the dependence between r and t such that $\alpha(r)\approx t\,r^{\varepsilon/2}$, where ε is the constant in (9.2). The rest is immediate from (5.3), since clearly $r^{-\varepsilon/2}\leq O(1)\,t^{\varepsilon'}$ for some $\varepsilon'>0$.

A consequence of this argument, which applies without assuming (9.2), is that for bond percolation on $\mathbb{Z}^2$ we have

$$\mathbf{P}\big[V_{t,R}\cap V_{0,R}\big]\leq O(t^{-1})\,\mathbf{P}\big[V_{t,R}\big]^2.$$

This gives yet another illustration as to how close the result for $\mathbb{Z}^2$ seems to be — if the t^{-1} term was improved to $t^{-1+\varepsilon}$, that would have been enough. Consequently, significant improvements in the algorithm or in (1.3) would also be sufficient.

10. **Some open questions**

Following are a few questions and open problems suggested by the present paper.

(1) For the results in Theorems 1.15 and 1.16, what is the Hausdorff dimension of the set of exceptional times in question? We tend to believe that the answer is the upper bound. In particular, is the upper bound of 31/36 in Theorem 1.9 the correct answer?

(2) Prove that there exist exceptional times for percolation on the square lattice (see Section 9 for a discussion).

(3) What is the best γ for which Theorems 1.6 and 1.7 hold?

(4) What is the best revealment of an algorithm determining the event Q_R in Theorem 4.1?

(5) What is the sharp form of Theorem 1.8?

(6) What are the properties of the infinite cluster at an exceptional time at which it exists? For example, what is the growth rate of the number of vertices in the Euclidean disk of radius r around the origin which belong to the cluster of the origin at the first time $t\geq 0$ in which the cluster is infinite? Is the growth rate the same at all exceptional times?

(7) What is the relationship between the exceptional infinite cluster and the incipient infinite cluster?

Note added in proofs. Questions (1), (2), and part of (3) have been answered by C. Garban, G. Pete, and O. Schramm. Question (7) has been answered by A. Hammond, G. Pete, and O. Schramm.

Appendix A. Quasi-multiplicativity

In this appendix, we discuss the k-arm probabilities and prove that they satisfy the corresponding analogue of the relation (3.1). For $R > 0$, let H_R be the union of the hexagons intersecting $B(0, R)$. For $R > r > 0$ let $A_j(r, R)$ denote the event that there are at least j crossings from ∂H_r to ∂H_R, of alternating colors. The following result refers to critical site percolation on the triangular grid and critical bond percolation on the square grid.

PROPOSITION A.1. *Let $j > 0$ be even. There is a constant C, depending only on j, such that for all $r < r' < r''$*

$$C^{-1}\,\mathbf{P}\big[A_j(r, r'')\big] \le \mathbf{P}\big[A_j(r, r')\big]\,\mathbf{P}\big[A_j(r', r'')\big] \le C\,\mathbf{P}\big[A_j(r, r'')\big], \tag{A.1}$$

and $\mathbf{P}\big[A_j(r, 2r)\big] > 1/C$ if $\mathbf{P}\big[A_j(r, 2r)\big] > 0$ (i.e., if r is large enough to allow j crossings). Moreover, a corresponding statement holds for critical bond percolation on the square grid which alternate between primal and dual crossings.

This theorem would have been a useful tool in [31], had it been available. Instead, the authors of that paper proved a weaker form of this which was good enough for their purposes. Our proof below uses techniques from [19], [24] and [31]. Indeed, the entire results of the appendix do follow from the ideas of [19]. We include them here for completeness, and for ease of reference. Additionally, though the basic ideas are the same, in several respects our treatment is a bit different from [19].

Below, we will work in the setting of the triangular grid. The proof for the square grid is essentially the same. In the setting of the triangular grid, there is the color exchange trick [20], [1], which shows that the probability for having alternating crossings is comparable to the probability of any color sequence as long as both colors are present. In the setting of the square grid, as far as we know, such a trick does not exist. At the end of the appendix we will explain how the proof of Proposition A.1 can be generalized to any color sequence.

In the following, an interface from ∂H_r to ∂H_R is an oriented simple path in the hexagonal grid that has one color of hexagons adjacent to it on the right, and the opposite color adjacent to it on the left. Thus, it is the common boundary of a black crossing and a white crossing.

For $R > r > 1$, consider all the interfaces crossing from ∂H_r to ∂H_R, and define $s(r, R)$ to be the least distance between any pair of endpoints of two interfaces on ∂H_R. If there are no interfaces, we take $s(r, R) = \infty$. Note that $s(r, R)$ is monotone nonincreasing in r. This quantity will roughly measure the "quality" of the interfaces; when $s(r, R)$ is comparable to R, the interfaces are well separated, and, as we will see, easier to extend.

LEMMA A.2. *For all $a \in (0, 1)$, $R > 0$, $\delta > 0$,*

$$\mathbf{P}\big[s(a\,R, R) < \delta\,R\big] \le C\,\delta^{\varepsilon},$$

where $C = C(a)$ is a constant depending only on a and $\varepsilon > 0$ is an absolute constant.

The lemma probably follows from [19, Lemma 2], but since the proof is rather short, we include a proof for completeness.

Proof. We prove this in the case $a = 1/2$. The general case is essentially the same. Let $\alpha \subset \partial H_R$ be an arc of diameter $R/3$, and let Y be the set of points in H_R at distance at most $R/3$ from α. Let α_1 be one of the two arcs in $\partial Y \cap \partial H_R \setminus \alpha$. Let $\beta_1, \beta_2, \ldots, \beta_k$ be the interfaces crossing from $\partial Y \setminus \partial H_R$ to α, ordered so that for $i_1 < i_2 \le k$, the interface β_{i_1} separates α_1 from β_{i_2} in $\overline{Y}$. Fix a positive integer i and condition on $i \le k$ and on β_i. Let $\hat{\beta}_i$ denote the union of the hexagons adjacent to β_i. Then the percolation in the connected component Y_i of $Y \setminus \hat{\beta}_i$ separated from α_1 by β_i remains unbiased. Suppose that the hexagons in $\hat{\beta}_i$ adjacent to Y_i are white, say. Let z_i denote the endpoint of β_i on α. The RSW theorem implies that there is some constant $\varepsilon > 0$ such that with conditional probability $1 - O(1)\,\delta^{\varepsilon}$ there is a white crossing in $Y_i \setminus B(z_i, \delta\,R)$ from $\partial\hat{\beta}_i$ to ∂H_R. On that event, it is clear that if $k \ge i + 1$, then $|z_i - z_{i+1}| \ge \delta\,R$. We conclude that

$$\mathbf{P}\Big[k \ge i+1,\ |z_i - z_{i+1}| \le \delta\,R \,\Big|\, k \ge i,\ \beta_i\Big] \le O(1)\,\delta^{\varepsilon}.$$

The RSW theorem also implies that there is conditional probability bounded away from zero that $k = i$ given $k \ge i$ and β_i (this would be guaranteed by an appropriate crossing in Y_i from $\partial\hat{\beta}_i$ to $\partial H_R \setminus \alpha$). Therefore, $\mathbf{P}\big[k \ge i\big] \le c^i$ for some constant $c < 1$. Thus,

$$\begin{aligned}\mathbf{P}\Big[k \ge i+1,\ |z_i - z_{i+1}| \le \delta\,R\Big] \\ = \mathbf{P}\Big[k \ge i+1,\ |z_i - z_{i+1}| \le \delta\,R \,\Big|\, k \ge i\Big]\,\mathbf{P}\big[k \ge i\big] = O(1)\,c^i\,\delta^{\varepsilon}.\end{aligned}$$

We sum this over all $i = 1, 2, \ldots$, and over an appropriate covering of ∂H_R by $O(1)$ arcs α of diameter $R/3$. The lemma follows. □

Next, we prove a statement saying, roughly, that if the crossings are "reasonably good", then there is a conditioned probability bounded away from zero that they extend and the extensions are "very good".

LEMMA A.3. *For every $j > 0$ even, there is a constant $\bar{\delta} = \bar{\delta}(j) > 0$ such that for every $\delta > 0$ there is some constant $c(\delta) > 0$, depending only on δ, such that when $R > r$,*

$$\mathbf{P}\Big[A_j(r, 4R) \cap \{s(r, 4R) > 4\bar{\delta}\, R\} \;\Big|\; A_j(r, R) \cap \{s(r, R) > \delta\, R\}\Big] > c(\delta)\,.$$

Proof. Set $S = \overline{H_R \setminus H_r}$. We assume that $A_j(r, R)$ holds and that $s(r, R) > \delta\, R$. Let $\gamma_0, \ldots, \gamma_{k-1}$ ($k \geq j$) be the collection of all interfaces crossing from ∂H_r to ∂H_R, in counterclockwise order, where we choose the indexes so that γ_0 has white hexagons on the right-hand side. (The interfaces are oriented from ∂H_r to ∂H_R.) In the following, we set $\gamma_i := \gamma_{i'}$ when $i \notin \{0, 1, \ldots, k-1\}$ and $i' = i \bmod k$. Set $\Gamma = \bigcup_{i \in \mathbb{N}} \gamma_i$. How does conditioning on the interfaces $\gamma_0, \ldots, \gamma_{k-1}$ affect the percolation process? Note the fact that there are no more than k interfaces means that for each $i \in \mathbb{N}$ there is a white crossing in $S \setminus \Gamma$ from the right-hand side of γ_{2i} to the left side of γ_{2i-1} and a black crossing in $S \setminus \Gamma$ from the left side of γ_{2i} to the right side of γ_{2i+1}. Otherwise, the configuration is unbiased on hexagons that are not adjacent to these interfaces.

Let z_i be the endpoint of γ_i on ∂H_R, $i \in \mathbb{N}$. For $i = 0, 1, \ldots, j-1$, let w_i be a point ∂H_R that is roughly in the center of the counterclockwise arc from z_i to z_{i+1}. Then $|w_i - z_i| \geq \delta\, R/5$, and the same is true for $|w_i - z_{i+1}|$. It is easy to see that there exist disjoint simple paths $\beta_0, \ldots, \beta_{j-1}$ satisfying the following. (See Figure A.1.) (1) Each β_i is a path in $\overline{H_{4R} \setminus H_R}$ from w_i to a point $w_i' \in \partial H_{4R}$. (2) The

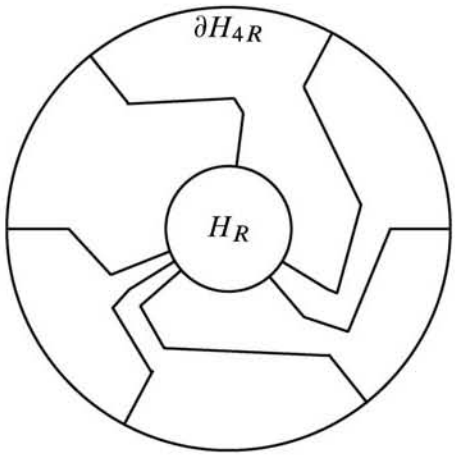


Figure A.1. The paths β_i.

points w_i' are roughly equally spaced on ∂H_{4R}. (3) Each $\beta_i \cap H_{2R}$ is contained in the line through the origin containing w_i, and each $\beta_i \setminus H_{3R}$ is contained in the line through the origin containing w_i'. (4) The distance from each of these paths to any other path is at least $c_1\,\delta\,R$, where $c_1 \in (0, 1/5)$ is some universal constant. (5) Each β_i has length at most a constant times R, where the constant may depend on j. For each $i \in \{0, 1, \ldots, j-1\}$ let α_i be the arc of a circle whose center is z_i, that has w_i as endpoint, that has the other endpoint on $\gamma_i \cup \gamma_{i+1}$, that is otherwise disjoint from $\gamma_i \cup \gamma_{i+1}$ and is contained in S.

Let $\hat{\beta}_i$ be the connected component containing w_i' of the complement of Γ in the $c_1\,\delta\,R/20$ neighborhood of $\beta_i \cup \alpha_i$. If $i \in \{0, 1, \ldots, j-1\}$ is odd, let F_i denote the event that there is a white crossing in $\hat{\beta}_i$ from Γ to ∂H_{4R}. Similarly, if i is even, let F_i denote the event that there is a black crossing in $\hat{\beta}_i$ from Γ to ∂H_{4R}. It is easy to see that if $\bigcap_{i=0}^{j-1} F_i$ holds, then $A_j(r, 4R)$ holds as well. The RSW theorem implies that $\mathbf{P}\big[F_i\big]$ is bounded from below (depending on δ) for a percolation process that is unbiased. But, as we have seen, the percolation on S between γ_i and γ_{i+1} is only conditioned on having a crossing of the appropriate color. By the Harris-FKG inequality, this is positively correlated with F_i. By independence on disjoint sets, the different F_i are independent given Γ (assuming, as we may, that the distance between the different sets $\hat{\beta}_i$ is significantly larger than the scale of the lattice). We conclude that for some $c(\delta) > 0$,

$$\mathbf{P}\Big[A_j(r, 4\,R) \;\Big|\; A_j(r, R) \cap \{s(r, R) > \delta\,R\}\Big] > c(\delta)\,.$$

Taking care of the condition $s(r, 4R) \geq 4\,\bar{\delta}\,R$ is not too hard. Suppose that in the above we truncate the paths β_i and the neighborhoods $\hat{\beta}_i$ by intersecting them with $H_{3.5\,R}$. We then condition on the "leftmost" crossing in $\hat{\beta}_i$. See Figure A.2. All this takes place within $H_{3.5\,R}$. The conditional probability that these crossings in the $\hat{\beta}_i$'s connect to $\partial H_{4\,R}$ is bounded away from zero by a function of j only (specifically, not δ). Thus, Lemma A.2 and the monotonicity of $s(r, R)$ in r shows that if $\bar{\delta} = \bar{\delta}(j) > 0$ is chosen small, with conditional probability at least $1/2$ we are also likely to have $s(r, 4R) \geq 4\,\bar{\delta}\,R$, as required. □

Set

$$f(r, R) := \mathbf{P}\big[A_j(r, R)\big], \qquad g_\delta(r, R) := \mathbf{P}\big[A_j(r, R) \cap \{s(r, R) > \delta\,R\}\big].$$

LEMMA A.4. *There is a constant $C_1(j) > 0$, depending only on j, such that for $R \geq 4\,r$*

$$C_1(j)\,g_{\bar{\delta}}(r, R) \geq f(r, R)\,,$$

where $\bar{\delta} = \bar{\delta}(j)$ is the constant introduced in Lemma A.3.

Proof. We assume that $f(r, R) > 0$. Let $\delta > 0$ be small. Set $N = \log_4(R/r)$ and let $m = m_\delta$ be the largest integer in the range $0 \leq m \leq N-1$ such that

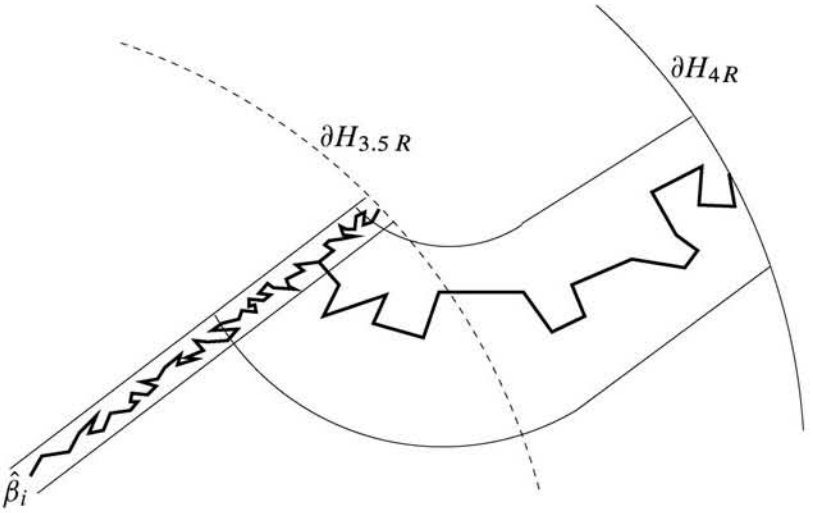


Figure A.2. A bent strip connecting with a channel. Indicated are the leftmost crossing of the intersection of the bent strip and the channel that connects to H_r and a reasonably likely crossing from it to ∂H_{4R} in the bent strip.

$g_\delta(r, 4^{-i}R) \le f(r, 4^{-i}R)/2$ holds for every integer i in the range $0 \le i < m$. Lemma A.2 and independence on disjoint sets imply

$$f(r, R) - g_\delta(r, R) \le C\,\delta^\varepsilon\, f(r, R/4),$$

and repeated applications of this inequality give

$$f(r, R) \le (2C)^m\,\delta^{\varepsilon m} f\left(r, 4^{-m}R\right). \tag{A.2}$$

We claim that if δ is a sufficiently small positive constant, then

$$f(r, 4^{-m}R) \le C_2(j)\, g_\delta(r, 4^{-m}R) \tag{A.3}$$

for some constant $C_2(j)$ depending only on j. If $m \le N-2$, this follows with $C_2(j) = 2$ from the definition of m. If $N-2 < m \le N-1$, then RSW easily implies $f(r, 4^{-m}R) \ge C_3(j)$ for some constant $C_3(j) > 0$, and $f(r, 4^{-m}R) - g_\delta(r, 4^{-m}R) \le C\,\delta^\varepsilon$ by Lemma A.2, which gives (A.3). On the other hand, repeated application of Lemma A.3 gives

$$c(\delta)\, c(\bar{\delta})^{m-1}\, g_\delta(r, 4^{-m}R) \le g_{\bar{\delta}}(r, R) \le f(r, R). \tag{A.4}$$

On combining this with (A.2) and (A.3), one obtains

$$c(\bar{\delta})\, C_2(j)\left(2\,C\,\delta^\varepsilon / c(\bar{\delta})\right)^m \ge c(\delta).$$

We choose δ sufficiently small so that $\delta^{\varepsilon} < c(\bar{\delta})/(4C)$. Then the above shows that m is bounded by a function of δ and j. The proof is now completed by combining (A.2), (A.3) and (A.4). □

Proof of Proposition A.1. The proof will be given only for the triangular grid, since the proof in the setting of bond percolation on the square grid is essentially the same. As we remarked above, when $\mathbf{P}\big[A_j(r,2r)\big] > 0$, the RSW theorem easily gives $\mathbf{P}\big[A_j(r,2r)\big] > 1/C(j)$, for some $C(j) > 0$.

We now assume that $r' > 8r$ and $r'' > 8r'$. Suppose that $A_j(r,r'/2) \cap \{s(r,r'/2) > \bar{\delta}\, r'/2\}$ holds. We also assume that the corresponding event occurs between $\partial H_{2r'}$ and $\partial H_{r''}$, but now we require that the interfaces be well separated on the inner boundary $\partial H_{2r'}$, instead of on the outer boundary. As in the proof of Lemma A.3, it is not too hard to see that conditioned on these events there is probability bounded away from zero (by a function of j) that the crossings between ∂H_r and $\partial H_{r'/2}$ will hook up nicely with the crossings between $\partial H_{2r'}$ and $\partial H_{r''}$. Basically, we only need to arrange that the channels $\hat{\beta}_i$ for the inner crossings will cross the corresponding channels of the outer crossings. The proof of the right-hand inequality in (A.1) now follows from Lemma A.4 and the corresponding statement for crossings with interfaces well-separated in the inner boundary, which is proved in the same way. If $r'' \le 8r'$, then the right-hand inequality in (A.1) is a consequence of Lemmas A.3 and A.4. A similar proof applies when $r' \le 8r$.

It now remains to prove the left-hand inequality in (A.1). If $r'' < 2r'$, then $\mathbf{P}\big[A_j(r',r'')\big]$ is bounded away from zero (if we assume $\mathbf{P}\big[A_j(r,r'')\big] > 0$) and we are done since $\mathbf{P}\big[A_j(r,r'')\big] \le \mathbf{P}\big[A_j(r,r')\big]$. Otherwise, we argue that

$$\mathbf{P}\big[A_j(r,r'')\big] \le \mathbf{P}\big[A_j(r,r')\big]\,\mathbf{P}\big[A_j(2r',r'')\big],$$

by independence on disjoint subsets, and

$$\mathbf{P}\big[A_j(r',2r')\big]\,\mathbf{P}\big[A_j(2r',r'')\big] \le C\,\mathbf{P}\big[A_j(r',r'')\big],$$

by the right-hand inequality in (A.1). Since $\mathbf{P}\big[A_j(r',2r')\big]$ is bounded away from zero, the left-hand inequality now follows. □

We now generalize Proposition A.1 to arbitrary sequences of crossings.

PROPOSITION A.5. *Let $j \ge 1$ be an integer, and fix a color sequence $X \in \{\text{black, white}\}^j$. The probabilities for the existence of j disjoint crossings whose colors match this sequence in counterclockwise order also satisfy the inequalities in Proposition* A.1. *The corresponding statement also holds in the setting of critical bond percolation on the square grid.*

Proof. We start with the easier case where all the colors in the sequence X are the same, say black. Suppose that $r' > 2r$ and $r'' > 2r'$. Consider the event that (a) there are j disjoint black crossings from ∂H_r to $\partial H_{r'}$ and (b) there are

j disjoint black crossings from $\partial H_{r'}$ to $\partial H_{r''}$ and (c) there are j disjoint black circuits separating $\partial H_{r'/2}$ from $\partial H_{r'}$ and (d) there are j disjoint black circuits separating $\partial H_{r'}$ from $\partial H_{2r'}$ and (e) there are j disjoint crossings from $\partial H_{r'/2}$ to $\partial H_{2r'}$. Note that if we choose any one path in each of (a)–(e), we can extract from the union a crossing from ∂H_r to $\partial H_{r''}$. To see that we actually have j disjoint crossings from ∂H_r to $\partial H_{r''}$, note that if we remove any $j-1$ hexagons, then there is still one path remaining in each of (a)–(e), and consequently, there is still a crossing from ∂H_r to $\partial H_{r''}$. Thus, Menger's theorem (see [8]) implies that when (a)–(e) all hold there are j disjoint crossings from ∂H_r to $\partial H_{r''}$. By the Harris-FKG inequality, the events (a)–(e) are all positively correlated. By RSW, events (c)–(e) have probabilities bounded away from zero (assuming that (a) has positive probability). The inequality corresponding to the right-hand inequality in (A.1) now follows. The corresponding left-hand inequality, as well as the cases where $r' \le 2r$ or $r'' \le 2r'$ are now proved as in the proof of Proposition A.1.

We now assume that both colors appear in X, and indicate the adaptations necessary in the proof of Proposition A.1 to generalize to the present setting. The quantity $s(r, R)$ needs to be defined slightly differently. In the modified definition of $s(r, R)$, still only interfaces between crossings of opposite colors are considered. Suppose that γ_1 and γ_2 are two adjacent interfaces from ∂H_r to ∂H_R, and that Q is the component of $H_R \setminus (H_r \cup \gamma_1 \cup \gamma_2)$ between them. Let $\mathrm{dist}(z, Z; Q)$ denote the infimal length of a path from z to Z in $\overline{Q}$. For $s > 0$ set $W(s) := \{z \in Q : \mathrm{dist}(z, \partial H_R; Q) \le s\}$. The *margin* between γ_1 and γ_2 is defined as the supremum of the set of $s > 0$ such that any path connecting γ_1 and γ_2 in $W(s)$ has length at least s. Now $s(r, R)$ is redefined as the smallest margin among any two consecutive interfaces.

Lemma A.2 is still valid with this new definition of $s(r, R)$. In fact, the only change needed in the proof is that instead of looking for a white crossing in $Y_i \setminus B(z_i, \delta R)$ from $\partial\hat{\beta}_i$ to ∂H_R, one looks for a crossing in $Y_i \setminus B(z_i', 2\delta R)$, where z_i' is the last point on the arc $\partial Y \cap \partial H_R$, directed away from α_1, that has distance at most δR from β_i. (Here, we assume that $\delta < 1/100$, say.)

We now explain how this new definition facilitates the obvious analogue of Lemma A.3. Indeed, suppose that γ_1 and γ_2 are two adjacent interfaces, there are at least $j_1 \le j$ disjoint black crossings in the sector Q of $H_R \setminus H_r$ between γ_1 and γ_2, and the margin between γ_1 and γ_2 is at least s. Let

$$Q_i := \{z \in W(s) : 2i\, s/(2\, j_1) \le \mathrm{dist}(z, \partial H_R; Q) \le (2i+1)\, s/(2\, j_1)\},$$
$$Q_i^* := \{z \in W(s) : 2i\, s/(2\, j_1) \le \mathrm{dist}(z, \gamma_1; Q) \le (2i+1)\, s/(2\, j_1)\},$$

where $W(s)$ is as above. We may then consider the event that in each Q_i, $i = 0, 1, \ldots, j_1 - 1$, there is a black crossing from γ_1 to γ_2, and in each Q_i^*, $i = 0, 1, \ldots, j_1 - 1$, there is a black crossing from ∂H_R to $\partial W(s) \setminus (\gamma_1 \cup \gamma_2 \cup \partial H_R)$,

and moreover, the latter crossings continue through well directed channels all the way to ∂H_{4R}, as in the proof of Lemma A.3. An application of Menger's theorem, as in the monochromatic case above, will then guarantee that at the end j_1 disjoint black crossings between γ_1 and γ_2 will extend all the way to H_{4R}. A compatible construction is applied to each of the other pairs of adjacent interfaces.

Of course, when we condition on the interfaces $\gamma_0, \gamma_1, \ldots, \gamma_{k-1}$, we do not know how many crossings we will have between each pair of adjacent interfaces. But $k = O(R/s(r, R))$, and so the number of distinct patterns in which crossings with color sequence type X can occur is bounded by a function of δ and j. (Specifically, a pattern for X is a specification of how many crossings are selected between each pair of adjacent interfaces to make up the sequence of crossings compatible with X. There may very well be additional crossings that are ignored.) Thus, the most likely pattern given the interfaces occurs with probability bounded away from zero given that there are crossings of color sequence X and the construction may be based on this most likely pattern. The occurrence of this pattern given the interfaces and the information about the color of hexagons adjacent to the right-hand sides of the interfaces will be a monotone function in the collection of white hexagons in the regions between interfaces that have white hexagons on their boundary, and monotone in black hexagons in the other regions. Thus, again, the Harris-FKG inequality is applicable. (We do not want to condition on the exact number of crossings between two adjacent interfaces, as this is not a monotone function of the configuration.)

Similarly, when we attempt to glue crossings between two different annuli, we condition on the interfaces, and then aim for the most likely pattern in each annulus. These are essentially the only modifications needed in the proof. □

Remark A.6. The analogous statements for critical percolation in cones and wedges also holds, with similar proofs. The wedge case is, in fact, easier.

Remark A.7. It is also clear that the above proof shows that if we prescribe j specific disjoint arcs on ∂H_R and require the crossings from ∂H_r to land on these arcs, with a prescribed color for every arc, the probability for this event is at least a positive constant times the probability to have j crossings with this sequential color pattern, where the constant only depends on the smallest angle at 0 subtended by any of the j arcs (provided that R is not too small, so that each of the arcs has at least one hexagon unshared with any other arc, say).

As a further application, we prove the following result about 5-arm and 6-arm exponents in $\mathbb{Z}^2$.

COROLLARY A.8. *Consider critical bond percolation on* $\mathbb{Z}^2$. *For* $R > r \geq 1$ *let* $F(r, R)$ *denote the event that there are five open crossings between distances* r

and R from 0 of types primal, primal, dual, primal, dual, in circular order. Then

$$C^{-1}(r/R)^2 \le \mathbf{P}\big[F(r,R)\big] \le C\,(r/R)^2, \tag{A.5}$$

where $C > 0$ is a universal constant. Moreover, the probability that there are six alternating crossings: primal, dual, primal, dual, primal, dual, between distances r and R is at most $C\,(r/R)^{2+\varepsilon}$, for some constant $\varepsilon > 0$. The same statement applies to any sequence obtained by inserting one additional primal or dual entry to the list (primal, primal, dual, primal, dual).

This result is essentially due to [20, Lemma 5] (in the context of site percolation on the triangular grid, though the proof is equally applicable to $\mathbb{Z}^2$). They omit some of the details, because the proof is long and the argument is similar to the proof of [19, Lemma 4]. Now, we can easily present an essentially complete and relatively short argument.

Proof. We begin with the basic argument from [20]. Divide the boundary of the circle $\partial B(0,R)$ into 5 equal arcs, $A_1,\dots,A_5$, in counterclockwise order. For concreteness, let's take each A_j as the arc between angles $(2j-1)\pi/5$ and $(2j+1)\pi/5$. For a vertex $v \in B(0,R)$, let F_v be the event that there are primal (open) crossings from v to A_1, A_3 and A_4 and dual crossings from dual vertices adjacent to v to A_2 and to A_5, and the primal crossings are disjoint, except at v. (By planarity, it follows that the dual crossings are disjoint.) We claim that F_v can happen for at most one vertex in $B(0,R/2)$. Indeed, suppose that $F_v \cap F_u$ holds, where v,u are vertices in $B(0,R/2)$. Let α_i, $i = 1,3,4$, denote some simple primal crossings from v to the arcs A_i, which are disjoint, except at v. Similarly, let α_i', $i = 1,3,4$, be the corresponding paths for u. Since $\alpha_1 \cup \alpha_3$ separates A_2 from A_5 in $B(0,R)$, it is clear that $u \in \alpha_1 \cup \alpha_3$. Similarly, $u \in \alpha_1 \cup \alpha_4$. Since $\alpha_3 \cap \alpha_4 = \{v\} \subset \alpha_1$, it follows that $u \in \alpha_1$. Let β_1 be the arc of α_1 from u to A_1. Since $\beta_1 \cup \alpha_3'$ is a path from A_1 to A_3, it contains v or separates v from A_2 or from A_5. The latter two possibilities are ruled out by the dual crossings to A_2 and A_5 starting at dual vertices adjacent to v. Thus, $v \in \beta_1 \cup \alpha_3'$, and similarly, $v \in \beta_1 \cup \alpha_4'$. But since $\alpha_3' \cap \alpha_4' = \{u\} \subset \beta_1$, we conclude that $v \in \beta_1$, which implies $u = v$.

We now claim that $F := \bigcup_{v \in B(0,R/2)} F_v$ has probability bounded away from 0. Consider the event that there is a crossing from A_1 to A_4, and consider the rightmost such crossing ℓ (in the sense that it separates any other crossing from A_5). If there is an open path from A_3 to ℓ, but no open path from A_3 to A_1 disjoint from ℓ, then F_w will hold, where w is the first vertex v along ℓ (when ℓ goes from A_1 to A_4) that connects to A_3 in the complement of ℓ. Thus, we need to show that there is probability bounded away from zero that this happens with $w \in B(0,R/2)$. Let L_1 be the line passing through the origin and the midpoint of A_1. Let L_2 and L_3 be lines parallel to L_1 at distance $R/20$ and $R/10$ from L_1, on the side of L_1

that contains A_5. By RSW, there is probability bounded away from zero for the existence of a dual-open crossing from A_1 to A_4 in the strip between L_2 and L_3. By conditioning on the leftmost such crossing (the one closest to L_1), it is easy to see that there is probability bounded away from zero that such a dual crossing exists and is also connected to A_5 by a dual-open path. If moreover we have a primal crossing from A_1 to A_4 in the strip between L_1 and L_2, which happens with probability bounded away from zero, then the rightmost primal crossing between A_1 and A_4 will be contained in the strip between L_1 and L_3. Conditioned on the latter event and on the latter crossing ℓ, it happens with probability bounded away from zero that there is a primal crossing from A_3 to ℓ whose endpoint on ℓ (which is its only point on ℓ) is within distance $R/5$ of the origin. Also, there is a dual crossing from A_2 to a dual vertex adjacent to ℓ that is within distance $R/5$ of the origin. On that event, F holds. Thus, $\mathbf{P}[F]$ is bounded away from zero.

It is easy to see that the proof of Proposition A.5 implies that $\mathbf{P}[F_v] \leq O(1)\,\mathbf{P}[F_w]$ for $v, w \in B(0, R/2)$. Since the events F_v are disjoint, and since their sum is of order 1, it follows that each F_v, $v \in B(0, R/2)$ has probability of order R^{-2}. In particular $R^2\,\mathbf{P}[F_0]$ is bounded away from zero and ∞. Now (A.5) follows from Remark A.7.

The statements regarding the 6-arm exponent now follow from Reimer's inequality [27]. Alternatively, we may also deduce them from Remark A.7, as follows. If we fix arcs $A_1, \ldots, A_6$ in counterclockwise order on the radius R circle, where the crossings are required to land, and we require a primal crossing to A_1 and a dual crossing to A_2, then we may condition on the most counterclockwise primal crossing γ connecting to A_1, then sequentially on the most clockwise crossings to $A_2, A_3, \ldots, A_5$ of the required type. The conditional probability for having yet another crossing to A_6 is still bounded by $O(1)\,(r/R)^{\varepsilon}$, for some constant $\varepsilon > 0$. Now we may apply Remark A.7, to complete the proof. □

Acknowledgments. We thank Harry Kesten for useful advice.

References

[1] M. AIZENMAN, S. DUPLANTIER, and A. AHARONY, Connectivity exponents and external perimeter in 2D percolation models, *Phys. Rev. Lett.* **83** (1999), 1359–1362.

[2] I. BENJAMINI, G. KALAI, and O. SCHRAMM, Noise sensitivity of Boolean functions and applications to percolation, *Inst. Hautes Études Sci. Publ. Math.* (1999), 5–43 (2001). MR 2001m:60016 Zbl 0986.60002

[3] I. BENJAMINI and O. SCHRAMM, Exceptional planes of percolation, *Probab. Theory Related Fields* **111** (1998), 551–564. MR 99i:60173 Zbl 0910.60076

[4] I. BENJAMINI, O. SCHRAMM, and D. B. WILSON, Balanced Boolean functions that can be evaluated so that every input bit is unlikely to be read, in STOC'05: *Proc. 37th Annual* ACM *Symposium on Theory of Computing*, ACM, New York, 2005, pp. 244–250. MR 2006g:68280

[5] J. BERG, R. MEESTER, and D. G. WHITE, Dynamic Boolean models, *Stochastic Process. Appl.* **69** (1997), 247–257. MR 98j:60134 Zbl 0911.60083

[6] E. I. BROMAN and J. E. STEIF, Dynamical stability of percolation for some interacting particle systems and ϵ-movability, *Ann. Probab.* **34** (2006), 539–576. MR 2007b:60233 Zbl 1107.82058

[7] F. CAMIA and C. M. NEWMAN, Two-dimensional critical percolation: the full scaling limit, *Comm. Math. Phys.* **268** (2006), 1–38. MR 2007m:82032 Zbl 1117.60086

[8] R. DIESTEL, *Graph Theory*, *Grad. Texts Math.* **173**, Springer-Verlag, New York, 1997. MR 1448665 Zbl 1086.05001

[9] S. N. EVANS, Local properties of Lévy processes on a totally disconnected group, *J. Theoret. Probab.* **2** (1989), 209–259. MR 90g:60069 Zbl 0683.60010

[10] P. J. FITZSIMMONS and R. K. GETOOR, On the potential theory of symmetric Markov processes, *Math. Ann.* **281** (1988), 495–512. MR 89k:60110 Zbl 0627.60067

[11] R. K. GETOOR and M. J. SHARPE, Naturality, standardness, and weak duality for Markov processes, *Z. Wahrsch. Verw. Gebiete* **67** (1984), 1–62. MR 86f:60093 Zbl 0553.60070

[12] G. GRIMMETT, *Percolation*, second ed., *Grundl. Math. Wissen.* **321**, Springer-Verlag, New York, 1999. MR 2001a:60114

[13] O. HÄGGSTRÖM, Y. PERES, and J. E. STEIF, Dynamical percolation, *Ann. Inst. H. Poincaré Probab. Statist.* **33** (1997), 497–528. MR 98m:60153

[14] T. HARA and G. SLADE, Mean-field behaviour and the lace expansion, in *Probability and Phase Transition* (*Cambridge*, 1993), *Sci.* **420**, Kluwer Acad. Publ., Dordrecht, 1994, pp. 87–122. MR 95d:82033 Zbl 0831.60107

[15] T. E. HARRIS, A lower bound for the critical probability in a certain percolation process, *Proc. Cambridge Philos. Soc.* **56** (1960), 13–20. MR 22 #6023 Zbl 0122.36403

[16] J. HAWKES, Some geometric aspects of potential theory, in *Stochastic Analysis and Applications* (*Swansea*, 1983), *Lecture Notes in Math.* **1095**, Springer-Verlag, New York, 1984, pp. 130–154. MR 86h:60146 Zbl 0558.60055

[17] J.-P. KAHANE, *Some Random Series of Functions*, second ed., *Cambridge Studies Adv. Mat.* **5**, Cambridge Univ. Press, Cambridge, 1985. MR 87m:60119 Zbl 0571.60002

[18] H. KESTEN, *Percolation Theory for Mathematicians*, *Progr. Prob. Statist.* **2**, Birkhäuser, Mass., 1982. MR 84i:60145 Zbl 0522.60097

[19] ———, Scaling relations for 2D-percolation, *Comm. Math. Phys.* **109** (1987), 109–156. MR 88k:60174

[20] H. KESTEN, V. SIDORAVICIUS, and Y. ZHANG, Almost all words are seen in critical site percolation on the triangular lattice, *Electron. J. Probab.* **3** (1998), no. 10, 75 pp. MR 99j:60155 Zbl 0908.60082

[21] H. KESTEN and Y. ZHANG, Strict inequalities for some critical exponents in two-dimensional percolation, *J. Statist. Phys.* **46** (1987), 1031–1055. MR 89g:60305 Zbl 0683.60081

[22] G. F. LAWLER, *Conformally Invariant Processes in the Plane*, *Math. Surveys Monogr.* **114**, Amer. Math. Soc., Providence, RI, 2005. MR 2006i:60003 Zbl 1074.60002

[23] G. F. LAWLER, O. SCHRAMM, and W. WERNER, One-arm exponent for critical 2D percolation, *Electron. J. Probab.* **7** (2002), 13 pp. MR 2002k:60204 Zbl 1015.60091

[24] ———, Sharp estimates for Brownian non-intersection probabilities, in *In and Out of Equilibrium* (*Mambucaba*, 2000), *Progr. Probab.* **51**, Birkhäuser, Boston, MA, 2002, pp. 113–131. MR 2003d:60162 Zbl 1011.60062

[25] Y. PERES, O. SCHRAMM, S. SHEFFIELD, and D. B. WILSON, Random-turn hex and other selection games, *Amer. Math. Monthly* **114** (2007), 373–387. MR 2008a:91039 Zbl 1153.91012

[26] Y. PERES and J. E. STEIF, The number of infinite clusters in dynamical percolation, *Probab. Theory Related Fields* **111** (1998), 141–165. MR 99e:60217 Zbl 0906.60069

[27] D. REIMER, Proof of the van den Berg-Kesten conjecture, *Combin. Probab. Comput.* **9** (2000), 27–32. MR 2001g:60017 Zbl 0947.60093

[28] O. SCHRAMM, Scaling limits of loop-erased random walks and uniform spanning trees, *Israel J. Math.* **118** (2000), 221–288. MR 2001m:60227 Zbl 0968.60093

[29] S. SMIRNOV, Critical percolation in the plane. I. Conformal invariance and Cardy's formula, II. Continuum scaling limit (long version), preprint. arXiv 0909.4499

[30] S. SMIRNOV, Critical percolation in the plane: conformal invariance, Cardy's formula, scaling limits, *C. R. Acad. Sci. Paris Sér. I Math.* **333** (2001), 239–244. MR 2002f:60193 Zbl 0985.60090

[31] S. SMIRNOV and W. WERNER, Critical exponents for two-dimensional percolation, *Math. Res. Lett.* **8** (2001), 729–744. MR 2003i:60173 Zbl 1009.60087

[32] M. TALAGRAND, Concentration of measure and isoperimetric inequalities in product spaces, *Inst. Hautes Études Sci. Publ. Math.* (1995), 73–205. MR 97h:60016 Zbl 0864.60013

(Received May 4, 2005)

E-mail address: schramm@microsoft.com
MICROSOFT CORPORATION, ONE MICROSOFT WAY, REDMOND WA 98052, UNITED STATES,
http://research.microsoft.com/en-us/um/people/schramm/

E-mail address: steif@math.chalmers.se
DEPARTMENT OF MATHEMATICS, CHALMERS UNIVERSITY OF TECHNOLOGY AND UNIVERSITY OF GOTHENBURG, GOTHENBURG, SWEDEN

http://www.math.chalmers.se/~steif/

Acta Math., 205 (2010), 19–104
DOI: 10.1007/s11511-010-0051-x
© 2010 by Institut Mittag-Leffler. All rights reserved

The Fourier spectrum of critical percolation

by

Christophe Garban
École normale supérieure de Lyon
Lyon, France

Gábor Pete
University of Toronto
Toronto, ON, Canada

Oded Schramm*
Microsoft Research
Redmond, WA, U.S.A.

Contents

1. Introduction 20
   1.1. Some general background 20
   1.2. The main result 22
   1.3. Applications to noise sensitivity 24
   1.4. Applications to dynamical percolation 26
   1.5. The scaling limit of the spectral sample 28
   1.6. A rough outline of the proof 28
2. Some basics 32
   2.1. A few general definitions 32
   2.2. Multi-arm events for percolation 33
   2.3. The spectral sample in general 35
3. First percolation spectrum estimates 38
4. The probability of a very small spectral sample 41
   4.1. The statement 41
   4.2. A local result 42
   4.3. Handling boundary issues 52
   4.4. The radial case 56
5. Partial independence in the spectral sample 60
   5.1. Setup and main statement 60
   5.2. Bounding the second moment 62
   5.3. Reformulation of the first-moment estimate 63
   5.4. Quasi-multiplicativity for coupled configurations 65

CG was partially supported by the ANR under the grant ANR-06-BLAN-0058. GP was partially supported by the Hungarian National Foundation for Scientific Research, grant T049398, and by an NSERC Discovery Grant. For large parts of the work, all three authors were at Microsoft Research.

* Deceased on September 1st, 2008.

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_14, © Springer Science+Business Media, LLC 2011

5.5. Proof of the first-moment estimate 71
5.6. A local result for general quads 73
5.7. The radial case 74
6. A large deviation result 76
7. The lower tail of the spectrum 77
7.1. Local version 77
7.2. Square version 79
7.3. Radial version 81
7.4. Tightness of the quad spectral sample 82
8. Applications to noise sensitivity 83
8.1. Noise sensitivity in a square and a quad 83
8.2. Resampling a fixed set of bits 86
9. Applications to dynamical percolation 91
10. Scaling limit of the spectral sample 93
11. Some open problems 97
Appendix A. An inequality for multi-arm probabilities 98
References 102

1. Introduction

1.1. Some general background

The Fourier expansion of functions on $\mathbb{R}^d$ is an indispensable tool with numerous applications. Likewise, the harmonic analysis of functions defined on the discrete cube $\{-1,1\}^d$ has found a host of applications; see the survey [KS]. Yet the Fourier expansion of some functions of interest is rather poorly understood. Here, we study the harmonic analysis of functions arising from planar percolation and answer most if not all of the previously posed problems regarding their Fourier expansions. We also derive some applications to the behavior of percolation under noise and in the study of dynamical percolation. It is hoped that some of the techniques introduced here will be helpful in the harmonic analysis of other functions.

Let $\mathcal{I}$ be some finite set, and let $\Omega=\{-1,1\}^{\mathcal{I}}$ be endowed with the uniform measure. The Fourier basis on Ω consists of all the functions of the form

$$\chi_S(\omega) := \prod_{i\in S} \omega_i,$$

where $S\subseteq\mathcal{I}$. (These functions are also sometimes called the *Walsh functions*.) It is easily seen to be an orthonormal basis with respect to the inner product $\mathbb{E}[fg]$. Therefore, for every $f:\Omega\to\mathbb{R}$, we have

$$f = \sum_{S\subseteq\mathcal{I}} \hat{f}(S)\chi_S, \tag{1.1}$$

where $\hat{f}(S):=\mathbb{E}[f\chi_S]$. If $\mathbb{E}[f^2]=1$, then the random variable $\mathscr{S}=\mathscr{S}_f\subseteq\mathcal{I}$ with distribution given by

$$\mathbb{P}[\mathscr{S}=S]=\hat{f}(S)^2$$

will be called the *Fourier spectral sample* of f. Due to Parseval's formula, this is indeed a probability distribution. The idea to look at this as a probability distribution was proposed in [BKS], though the study of the weights $\hat{f}(S)^2$ is "ancient", and boils down to the same questions in a different language. As noted there, important properties of the function f are encoded in the law of the spectral sample. For example, suppose that $f:\{-1,1\}^{\mathcal{I}}\to\{-1,1\}$. Let $x\in\{-1,1\}^{\mathcal{I}}$ be random and uniform, and let y be obtained from x by resampling([1]) each coordinate with probability ε independently, with $\varepsilon\in(0,1)$. Then y is referred to as an *ε-noise* of x. Since for $i\in\mathcal{I}$ we have $\mathbb{E}[x_iy_i]=1-\varepsilon$, it follows that $\mathbb{E}[\chi_S(x)\chi_S(y)]=(1-\varepsilon)^{|S|}$ for $S\subseteq\mathcal{I}$ and hence it easily follows by using the Fourier expansion (1.1) that

$$\mathbb{E}[f(x)f(y)]=\mathbb{E}[(1-\varepsilon)^{|\mathscr{S}_f|}]. \tag{1.2}$$

Thus, the stability or sensitivity of f to noise is encoded in the law of $|\mathscr{S}|$.

One mathematical model in which noise comes up is that of the Poisson dynamics on Ω, in which each coordinate is resampled according to a Poisson process of rate 1, independently. This is, of course, just the continuous time random walk on Ω. If x_t denotes this continuous time Markov process started at the stationary (uniform) measure on Ω, then x_t is just ε-noise of x_0, where $\varepsilon=1-e^{-t}$. Indeed, the Markov operator defined by

$$T_tf(x)=\mathbb{E}[f(x_t)\,|\,x_0=x]$$

is diagonalized by the Fourier basis:

$$T_t\chi_S=e^{-t|S|}\chi_S.$$

It is therefore hardly surprising that the behavior of $|\mathscr{S}_f|$ will play an important role in the study of the generally non-Markov process $f(x_t)$. These types of questions have been under investigation in the context of Bernoulli percolation [HPSt], [PS], [HP], [PSS], [Kh], other percolation type processes [BMW], [BS], [BrS], and also more generally [BHPS], [JS], [H], [KLM]. Estimates of the Fourier coefficients played an important role in the proof that the dynamical version of critical site percolation on the triangular grid a.s. has percolation times [SSt]. These estimates can naturally be phrased in terms of properties of the random variable $|\mathscr{S}|$.

([1]) In the definition of [BKS], each bit is flipped with probability ε, rather than resampled. This accounts for some discrepancies involving factors of 2. The present formulation generally produces simpler formulas.

Recall that the (random) set of *pivotals* of $f:\{-1,1\}^{\mathcal{I}}\to\{-1,1\}$ is the set of $i\in\mathcal{I}$ such that flipping the value of ω_i also changes the value of $f(\omega)$. It is easy to see [KKL] that the first moment of the number of pivotals of f is the same as the first moment of $|\mathscr{S}_f|$. Gil Kalai (personal communication) observed that the same is true for the second moment, but not for the higher moments. (We will recall the easy proof of this fact in §2.3.) This often facilitates an easy estimation of $\mathbb{E}[|\mathscr{S}_f|]$ and $\mathbb{E}[|\mathscr{S}_f|^2]$.

It is often the case that $\mathbb{E}[|\mathscr{S}|^2]$ is of the same order of magnitude as $\mathbb{E}[|\mathscr{S}|]^2$, and this implies that with probability bounded away from zero, the random variable $|\mathscr{S}|$ is of the same order of magnitude as its mean. However, what turns out to be much harder to estimate is the probability that $|\mathscr{S}|$ is positive and much smaller than its mean. (In particular, this is much harder than the analogous tightness result for pivotals, see Remark 1.7 below.) This is very relevant to applications; as can be seen from (1.2), the probability that $|\mathscr{S}|$ is small is what matters most in understanding the correlation of $f(x)$ with f evaluated on a noisy version of x. Likewise, in the dynamical setting, the lower tail of $|\mathscr{S}|$ controls the switching rate of f. Indeed, the primary purpose of this paper is getting good estimates on $\mathbb{P}[0<|\mathscr{S}_f|<s]$ for indicators of crossing events in percolation and deriving the consequences of such bounds.

1.2. The main result

We consider two percolation models: critical bond percolation on the square grid $\mathbb{Z}^2$ and critical site percolation on the triangular grid. See [G] and [W] for background. These two models are believed to behave essentially the same, but the mathematical understanding of the latter is significantly superior to the former due to Smirnov's theorem [Sm1] and its consequences. Fix some large $R>0$, and consider the event that (in either of these percolation models) there is an open (i.e., occupied) left-right crossing of the square $[0,R]^2$. Let f_R denote the ±1-indicator function of this event; that is, $f_R=1$ if there is a crossing and $f_R=-1$ when there is no crossing. In this case, $\mathcal{I}$ is the relevant set of bonds or sites, depending on whether we are in the bond or site model. The probability space is $\Omega=\{-1,1\}^{\mathcal{I}}$ with the uniform measure. Here, it is convenient that $p_c=\frac{1}{2}$ for both models, and so the relevant measure on Ω is uniform.

The paper [BKS] posed the problem of studying the law of $\mathscr{S}_{f_R}$. There, it was proved that $\mathbb{P}[0<|\mathscr{S}_{f_R}|<c\log R]\to0$, as $R\to\infty$, for some $c>0$. This had the implication that f_R is asymptotically noise sensitive; that is

$$\lim_{R\to\infty}(\mathbb{E}[f_R(x)f_R(y)]-\mathbb{E}[f_R(x)]\mathbb{E}[f_R(y)])=0,$$

when y is an ε-noisy version of x and $\varepsilon>0$ is fixed. It was also asked in [BKS] whether $\mathbb{P}[0<|\mathscr{S}|<R^\delta]\to0$ for some $\delta>0$. This was later proved in [SSt] with $\delta=\frac{1}{8}+o(1)$ in the

setting of the triangular lattice and with an unspecified $\delta>0$ for the square lattice. While certainly a useful step forward, these results were far from being sharp. But the issue is more than just a quantitative question. The most natural hypothesis is that $|\mathscr{S}|$ is always proportional to its mean when it is non-zero, or more precisely that

$$\sup_{R>1} \mathbb{P}[0<|\mathscr{S}_{f_R}|<t\mathbb{E}|\mathscr{S}_{f_R}|] \to 0 \quad \text{as } t \searrow 0. \tag{1.3}$$

To illustrate the fact that (1.3) is not a universal principle, we note that it does not hold, e.g., for the ±1-indicator function of the event that there is a left-right percolation in the square $[0,R]^2$ and this square has more open sites (or edges) than closed sites (or edges). We prove (1.3) in the present paper, and give useful bounds that are sharp up to constants on the left-hand side of (1.3). This is the content of our first theorem.

THEOREM 1.1. *As above, let $R>1$ and let f_R be the ±1-indicator function of the left-right crossing event of the square $[0,R]^2$ in critical bond percolation on $\mathbb{Z}^2$ or site percolation on the triangular grid. The spectral sample of f_R satisfies*

$$\mathbb{P}[0<|\mathscr{S}_{f_R}|<\mathbb{E}|\mathscr{S}_{f_r}|] \asymp \left(\frac{\mathbb{E}|\mathscr{S}_{f_R}|/R}{\mathbb{E}|\mathscr{S}_{f_r}|/r}\right)^2 \tag{1.4}$$

for every $r\in[1,R]$, and $\asymp$ denotes equivalence up to positive multiplicative constants.

To make the sharp bound (1.4) more explicit, one needs to discuss estimates for $\mathbb{E}|\mathscr{S}_{f_r}|$. The value of $\mathbb{E}|\mathscr{S}_{f_r}|$ is estimated by the probability of the so-called "alternating 4-arm event", which we will treat in detail in §2.2. For now, let us just mention that it is known that

$$\frac{\mathbb{E}|\mathscr{S}_{f_R}|}{\mathbb{E}|\mathscr{S}_{f_r}|} \leqslant C\left(\frac{R}{r}\right)^{1-\delta} \tag{1.5}$$

for any $R\geqslant r\geqslant1$ with some $\delta>0$ and $C<\infty$, and that, for the triangular lattice,

$$\frac{\mathbb{E}|\mathscr{S}_{f_R}|}{\mathbb{E}|\mathscr{S}_{f_r}|} \asymp \left(\frac{R}{r}\right)^{3/4+o(1)} \tag{1.6}$$

as $R/r\to\infty$ while $r\geqslant1$, which follows from Smirnov's theorem [Sm1] and the SLE-based analysis of the percolation exponents in [SW]. (This will be proved in §7.2.) From (1.6) and (1.4), we get for the triangular grid that

$$\mathbb{P}[0<|\mathscr{S}_{f_R}|\leqslant\lambda\mathbb{E}|\mathscr{S}_{f_R}|] \asymp \lambda^{2/3+o(1)}, \tag{1.7}$$

where λ may depend on R, but is restricted to the range $[(\mathbb{E}|\mathscr{S}_{f_R}|)^{-1},1]$. Here, the $o(1)$ represents a function of λ and R that tends to 0 as $\lambda\to0$, uniformly in R. This answers

Problem 5.1 from [Sch]. Even for the square lattice, (1.3) follows from Theorem 1.1 combined with (1.5). Below, we prove (7.6), which is a variant of (1.7) with slightly different asymptotics.

There is nothing particularly special about the square with regard to Theorem 1.1. The proof applies to every rectangle of a fixed shape (with the implied constants depending on the shape). For percolation crossings in more general shapes, Theorem 7.1 gives bounds on the behavior of $\mathscr{S}$ away from the boundary, and we also prove Theorem 7.4, which is some analog of (1.3). However, we chose not to go into the complications that would arise when trying to prove (1.4) in this general context. The issue is that $\mathscr{S}$ could possibly be of larger size near "peaks" of the boundary of $\mathcal{Q}$ (especially if it is fractal-like), so the boundary contributions might dominate $\mathbb{E}|\mathscr{S}|$ and might also have a complicated influence on the tail behavior. Still, as it turns out, $\mathscr{S}$ is unlikely to be close to the boundary, so we can ignore these effects to a certain extent and prove Theorem 7.4.

We also remark that $\mathbb{P}[\mathscr{S}_f=\varnothing]=\mathbb{E}[f]^2=\hat{f}(\varnothing)^2$ is generally easy to compute. The level-0 Fourier coefficient $\hat{f}(\varnothing)$ has more to do with the way the function is normalized than with its fundamental properties. For this reason, $|\mathscr{S}|=0$ is separated out in bounds such as (1.4).

At this point we mention that Theorem 7.3 gives the bound analogous to Theorem 1.1, but dealing with the spectrum of the indicator function for a percolation crossing from the origin to distance R away.

1.3. Applications to noise sensitivity

Figure 1.1 illustrates a sequence of percolation configurations, where each configuration is obtained from the previous one by applying some noise. The effect on the interfaces can be observed. Theorem 1.1 (or, in fact, its corollary (1.3)) implies the following sharp noise sensitivity estimate regarding such perturbations.

COROLLARY 1.2. *Suppose that y is an ε_R-noisy version of x, where $\varepsilon_R\in(0,1)$ may depend on R. If $\lim_{R\to\infty}\mathbb{E}|\mathscr{S}_{f_R}|\varepsilon_R=\infty$, then*

$$\lim_{R\to\infty}(\mathbb{E}[f_R(x)f_R(y)]-\mathbb{E}[f_R(x)]\mathbb{E}[f_R(y)])=0, \tag{1.8}$$

while if $\lim_{R\to\infty}\mathbb{E}|\mathscr{S}_{f_R}|\varepsilon_R=0$, then

$$\lim_{R\to\infty}(\mathbb{E}[f_R(x)f_R(y)]-\mathbb{E}[f_R(x)^2])=0. \tag{1.9}$$

The second of these statements is actually obvious from (1.2) and Jensen's inequality, and is brought here only to complement the first claim. Of course, (1.8) just

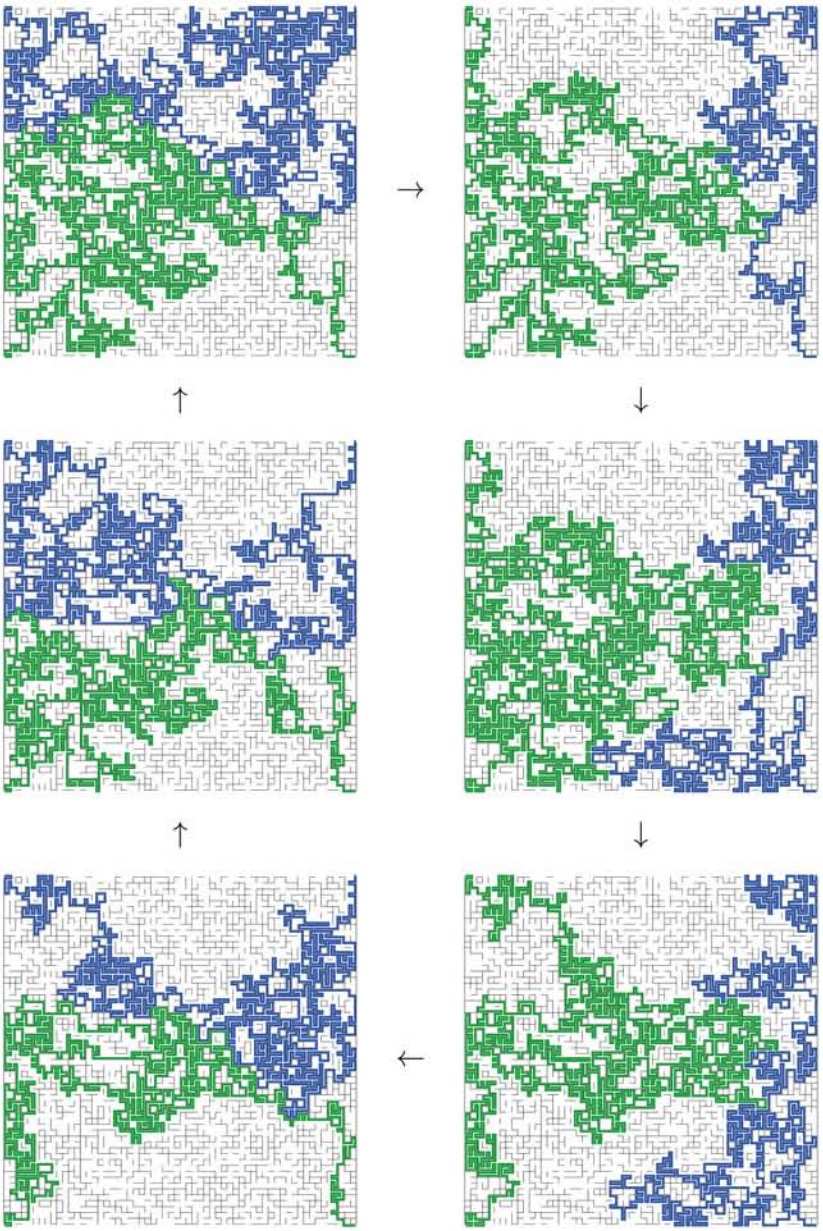


Figure 1.1. Interfaces in percolation on $\mathbb{Z}^2$. Each successive pair of configurations are related by a noise of about 0.04, which results in about one in every 50 bits being different. The squares are of size about 60×60. The perturbations follow each other cyclically, flipping the edges in three independently chosen subsets, S_1, S_2, S_3, S_1, S_2, S_3, arriving back where we started.

means that having a crossing in x is asymptotically uncorrelated with having a crossing in y, while (1.9) means that with probability going to 1, a crossing in x occurs if and only if there is a crossing in y. Although $f_R(x)^2=1$, we find the form of (1.9) more suggestive, since this is the statement that generalizes to other situations. Likewise, $\lim_{R\to\infty}\mathbb{E}[f_R(x)]=0$, but (1.8) is more suggestive.

In Corollary 8.1, we prove a generalization of Corollary 1.2 for the crossing function $f_{R\mathcal{Q}}$, where $\mathcal{Q}$ is an arbitrary fixed "quad", i.e., a domain homeomorphic to a disk with four marked points on its boundary.

In a forthcoming paper on the scaling limit of dynamical percolation [GPS2] (see also [GPS1]), we plan to show that for critical percolation on the triangular grid, whenever $\lim_{R\to\infty}\mathbb{E}|\mathscr{S}_{f_R}|\varepsilon_R$ exists and is in $(0,\infty)$, then $\lim_{R\to\infty}\mathbb{E}[f_R(x)f_R(y)]$ also exists and is strictly between the limits of $\mathbb{E}[f_R(x)]^2$ and $\mathbb{E}[f_R(x)^2]$.

The following theorem proves Conjecture 5.1 from [BKS]. With minor adaptations, it follows from the proof of Theorem 1.1.

THEOREM 1.3. *Consider bond percolation on $\mathbb{Z}^2$. Let x be a critical percolation configuration, and let z be another critical percolation configuration, which equals x on the horizontal edges, but is independent of x on the vertical edges. Then having a left-right crossing in x is asymptotically independent from having a left-right crossing in z. Moreover, the same holds true if "horizontal" and "vertical" are interchanged (but still considering left-to-right crossings).*

This is just a special case of Proposition 8.2, a quite general result on "sensitivity to selective noise".

1.4. Applications to dynamical percolation

First, recall that in dynamical percolation the random bits determining the percolation configuration are refreshed according to independent Poisson clocks of rate 1. Dynamical percolation was proposed by Itai Benjamini in 1992, and later independently by Häggström, Peres and Steif, who wrote the first paper on the subject [HPSt]. Since then, dynamical percolation and other dynamical random processes have been the focus of several research papers; see the references in §1.1. In response to a question in [HPSt], it was proved in [SSt] that critical dynamical site percolation on the triangular grid a.s. has times at which the origin is in an infinite connected percolation component. Such times are called "exceptional", since they necessarily have measure zero. The paper [SSt] also showed that the dimension of the set of exceptional times is a.s. in $\left[\frac{1}{6},\frac{31}{36}\right]$, and conjectured that it is a.s. $\frac{31}{36}$. Likewise, it was conjectured there that the set of times at

which an infinite occupied as well as an infinite vacant cluster coexist is a.s. $\frac{2}{3}$, but [SSt] only proved that $\frac{2}{3}$ is an upper bound, without establishing the existence of such times. Similarly, [SSt] proved that the set of times at which there is an infinite percolation component in the upper half-plane has Hausdorff dimension at most $\frac{5}{9}$, but did not prove that the set is non-empty. There were numerous other lower and upper bounds of this type in [SSt], some of them having to do with dynamical percolation in wedges and cones, which will not be discussed in this paper. We can now prove most of these conjectures regarding the Hausdorff dimensions of exceptional times in the setting of the triangular grid, and for each case corresponding to a monotone event, the Hausdorff dimension a.s. equals the previously known upper bound. In particular, we have the following result.

THEOREM 1.4. *In the setting of dynamical critical site percolation on the triangular grid, we have the following a.s. values for the Hausdorff dimensions.*

(1) *The set of times at which there is an infinite cluster a.s. has Hausdorff dimension* $\frac{31}{36}$.

(2) *The set of times at which there is an infinite cluster in the upper half-plane a.s. has Hausdorff dimension* $\frac{5}{9}$.

(3) *The set of times at which an infinite occupied cluster and an infinite vacant cluster coexist a.s. has Hausdorff dimension at least* $\frac{1}{9}$.

The reason that in (1) and (2) the results agree with the conjectured upper bound from [SSt] is that the upper bound is dictated by $\mathbb{E}[|\mathscr{S}_f|]$ (which is generally not hard to compute), while the tail estimate given in (1.4) and its analogs give sufficient estimates to bound the probability that $|\mathscr{S}_f|$ is much smaller than its expectation. Here, f is the indicator function of some crossing event, which may vary from one application to another. We cannot calculate the exact dimension in statement (3) because we use the monotonicity of f in an essential way (though at only one point), and the event that both vacant and occupied percolation crossings occur is not monotone. The lower bound $\frac{1}{9}$ comes from using the $\frac{5}{9}$ result in the upper and lower half-planes separately, in which case "codimensions add" by independence.

See §9 for further results and an explanation as to how these numbers are calculated.

The paper [SSt] came quite close to proving that exceptional times exist for dynamical critical bond percolation on $\mathbb{Z}^2$, but was not able to do it. Now, we close this gap.

THEOREM 1.5. *A.s. there are exceptional times at which dynamical critical bond percolation on* $\mathbb{Z}^2$ *has infinite clusters, and the Hausdorff dimension of the set of such times is a.s. positive.*

A paper in preparation, [HPS], defines a natural local time measure on the set of exceptional times for percolation, and proves that the configuration at a "typical

exceptional time", chosen with respect to this local time, has the distribution of Kesten's incipient infinite cluster [K1].

There is one more application to dynamical percolation that we will presently mention. This has to do with the scaling limit of dynamical percolation, as introduced in [Sch], and whose existence we plan to show in [GPS2] (see also [GPS1]). In this scaling limit, time and space are both scaled, and the relationship between their scaling is chosen in such a way that the event of the existence of a percolation crossing of the unit square at time 0 and at one unit of time later have some fixed correlation strictly between 0 and 1. Consequently, as space is shrinking, time is expanding. We leave it as an exercise to the reader to verify that the ratio between the scaling of time and of space can be worked out directly from the law of $|\mathcal{S}_{f_R}|$. An easy consequence of (1.3) is that in the dynamical percolation scaling limit, the correlation between having a left-right crossing of the square at time 0 and at time t goes to zero as $t\to\infty$; see (8.7). In fact, based on [SS] and estimates such as (1.3) and its generalizations to other domains, it can be shown that the dynamical percolation scaling limit is ergodic with respect to the time. These results answer Problem 5.3 from [Sch].

1.5. The scaling limit of the spectral sample

The study of the scaling limit of $\mathcal{S}$ was suggested by Gil Kalai [Sch, Problem 5.2] (see also [BKS, Problem 5.4]). The idea is that we can think of $\mathcal{S}_{f_R}$ as a random subset of the plane, and consider the existence of the weak limit as $R\to\infty$ of the law of $R^{-1}\mathcal{S}_{f_R}$. Boris Tsirelson [T] addressed this problem more generally within his theory of noises, dealing with various functions f that are not necessarily related to percolation. It follows from Tsirelson's theory and from [SS] that the scaling limit of $\mathcal{S}_{f_R}$ exists. In §10, we explain this, and prove the following result.

THEOREM 1.6. *In the setting of the triangular grid, the limit in law of* $R^{-1}\mathcal{S}_{f_R}$ *exists. It is a.s. a Cantor set of dimension* $\frac{3}{4}$.

The conformal invariance of the scaling limit of $\mathcal{S}_{f_R}$ in the setting of the triangular grid is also proved in §10. These results answer a problem posed by Gil Kalai [Sch, Problem 5.2].

1.6. A rough outline of the proof

The proof of Theorem 1.1 does not follow the same general strategy as the proof of the non-sharp bounds given in [SSt]. The lower bound for the left-hand side in (1.4) is rather easy, and so we only discuss here the proof of the upper bound. Fix some $r\in[1,R]$

and subdivide the square $[0,R]^2$ into subsquares of side length r (suppose that r divides R, say). Let $\mathcal{S}(r)$ denote the set of these subsquares that intersect $\mathcal{S}_{f_R}$. In §4 we estimate the probability that $|\mathcal{S}(r)|=k$ when k is small (for example, $k=O(\log(R/r))$). The argument is based on building a rough geometric classification of all the possible configurations of $\mathcal{S}(r)$, applying a bound for each class, and summing over the different classes. The bound obtained in this way is

$$\mathbb{P}[|\mathcal{S}(r)|=k] \leqslant \exp(O(1)\log^2(k+2)) \left(\frac{\mathbb{E}|\mathcal{S}_{f_R}|/R}{\mathbb{E}|\mathcal{S}_{f_r}|/r} \right)^2, \tag{1.10}$$

and has the optimal dependence on R and r, but a rather bad dependence on k.

Here is a naive strategy for getting from (1.10) to (1.4), which does not seem to work. Fix some $r\times r$ square B. Suppose that we are able to show that conditioned on the intersection of $\mathcal{S}$ with some set W in the complement of the r-neighborhood of B, and conditioned on $\mathcal{S}$ intersecting B, we have a probability bounded away from 0 that $|\mathcal{S}\cap B|>\mathbb{E}|\mathcal{S}_{f_r}|$. We can then restrict to a sublattice of $r\times r$ squares that are at mutual distance at least r and easily show by induction that the probability that $\mathcal{S}$ intersects at least k' of the squares in the sublattice but has size less than $\mathbb{E}|\mathcal{S}_{f_r}|$ is exponentially small in k'. We may then take a bounded set of such sublattices, which covers every one of the $r\times r$ squares in our initial tiling of $[0,R]^2$. Thus, using the exponentially small bound in k' when $|\mathcal{S}(r)|$ is large enough, while (1.10) when $|\mathcal{S}(r)|$ is small, we obtain the required bound on $\mathbb{P}[0<|\mathcal{S}|<\mathbb{E}|\mathcal{S}_{f_r}|]$. The reason that this strategy fails is that there are presently no good tools to understand the conditional law of $\mathcal{S}\cap B$ given $\mathcal{S}\cap W$. Refusing to give up, we observe that, as explained in §2.3, the law of $B\cap\mathcal{S}$ conditioned on the "negative information" $\mathcal{S}\cap W=\varnothing$ can be described. Based on this, we amend the above strategy, as follows. We pick a random set $\mathcal{Z}\subseteq\mathcal{I}$ independent of $\mathcal{S}$, where each $i\in\mathcal{I}$ is put in $\mathcal{Z}$ with probability about $1/\mathbb{E}|\mathcal{S}_{f_r}|$ independently. The reason for using this sparse random set is that for any $r\times r$ square B, either $\mathcal{S}\cap B\cap\mathcal{Z}$ is empty, and thus we gained only "negative information" about $\mathcal{S}$, or $|\mathcal{S}\cap B|$ is likely to be as large as $\mathbb{E}|\mathcal{S}_{f_r}|$. So, as we will see in a second, we can hope to get a good upper bound on $\mathbb{P}[\mathcal{S}\neq\varnothing=\mathcal{Z}\cap\mathcal{S}]$, which would almost immediately give a constant times the same bound on $\mathbb{P}[0<|\mathcal{S}|<\mathbb{E}|\mathcal{S}_{f_r}|]$.

In §5 we show that for an $r\times r$ square B and the random $\mathcal{Z}$ as above, if we condition on $\mathcal{S}\cap B\neq\varnothing$ and on $\mathcal{S}\cap W=\varnothing$, where $W\subseteq B^c$, then with probability bounded away from 0 we have $\mathcal{S}\cap B'\cap\mathcal{Z}\neq\varnothing$, where B' is a square of $\frac{1}{3}$ the side length that is concentric with B; namely,

$$\mathbb{P}[\mathcal{S}\cap B'\cap\mathcal{Z}\neq\varnothing \,|\, \mathcal{S}\cap W=\varnothing, \mathcal{S}\cap B\neq\varnothing] > a > 0 \tag{1.11}$$

for some constant a. This is based on a second-moment argument, but to carry it through we have to resort to rather involved percolation arguments. A key observation here is to interpret these conditional events for the spectral sample in terms of percolation events for a coupling of two configurations (which are independent on the set W but coincide elsewhere). An important step is to prove a quasi-multiplicativity property for arm-events in the case of this system of coupled configurations.

Again, there is a simple naive strategy based on (1.11) and (1.10) to get an upper bound for $\mathbb{P}[\mathcal{S}\neq\varnothing=\mathcal{Z}\cap\mathcal{S}]$. One may try to check sequentially if $B'\cap\mathcal{Z}\cap\mathcal{S}\neq\varnothing$ for each of the $r\times r$ squares B, and as long as a non-empty intersection has not been found, the probability to detect a non-empty intersection is proportional to the conditional probability that $\mathcal{S}\cap B\neq\varnothing$. However, the trouble with this strategy is that the conditional probability of $\mathcal{S}\cap B\neq\varnothing$ varies with time, and the bound (1.10) does not imply a similar bound for the sum of these conditional probabilities, since each time the conditioning is different. Hence we cannot conclude that $\mathcal{S}\cap B\neq\varnothing$ happens many times during the sequential checking. And this issue cannot be solved by first conditioning on having a large total number of non-empty intersections, because we cannot handle such a "positive conditioning".

The substitute for this naive strategy is a large deviation estimate that we state and prove in §6, namely, Proposition 6.1. This result is somewhat in the flavor of the Lovász local lemma and the domination of product measures result of [LSS] (which proves and uses an extension of the Lovász lemma), since it gives estimates for probabilities of events with a possibly complicated dependence structure, and more specifically, it says that certain positive conditional marginals ensure exponential decay for the probability of complete failure, similarly to what would happen in a product measure. However, our situation is more complicated than [LSS]: we have two random variables $x, y\in\{0,1\}^n$ instead of one random field, and we have only a much smaller set of positive conditional marginals to start with. In the application, x_i is the indicator of the event that $\mathcal{S}$ intersects the ith $r\times r$ square, and y_i is the indicator of the event that $\mathcal{S}\cap\mathcal{Z}$ intersects that square. The assumption (1.11) then translates to

$$\mathbb{P}[y_j=1\,|\,y_i=0 \text{ for all } i\in I]\geqslant a\mathbb{P}[x_j=1\,|\,y_i=0 \text{ for all } i\in I],\quad j\notin I\subset[n],$$

and the proposition tells us that under these assumptions we have

$$\mathbb{P}[y=0\,|\,X>0]\leqslant a^{-1}\mathbb{E}[e^{-aX/e}\,|\,X>0], \tag{1.12}$$

where $X=\sum_i x_i$. In our application $X=|\mathcal{S}(r)|$, and thus (1.12) combines with (1.11) to yield the desired bound. The proof of Theorem 1.1 is completed in this way in §7.

Remark 1.7. As we mentioned earlier, the above results about the tightness of the spectrum when normalized by its mean (as well as the up to constants optimal results on the lower tail) are much harder to achieve than their analogs for the set of pivotal points, even though the two are intimately related. Indeed, the techniques of this paper easily adapt to the set $\mathscr{P}_{f_R}$ of pivotal points of the left-right crossing of the square $[0,R]^2$ and give tightness results on the number of pivotal points $|\mathscr{P}_{f_R}|$. For instance, they imply the following analog of Theorem 1.1:

$$\mathbb{P}[0<|\mathscr{P}_{f_R}|<\mathbb{E}|\mathscr{P}_{f_r}|]\asymp\left(\frac{R}{r}\right)^2\alpha_6(r,R).$$

The appearance of $\alpha_6(r,R)$ in place of $\alpha_4(r,R)^2$ will be explained in Remark 4.6. In the case of the triangular grid, this gives the following estimate on the lower tail:

$$\limsup_{R\to\infty}\mathbb{P}[0<|\mathscr{P}_{f_R}|\leqslant\lambda\mathbb{E}|\mathscr{P}_{f_R}|]=\lambda^{11/9+o(1)},$$

as $\lambda\to 0$. It turns out that in the case of the pivotal points $\mathscr{P}_{f_R}$ (where the i.i.d. structure of the percolation configuration helps), there is a more direct route towards the tightness of $|\mathscr{P}_{f_R}|/\mathbb{E}|\mathscr{P}_{f_R}|$ on $(0,\infty)$ than the route we needed to follow for the spectrum $\mathscr{S}$. We do not give more details, since we will not use this pivotal tightness in this paper.

One might think that this good control on the lower tail of the number of pivotals should imply that if one waits long enough, the system must switch many times between having and not having a left-right crossing, and then this should imply an almost complete decorrelation. However, assuming the first implication for now, the second one still remains completely unclear: even if there are a lot of switches, it might well be that the system started, say, from having the left-right crossing, will remember this for a long time in the sense that it will move back more quickly from not having the crossing to having the crossing, than vice versa. In this case, at a given time it would be significantly more likely (by a constant factor) that an even number of switches have happened so far, i.e., the system would not have lost most of the correlation. We do not see how to rule out this scenario by studying pivotals only, that is why the Fourier spectrum is so useful.

Acknowledgments. We are grateful to Gil Kalai for inspiring conversations and for his observation (2.11), and to Jeff Steif for insights regarding the dimensions of exceptional times in dynamical percolation and for numerous comments on the manuscript. We thank Vincent Beffara for permitting us to include Proposition A.1, Alan Hammond, Pierre Nolin and Wendelin Werner for motivating and enlightening discussions, and Erik Broman and Alan Hammond for suggestions on how to improve readability at several places. Finally, we are extremely grateful to a referee for his or her careful reading and many detailed comments.

2. Some basics

2.1. A few general definitions

In this paper we consider site percolation on the triangular grid as well as bond percolation on $\mathbb{Z}^2$, both at the critical parameter $p=\frac{1}{2}$.

In the case of site percolation on the triangular grid T, a percolation configuration ω is just the set of sites which are open. However, we often think of ω as a coloring of the plane by two colors: in the hexagonal grid dual to T, a hexagon is colored white if the corresponding site is in ω, while the other hexagons are colored black. If A and B are subsets of the plane, we say that there is a *crossing* in ω from A to B if there is a continuous path with one endpoint in A and the other endpoint in B that is contained in the closure of the union of the white hexagons. Likewise, a *dual crossing* corresponds to a path contained in the closure of the black hexagons.

In the case of bond percolation on $\mathbb{Z}^2$, there is a similar coloring of the plane by two colors which has the "correct" connectivity properties. In this case, we color by white all the points that are within L^∞ distance of $\frac{1}{4}$ from all the vertices of $\mathbb{Z}^2$ and all the points that are within L^∞ distance of $\frac{1}{4}$ from the edges in ω, and color by black the closure of the complement of the white colored points.

Regardless of the grid, the set of points whose color is determined by ω_x will be called the *tile* of x. In the case of the square grid, we also have tiles with deterministic color, namely, each square of side length $\frac{1}{2}$ centered at a vertex of $\mathbb{Z}^2$ and each square of side length $\frac{1}{2}$ concentric with a face of $\mathbb{Z}^2$. Thus, in either case we have a tiling of the plane by hexagons or squares where each tile consists of a connected set of points whose colors always agree.

A *quad* $\mathcal{Q}$ is a subset of the plane homemorphic to the closed unit disk together with a distinguished pair of disjoint closed arcs on $\partial\mathcal{Q}$. We say that ω has a *crossing* of $\mathcal{Q}$ if the two distinguished arcs can be connected by a white path inside $\mathcal{Q}$.

If $\mathcal{A}$ is an event, then the ±1 *indicator* function of $\mathcal{A}$ is the function $2\cdot1_{\mathcal{A}}-1$, which is 1 on $\mathcal{A}$ and -1 on $\neg\mathcal{A}$. The ±1-indicator function for the event that a quad $\mathcal{Q}$ is crossed will be denoted by $f_{\mathcal{Q}}$.

We use $\mathcal{I}$ to denote the set of bits in ω; that is, in the context of the triangular grid $\mathcal{I}$ is the set of vertices of the grid, and in the context of $\mathbb{Z}^2$ it denotes the set of edges. Although $\mathcal{I}$ is not finite in these cases, the functions we consider will only depend on finitely many bits in $\mathcal{I}$, and so the Fourier–Walsh expansion (1.1) still holds. Moreover, for L^2 functions depending on infinitely many bits we still have (1.1), except that the summation is restricted to finite $S\subseteq\mathcal{I}$.

Since we will be considering $\mathscr{S}$ as a geometric object, we find it convenient to think

of $\mathcal{I}$ as a discrete set in the plane. In the context of the triangular grid, this is anyway the case, but for the square grid we will implicitly associate each edge of $\mathbb{Z}^2$ with its center; so $\mathcal{I}$ can be considered as the set of centers of the relevant edges. This way, any subset of the plane also represents a subset of the bits. Note however that e.g. the crossing function $f_{\mathcal{Q}}$ usually depends on more bits than the ones contained in $\mathcal{Q}$.

For $z\in\mathbb{R}^2$ and $r\geqslant 0$, the set $z+[-r,r)^2$ will be called the square of radius r centered at z. Furthermore, we let $B(z,r)$ denote the union of the tiles whose center is contained in $z+[-r,r)^2$, and will refer to $B(z,r)$ as a *box* of radius r. One reason for using these boxes (instead of round balls, say) is that the plane can be tiled perfectly with them.

2.2. Multi-arm events for percolation

In many different studies of percolation, the multi-arm events play a central role. We now define these events (a word of caution—there are a few different natural variants to these definitions), and discuss the asymptotics of their probabilities.

Let $A\subset\mathbb{R}^2$ be some topological annulus in the plane, and let $j\in\mathbb{N}_+$. If j is even, then the *j-arm event* in A is the event that there are j disjoint monochromatic paths joining the two boundary components of A, and these paths in circular order are alternating between white and black. If j is odd, the definition is similar, except that the order of the colors is required to be (in circular order) alternating between white and black with one additional white crossing.

In most papers, the restriction that the colors are alternating is relaxed to the requirement that not all crossings are of the same color. Indeed, it is known that if A is an annulus, $A=B(0,R)\setminus B(0,r)$, then in the setting of critical site percolation on the triangular grid the circular order of the colors effects the probability of the event by at most a constant factor (which may depend on j), provided that in the case $j>1$ there is at least one required crossing from each color [ADA]. However, since it appears that the corresponding result for the square grid has not been worked out, we have opted to impose the alternating colors restriction.

We let $\alpha_j(A)$ denote the probability of the j-arm event in A. For the case $A=B(0,R)\setminus B(0,r)$, write $\alpha_j(r,R)$ for $\alpha_j(A)$. Note that $\alpha_j(r,R)=0$ if $r\ll j<R$. We use $\alpha_j(R)$ as a shorthand for $\alpha_j(2j,R)$. We will also adopt the convention that $\alpha_j(r,R)=1$ if $r\geqslant R$.

We now review some of the results concerning these arm events. The Russo–Seymour–Welsh (RSW) estimates imply that

$$\frac{s^{-a_j}}{C_j}\leqslant\alpha_j(r,sr)\leqslant C_j s^{-1/a_j} \tag{2.1}$$

for all $r>r(j)$ and $s>1$, where $r(j), C_j, a_j>1$ depend only on j. Another important property of these arm events is *quasi-multiplicativity*, namely,

$$\frac{\alpha_j(R)}{C_j} \leqslant \alpha_j(r)\alpha_j(r,R) \leqslant C_j\alpha_j(R) \tag{2.2}$$

for $r(j)<r<R$, where, again, $r(j), C_j>1$ depend only on j. This was proved in [K2]; see [SSt, Proposition A.1] and [N, §4] for concise proofs. The above properties in particular give for $r<r'<R'<R$ that

$$C_j^{-1}\left(\frac{Rr'}{R'r}\right)^{a_j^{-1}} \alpha_j(r,R) \leqslant \alpha_j(r',R') \leqslant C_j\left(\frac{Rr'}{R'r}\right)^{a_j} \alpha_j(r,R), \tag{2.3}$$

with possibly different constants C_j.

Of the multi-arm events, the most relevant to this paper is the 4-arm event, due to its relation to pivotality for the crossing event in a quad $\mathcal{Q}$. In particular, for closed $B\subset\mathbb{R}^2$, we will use $\alpha_\square(B,\mathcal{Q})$ to denote the probability of having four arms in $\mathcal{Q}\setminus B$, the white arms connecting ∂B to the two distinguished arcs on $\partial\mathcal{Q}$ and the black arms to the complementary arcs. If $B\cap\partial\mathcal{Q}\neq\varnothing$, then the arms connecting B to the arcs of $\partial\mathcal{Q}$ which B intersects are considered as present. Quasi-multiplicativity often generalizes easily to this quantity; for example, if $\mathcal{Q}$ is an $R\times R$ square with two opposite sides being the distinguished arcs, B is a radius r box anywhere in $\mathcal{Q}$, and $x\in B$ is at a distance at least cr from ∂B, then

$$\frac{\alpha_\square(x,\mathcal{Q})}{\alpha_\square(B,\mathcal{Q})} \asymp \alpha_4(x,B) \asymp \alpha_4(r), \tag{2.4}$$

with the implied constants depending only on c. (A more general version of this will also be proved in §5.5.) Here and in the following, when we write $\alpha_\square(x,\mathcal{Q})$ or $\alpha_4(x,B)$, we are referring to the corresponding 4-arm event from the tile of x to $\partial\mathcal{Q}$ or ∂B (with or without paying attention to any distinguished arms on the boundary, respectively).

Let us also recall what is known about α_4 quantitatively. For site percolation on the triangular lattice, by [SW], we have

$$\alpha_4(r,R) = \left(\frac{r}{R}\right)^{5/4+o(1)} \tag{2.5}$$

as $R/r\to\infty$ while $1\leqslant r\leqslant R$. Similar relations are known for $j\neq 4$ [LSW2], [SW]. For bond percolation on the square grid, we presently have weaker estimates; in particular,

$$C^{-1}\left(\frac{r}{R}\right)^{2-\varepsilon} \leqslant \alpha_4(r,R) \leqslant C\left(\frac{r}{R}\right)^{1+\varepsilon} \tag{2.6}$$

for some fixed constants $C,\varepsilon>0$ and every $1\leqslant r\leqslant R$. The left inequality can be obtained by combining $\alpha_5(r,R)\asymp(r/R)^2$ (see [KSZ, Lemma 5] or [SSt, Corollary A.8]), the RSW estimate $\alpha_1(r,R)<O(1)(r/R)^\varepsilon$, and, finally, the relation $\alpha_1(r,R)\alpha_4(r,R)\geqslant\alpha_5(r,R)$ (which follows from Reimer's inequality [R], or from Proposition A.1 in our appendix). The right-hand inequality in (2.6) follows from [K2]; see also [BKS, Remark 4.2].

2.3. The spectral sample in general

This subsection derives some formulas and estimates for $\mathbb{P}[\mathscr{S}\subseteq A]$, for the distribution of $\mathscr{S}\cap A$, and for $\mathbb{P}[\mathscr{S}\cap A=\varnothing\neq\mathscr{S}\cap B]$. We also briefly present an estimate of the variation distance between the laws of $\mathscr{S}_f$ and of $\mathscr{S}_g$ in terms of $\|f-g\|$. (We generally use $\|\cdot\|$ to denote the L^2 norm.) Moreover, the definition of the set of pivotals $\mathscr{P}$ is recalled and some relations between $\mathscr{P}$ and $\mathscr{S}$ are discussed.

As before, let $\Omega:=\{-1,1\}^{\mathcal{I}}$, where $\mathcal{I}$ is finite. Recall that for $f:\Omega\to\mathbb{R}$ with $\|f\|=1$, we consider the random variable $\mathscr{S}_f$ whose law is given by $\mathbb{P}[\mathscr{S}=S]=\hat{f}(S)^2$. More generally, if $\|f\|>0$, we use the law given by

$$\mathbb{P}[\mathscr{S}=S]=\frac{\hat{f}(S)^2}{\|f\|^2},$$

but will also consider the un-normalized measure given by

$$\mathbb{Q}[\mathscr{S}=S]=\hat{f}(S)^2.$$

(If we wish to indicate the function f, we may write $\mathbb{Q}_f$ in place of $\mathbb{Q}$.)

Now suppose that $f,g:\Omega\to\mathbb{R}$. We argue that if $\|f-g\|$ is small, then the law of $\mathscr{S}_f$ is close to the law of $\mathscr{S}_g$, as follows. If we want to compare the "laws" under $\mathbb{Q}$, then

$$\begin{aligned}\sum_{S\subseteq\mathcal{I}}|\hat{f}(S)^2-\hat{g}(S)^2| &= \sum_S|\hat{f}-\hat{g}|\,|\hat{f}+\hat{g}|\\ &\leqslant\left(\sum_S(\hat{f}-\hat{g})^2\right)^{1/2}\left(\sum_S(\hat{f}+\hat{g})^2\right)^{1/2}\\ &=\|f-g\|\,\|f+g\|,\end{aligned}\tag{2.7}$$

where the inequality is due to Cauchy–Schwarz' inequality and the final equality is an application of Parseval's identity. For the laws under $\mathbb{P}$, when $\|f\|=\|g\|=1$ does not hold, a slightly more complicated version of this argument gives that if $\|f-g\|\leqslant\varepsilon\max\{\|f\|,\|g\|\}$ for some $\varepsilon\in(0,1)$, then

$$\sum_{S\subseteq\mathcal{I}}\left|\frac{\hat{f}(S)^2}{\|f\|^2}-\frac{\hat{g}(S)^2}{\|g\|^2}\right|\leqslant\frac{4\varepsilon}{(1-\varepsilon)^2}.$$

We will not use this version, and hence omit the details of its derivation.

For $A\subseteq\mathcal{I}$, let ω_A denote the restriction of ω to A, and let $\mathcal{F}_A$ denote the σ-field of subsets of Ω generated by ω_A. We use the notation $A^c:=\mathcal{I}\setminus A$ for the complement of A. Observe that for $A\subseteq\mathcal{I}$,

$$\mathbb{E}[\chi_S\,|\,\mathcal{F}_A]=\begin{cases}\chi_S, & \text{if } S\subseteq A,\\ 0, & \text{otherwise.}\end{cases}\tag{2.8}$$

It follows from this and (1.1) that

$$g := \mathbb{E}[f \mid \mathcal{F}_A] = \sum_{S \subseteq A} \hat{f}(S)\chi_S.$$

Thus

$$\hat{g}(S) = \begin{cases} \hat{f}(S), & \text{if } S \subseteq A, \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, Parseval's formula implies that

$$\mathbb{Q}[\mathscr{S} \subseteq A] = \mathbb{E}[\mathbb{E}[f \mid \mathcal{F}_A]^2]. \tag{2.9}$$

In principle, this describes the distribution of $\mathscr{S}$ in terms of f, and indeed we will extract information about $\mathscr{S}$ from this formula and its consequences.

Using (1.1) and (2.8), one obtains for $S \subseteq A \subseteq \mathcal{I}$ that

$$\mathbb{E}[f\chi_S \mid \mathcal{F}_{A^c}] = \sum_{S' \subseteq A^c} \hat{f}(S \cup S')\chi_{S'}.$$

This gives

$$\mathbb{E}[\mathbb{E}[f\chi_S \mid \mathcal{F}_{A^c}]^2] = \sum_{S' \subseteq A^c} \hat{f}(S \cup S')^2 = \mathbb{Q}[\mathscr{S} \cap A = S],$$

which implies the following lemma from [LMN]. Roughly, the lemma says that in order to sample the random variable $\mathscr{S} \cap A$, one can first pick a random sample of ω_{A^c}, and then take a sample from the spectral sample of the function we get by plugging in these values for the bits in A^c.

LEMMA 2.1. *Suppose that $f: \Omega \to \mathbb{R}$ and $A \subseteq \mathcal{I}$. For $x \in \{-1,1\}^A$ and $y \in \{-1,1\}^{A^c}$, write $g_y(x) := f(\omega(x,y))$, where $\omega(x,y)$ is the element of Ω whose restriction to A is x and whose restriction to A^c is y. Then, for every $S \subseteq A$, we have*

$$\mathbb{Q}[\mathscr{S}_f \cap A = S] = \mathbb{E}[\hat{g}_y(S)^2] = \mathbb{E}[\mathbb{Q}_{g_y}[\mathscr{S}_{g_y} = S]]. \qquad \square$$

For any $\omega \in \Omega$ and any $A \subseteq \mathcal{I}$, let ω_A^+ denote the element of Ω that is equal to 1 in A and equal to ω outside of A. Similarly, let ω_A^- denote the element of Ω that is equal to -1 in A and equal to ω outside of A. An $i \in \mathcal{I}$ is said to be *pivotal* for $f: \Omega \to \mathbb{R}$ and ω if $f(\omega_{\{i\}}^+) \neq f(\omega_{\{i\}}^-)$. Let $\mathscr{P} = \mathscr{P}_f$ denote the (random) set of pivotals.

It is known from [KKL] that for functions $f: \Omega \to \{-1,1\}$,

$$\mathbb{E}[|\mathscr{S}|] = \mathbb{E}[|\mathscr{P}|]. \tag{2.10}$$

Gil Kalai (private communication) further observed that it also holds for the second moment

$$\mathbb{E}[|\mathscr{S}|^2]=\mathbb{E}[|\mathscr{P}|^2], \tag{2.11}$$

but this does not extend to higher moments. (Note that in these formulas, the expectation $\mathbb{E}$ is with respect to different measures on the left-hand and right-hand sides.) To prove (2.10) and (2.11), consider some $i,j\in\mathcal{I}$. In Lemma 2.1, if we take $A=\{i\}$, then g_y is a constant function (of x) unless $i\in\mathscr{P}$, while $g_y=\pm\chi_{\{i\}}$ if $i\in\mathscr{P}$. Therefore, the lemma gives

$$\mathbb{P}[i\in\mathscr{S}]=\mathbb{P}[\mathscr{S}\cap\{i\}=\{i\}]=\mathbb{P}[i\in\mathscr{P}], \tag{2.12}$$

which sums to give (2.10). Similarly, one can show that

$$\mathbb{P}[i,j\in\mathscr{S}]=\mathbb{P}[i,j\in\mathscr{P}] \tag{2.13}$$

by using Lemma 2.1 with $A=\{i,j\}$ to reduce (2.13) to the case where $\Omega=\{-1,1\}^2$, which easily yields to direct inspection. Now (2.11) follows by summing (2.13) over i and j.

As we have discussed in the introduction, (1.2) immediately shows that the distribution of the size $|\mathscr{S}_f|$ governs the sensitivity of $f:\Omega\to\{-1,1\}$ to ε-noise: if $\varepsilon\mathbb{E}[|\mathscr{S}|]$ is small, then the correlation is always close to 1, while the ε needed for decorrelation is given by the lower tail of $|\mathscr{S}|$. One can also ask finer questions, concerning "sensitivity to selective noise": which deterministic subsets $\mathcal{U}\subseteq\mathcal{I}$ have the property that knowing the bits in $\mathcal{U}$ (and having unknown independent random bits in $\mathcal{U}^c$) we can predict f with probability close to 1, and which subsets $\mathcal{U}$ give almost no information on f? In §8.2 we will see, using (2.9), that the first case ($\mathcal{U}$ is "almost decisive") happens if and only if $\mathbb{P}[\mathscr{S}\subseteq\mathcal{U}]$ is close to 1, while the second case ($\mathcal{U}$ is "almost clueless") happens if and only if $\mathbb{P}[\varnothing\neq\mathscr{S}\subseteq\mathcal{U}]$ is close to 0. For percolation, we will be able to understand both cases quite well, and will also see that although the pivotal and spectral sets, $\mathscr{P}$ and $\mathscr{S}$, are different in many ways, it is reasonable to conjecture that they have the same decisive and clueless sets (see Remark 8.6). There is probably a similar phenomenon for many Boolean functions.

We now derive estimates for $\mathbb{Q}[\mathscr{S}\cap B\neq\varnothing=\mathscr{S}\cap W]$, when B and W are disjoint subsets of the bits. Define $\Lambda_B=\Lambda_{f,B}$ as the event that B is pivotal for f. More precisely, Λ_B is the set of $\omega\in\Omega$ such that there is some $\omega'\in\Omega$ that agrees with ω on B^c while $f(\omega)\neq f(\omega')$. Also define $\lambda_{B,W}:=\mathbb{P}[\Lambda_B\,|\,\mathcal{F}_{W^c}]$.

LEMMA 2.2. *Let $\mathscr{S}=\mathscr{S}_f$ be the spectral sample of some $f:\Omega\to\mathbb{R}$, and let W and B be disjoint subsets of $\mathcal{I}$. Then*

$$\mathbb{Q}[\mathscr{S}\cap B\neq\varnothing=\mathscr{S}\cap W]\leqslant 4\|f\|_\infty^2\mathbb{E}[\lambda_{B,W}^2].$$

Proof. From (2.9),

$$\begin{aligned}\mathbb{Q}[\mathscr{S}\cap B\neq\varnothing,\mathscr{S}\cap W=\varnothing]&=\mathbb{Q}[\mathscr{S}\subseteq W^c]-\mathbb{Q}[\mathscr{S}\subseteq(W\cup B)^c]\\&=\mathbb{E}[\mathbb{E}[f\,|\,\mathcal{F}_{W^c}]^2-\mathbb{E}[f\,|\,\mathcal{F}_{(W\cup B)^c}]^2]\\&=\mathbb{E}[(\mathbb{E}[f\,|\,\mathcal{F}_{W^c}]-\mathbb{E}[f\,|\,\mathcal{F}_{(W\cup B)^c}])^2].\end{aligned}\tag{2.14}$$

On the complement of Λ_B, we have $f=\mathbb{E}[f\,|\,\mathcal{F}_{B^c}]$. Therefore,

$$-2\|f\|_\infty 1_{\Lambda_B}\leqslant f-\mathbb{E}[f\,|\,\mathcal{F}_{B^c}]\leqslant 2\|f\|_\infty 1_{\Lambda_B}.$$

Taking conditional expectations throughout, we get

$$-2\|f\|_\infty\lambda_{B,W}\leqslant\mathbb{E}[f\,|\,\mathcal{F}_{W^c}]-\mathbb{E}[\mathbb{E}[f\,|\,\mathcal{F}_{B^c}]\,|\,\mathcal{F}_{W^c}]\leqslant 2\|f\|_\infty\lambda_{B,W}.$$

Note that $\mathbb{E}[\mathbb{E}[f\,|\,\mathcal{F}_{B^c}]\,|\,\mathcal{F}_{W^c}]=\mathbb{E}[f\,|\,\mathcal{F}_{(B\cup W)^c}]$, since our measure on Ω is i.i.d. Thus, the above gives

$$|\mathbb{E}[f\,|\,\mathcal{F}_{W^c}]-\mathbb{E}[f\,|\,\mathcal{F}_{(B\cup W)^c}]|\leqslant 2\|f\|_\infty\lambda_{B,W}.$$

An appeal to (2.14) now completes the proof. □

Taking $W=\varnothing$ yields $\mathbb{Q}\big[\mathscr{S}\cap B\neq\varnothing\big]\leqslant 4\|f\|_\infty\mathbb{P}[\Lambda_B]$, since $\Lambda_{B,\varnothing}=1_{\Lambda_B}$. Taking $W=B^c$ yields $\mathbb{Q}\big[\varnothing\neq\mathscr{S}\subseteq B\big]\leqslant 4\|f\|_\infty\mathbb{P}[\Lambda_B]^2$, since $\Lambda_{B,B^c}=\mathbb{P}[\Lambda_B]$. See also Lemma 3.2 and the discussion afterwards.

Remark 2.3. In the special case when f is monotone and ±1-valued, and $B=\{x\}$ is a single bit, the lemma takes a more precise form, as we will prove in §5.3:

$$\mathbb{P}[x\in\mathscr{S},\mathscr{S}\cap W=\varnothing]=\mathbb{E}[\lambda_{x,W}^2].$$

What turns out to be important in §5 is that, in the context of percolation, the quantity $\mathbb{E}[\lambda_{B,W}^2]$ can be studied and controlled when B is a box and $W\subseteq\mathcal{I}\setminus B$ is arbitrary. Likewise, in §4, we use a variant of Lemma 2.2 in which we look at the event that $\mathscr{S}$ intersects a collection of boxes and is disjoint from some collection of annuli.

3. First percolation spectrum estimates

We will now consider the special case of $\mathscr{S}_f$ when $f=f_\mathcal{Q}$ for a quad $\mathcal{Q}\subset\mathbb{R}^2$. As noted in §2.1, we will be considering $\mathcal{I}$ and $\mathscr{S}\subseteq\mathcal{I}$ as subsets of the plane. When $R>0$, we will use the notation $R\mathcal{Q}$ to denote the quad obtained from $\mathcal{Q}$ by scaling by a factor of R about 0.

LEMMA 3.1. (First and second moments) *Let $\mathcal{Q}\subset\mathbb{R}^2$ be some quad and let U be an open set whose closure is contained in the interior of $\mathcal{Q}$. Let $\mathcal{S}:=\mathcal{S}_{f_{R\mathcal{Q}}}$ be the spectral sample for the ±1-indicator function of crossing $R\mathcal{Q}$. There are constants $C, R_0>0$, depending only on $\mathcal{Q}$ and U, such that for all $R>R_0$,*

$$C^{-1}R^2\alpha_4(R)\leqslant\mathbb{E}[|\mathcal{S}_{f_{R\mathcal{Q}}}\cap RU|]\leqslant CR^2\alpha_4(R)$$

and

$$\mathbb{E}[|\mathcal{S}_{f_{R\mathcal{Q}}}\cap RU|^2]b\leqslant C\mathbb{E}[|\mathcal{S}_{f_{R\mathcal{Q}}}\cap RU|]^2.$$

The reason for the appearance of α_4 in the first moment is the following. We know from (2.12) that $\mathbb{P}[x\in\mathcal{S}]=\mathbb{P}[x\in\mathscr{P}]$ for $\mathscr{P}:=\mathscr{P}_{f_{R\mathcal{Q}}}$. In order for x to be pivotal for $f_{R\mathcal{Q}}$ it is necessary and sufficient that there are white paths in $R\mathcal{Q}$ from the tile of x to the two distinguished arcs on $R\partial\mathcal{Q}$ and two black paths in $R\mathcal{Q}$ from the tile of x to the complementary arcs of $R\partial\mathcal{Q}$. These four paths form the 4-arm event in the annulus between the tile of x and $R\partial\mathcal{Q}$. Moreover, it is well known (and follows from quasi-multiplicativity arguments; see [K2], [W]) that

$$\mathbb{P}[x\in\mathscr{P}_{f_{R\mathcal{Q}}}]=\alpha_\square(x,R\mathcal{Q})\asymp\alpha_4(R)\quad\text{for all } x\in\mathcal{I}\cap RU. \tag{3.1}$$

Here and in the proof below, we use the notation $g\asymp g'$ to mean that there is a constant $c>0$ (which may depend on U and $\mathcal{Q}$), such that $g\leqslant cg'$ and $g'\leqslant cg$. Likewise, the $O(\cdot)$ notation will involve constants that may depend on $\mathcal{Q}$ and U.

Proof. From (2.12) and (3.1) we get

$$\mathbb{P}[x\in\mathcal{S}]\asymp\alpha_4(R)\quad\text{for all } x\in\mathcal{I}\cap RU. \tag{3.2}$$

The first claim of the lemma is obtained by summing over $x\in\mathcal{I}\cap RU$.

Now consider $x,y\in\mathcal{I}\cap RU$. Let a be the distance from $\overline{U}$ to $\partial\mathcal{Q}$. Thus $a>0$. Then by (2.13), we have $\mathbb{P}[x,y\in\mathcal{S}]=\mathbb{P}[x,y\in\mathscr{P}]$. In order for $x,y\in\mathscr{P}$ it is necessary that the 4-arm event occurs from the tile of x to distance $\left(\frac{1}{3}|x-y|\right)\wedge(aR)$ away, and from the tile of y to distance $\left(\frac{1}{3}|x-y|\right)\wedge(aR)$ away, and from the circle of radius $2|x-y|$ around $\frac{1}{2}(x+y)$ to distance aR away (if $2|x-y|<aR$). By independence on disjoint subsets of $\mathcal{I}$, this (together with the regularity properties of the 4-arm probabilities (2.3)) gives, for R sufficiently large, that

$$\mathbb{P}[x,y\in\mathcal{S}]\leqslant O(1)\alpha_4(|x-y|)^2\alpha_4(|x-y|,R).$$

Using the quasi-multiplicativity property of α_4, this gives

$$\mathbb{P}[x,y\in\mathcal{S}]\leqslant O(1)\frac{\alpha_4(R)^2}{\alpha_4(|x-y|,R)}\quad\text{for all } x,y\in\mathcal{I}\cap RU. \tag{3.3}$$

The number of pairs $x,y\in\mathcal{I}\cap U$ such that $|x-y|\in[2^n,2^{n+1})$ is $O(R^2)2^{2n}$, and is zero if $|x-y|>R\,\mathrm{diam}(\mathcal{Q})$. Therefore, we get from (3.3) and the regularity property (2.3) that

$$\mathbb{E}[|\mathcal{S}\cap RU|^2]\leqslant O(R^2)\alpha_4(R)^2\sum_{n=0}^{\log_2(R)+O(1)}\frac{2^{2n}}{\alpha_4(2^n,R)}.$$

From (2.6) we get $2^{2n}/\alpha_4(2^n,R)\leqslant O(1)R^{2-\varepsilon}2^{\varepsilon n}$. Hence the sum over n is at most $O(R^2)$, and we obtain the desired bound on the second moment. □

Note that $\mathbb{E}[|\mathcal{S}\cap RU|]\to\infty$ as $R\to\infty$, which follows from Lemma 3.1 and (2.6). Moreover, by the standard Cauchy–Schwarz second-moment argument (also called the Paley–Zygmund inequality), the above lemma implies that for some constant $c>0$ (which may depend on $\mathcal{Q}$ and U),

$$\mathbb{P}[|\mathcal{S}\cap RU|>c\mathbb{E}|\mathcal{S}\cap RU|]>c\quad\text{for all } R>0.$$

We also note the following lemma.

Lemma 3.2. *Let $\mathcal{Q}\subset\mathbb{R}^2$ be a quad, and set $\mathcal{S}=\mathcal{S}_{f_\mathcal{Q}}$. Let B be some union of tiles such that $B\cap\mathcal{Q}$ is non-empty and connected. Then*

$$\mathbb{P}[\mathcal{S}\cap B\neq\varnothing]\leqslant 4\alpha_\square(B,\mathcal{Q}) \tag{3.4}$$

and

$$\mathbb{P}[\varnothing\neq\mathcal{S}\subseteq B]\leqslant 4\alpha_\square(B,\mathcal{Q})^2. \tag{3.5}$$

When B is a single tile, corresponding to $x\in\mathcal{I}$, we have

$$\mathbb{P}[x\in\mathcal{S}]=\alpha_\square(x,\mathcal{Q})\quad\textit{and}\quad\mathbb{P}[\mathcal{S}=\{x\}]=\alpha_\square(x,\mathcal{Q})^2. \tag{3.6}$$

Proof. Note that $\mathbb{P}[\Lambda_B]=\alpha_\square(B,\mathcal{Q})$, since Λ_B holds if and only if the 4-arm event from B to the corresponding arcs on $\partial\mathcal{Q}$ occurs. Since $\lambda_{B,\varnothing}=1_{\Lambda_B}$, the first claim follows from Lemma 2.2 with $W=\varnothing$. Similarly, (3.5) follows by taking $W=B^c$. The identity $\mathbb{P}[x\in\mathcal{S}]=\alpha_\square(x,\mathcal{Q})$ follows from (2.12). Finally, the right-hand identity in (3.6) can be derived from (2.14) with $B=\{x\}$ and $W=\mathcal{I}\setminus B$. Alternatively, it also follows from $\mathbb{P}[\{x\}=\mathcal{S}]=\mathbb{P}[x\in\mathcal{S}]^2$, which holds for arbitrary monotone $f:\Omega\to\{-1,1\}$. □

As we will see in §5, both inequalities in Lemma 3.2 are actually approximate equalities when $B\subset\mathcal{Q}$. The main reason for this is a classical arm separation phenomenon, see e.g. [SSt, Appendix, Lemma A.2], which roughly says that conditioned on having four arms connecting ∂B to the appropriate boundary arcs on $\partial\mathcal{Q}$, with positive conditional probability, these arms are "well separated" on ∂B. On this event, a positive

proportion of the ω_B configurations enable the crossing of $\mathcal{Q}$, while a positive proportion disable all crossings. However, for radial crossings, the estimates from Lemma 2.2 are not sharp—see Lemma 4.8 and the discussion afterwards.

Lemma 3.2 has the following immediate consequence. Let $\mathcal{Q}$ be the $R\times R$ square with two opposite sides as distinguished boundary arcs. Then, for a box $B\subseteq\mathcal{Q}$ of radius r and a concentric sub-box B' of radius $\frac{1}{3}r$, if $B'\cap\mathcal{I}\neq\varnothing$, then

$$\mathbb{E}\big[|\mathcal{S}\cap B'|\,\big|\,\mathcal{S}\cap B\neq\varnothing\big]=\sum_{x\in B'}\frac{\mathbb{P}[x\in\mathcal{S}]}{\mathbb{P}[\mathcal{S}\cap B\neq\varnothing]}\geqslant\sum_{x\in B'}\frac{\alpha_\square(x,\mathcal{Q})}{4\alpha_\square(B,\mathcal{Q})}\asymp|B'|\alpha_4(r)\asymp r^2\alpha_4(r),\tag{3.7}$$

where we used the quasi-multiplicativity result (2.4). This result already suggests that $\mathcal{S}$ has self-similarity properties that a random fractal-like object should have, and it should be unlikely that it is very small. This idea will, in fact, be of key importance to us, and will be developed in §5.

4. The probability of a very small spectral sample

4.1. The statement

In this section, we study the Fourier spectrum of the ±1 indicator function f of having a crossing of a square, or more generally, of a quad $\mathcal{Q}$.

Divide the plane into a lattice of $r\times r$ subsquares, that is, $r\mathbb{Z}^2$, and for any set of bits $S\subseteq\mathcal{I}$ define

$$S_r:=\{\text{those } r\times r \text{ squares that intersect } S\}.$$

In particular, $\mathcal{S}_r$ is the set of r-squares whose intersection with the spectral sample $\mathcal{S}$ of f is non-empty. The following is an estimate for the probability that $\mathcal{S}$ is very small, or, more generally, that $\mathcal{S}_r$ is very small.

Proposition 4.1. *Let $\mathcal{S}$ be the spectral sample of $f=f_{[0,R]^2}$, the ±1-indicator function of the left-right crossing of the square $[0,R]^2$. For $g(k):=2^{\vartheta\log_2^2(k+2)}$, with $\vartheta>0$ large enough, and $\gamma_r(R):=(R/r)^2\alpha_4(r,R)^2$,*

$$\mathbb{P}[|\mathcal{S}_r|=k]\leqslant g(k)\gamma_r(R)\quad\textit{for all } k,R,r\in\mathbb{N}_+.$$

The square factor $(R/r)^2$ in the definition of γ_r reflects the 2-dimensionality of the ambient space—it corresponds to the number of different ways to choose a square of size r inside $[0,R]^2$. The factor $\alpha_4(r,R)^2$ comes from the second inequality in Lemma 3.2. In fact, since that lemma will turn out to be sharp up to a constant factor (when the

r-square is not close to the boundary of $[0,R]^2$), the $k=1$ case of Proposition 4.1 is also sharp. When we use this proposition, this will be important, as well as the fact that the dependence on k is subexponential.

Recall from (2.5) that for critical site percolation on the triangular lattice, we have $\gamma_r(R)=(r/R)^{1/2+o(1)}$, as $R/r\to\infty$. Also, note that for critical bond percolation on $\mathbb{Z}^2$ we have $\gamma_r(R)<O(1)(r/R)^{\varepsilon}$ for some $\varepsilon>0$ by (2.6).

For either lattice, the $r=1$ case of the proposition implies that for arbitrary $C,a>0$, we have $\lim_{R\to\infty}\mathbb{P}[0<|\mathscr{S}|<C(\log R)^a]=0$. This is already stronger than the $a=1$ and small C result of [BKS], whose proof used more analysis but less combinatorics.

In order to demonstrate the main ideas of the proof of the proposition in a slightly simpler setting in which considerations having to do with the boundary of the square do not appear, we will first state and prove a "local" version of this proposition. Then we will see in §4.3 that the boundary of the square indeed has no significant effect; the main reason for this is that the spectral sample "does not like" to be close to the boundary. (Let us note here that this phenomenon holds in any quad, as we will see in §7.4. However, the exact computations that are needed to get the bounds of Proposition 4.1 apply only to the case of the square, whose boundary is easy to handle.)

In §4.4, we will prove a radial version of Proposition 4.1, which will be similar to the square case, using one main additional trick.

4.2. A local result

PROPOSITION 4.2. *Consider some quad $\mathcal{Q}$, and let $\mathscr{S}$ be the spectral sample of $f=f_{\mathcal{Q}}$, the ±1-indicator function for the crossing event in $\mathcal{Q}$. Let $U'\subset U\subset\mathcal{Q}$, let R denote the diameter of U, let $a\in(0,1)$, and suppose that the distance from U' to the complement of U is at least aR. Let $\mathcal{S}(r,k)$ be the collection of all sets $S\subseteq\mathcal{I}$ such that $|(S\cap U)_r|=k$ and $S\cap(U\setminus U')=\varnothing$. Then, for g and γ_r as in Proposition* 4.1, *we have*

$$\mathbb{P}[\mathscr{S}\in\mathcal{S}(r,k)]\leqslant c_a g(k)\gamma_r(R)\quad \textit{for all } k,r\in\mathbb{N}_+,$$

where c_a is a constant that depends only on a.

We preface the proof of the proposition with a rough sketch of the main ideas. When $S\in\mathcal{S}(r,k)$ and k is small, the set $(S\cap U)_r$ has to consist of one or very few "clusters" of r-squares that are small (since their total cardinality is k, and the elements of each cluster are close to each other) and well separated from each other (otherwise they would not be distinct clusters). The probability that $(\mathscr{S}\cap U)_r$ has just one cluster, contained in a specific small box B of Euclidean diameter anywhere between r and $g(O(k))r$,

can be estimated by (3.5) of Lemma 3.2, at least if we assume $U=\mathcal{Q}$. For general U, we will soon prove a generalization of (3.5), Lemma 4.3. Then, one may sum over an appropriate collection of such small boxes B to get a reasonable bound for the probability that $\operatorname{diam}(\mathscr{S}\cap U)$ is small while $\mathscr{S}\cap U\neq\varnothing$. To deal with the case where $(\mathscr{S}\cap U)_r$ has a few different well-separated clusters, we will again use Lemma 4.3. The more involved part of the proof will be to classify the possible cluster structures of $\mathscr{S}$ and sum up the bounds corresponding to each possibility.

Proof. Let $\mathcal{A}$ be a finite collection of disjoint (topological) annuli in the plane; we call this an *annulus structure*. We say that a set $S\subset\mathbb{R}^2$ is *compatible* with $\mathcal{A}$ (or vice versa) if it is contained in $\mathbb{R}^2\setminus\bigcup\mathcal{A}$ and intersects the inner disk of each annulus in $\mathcal{A}$. Define $h(\mathcal{A})$ as the probability that each annulus in $\mathcal{A}$ has the 4-arm event. By independence on disjoint sets, we have

$$h(\mathcal{A})=\prod_{A\in\mathcal{A}}h(A). \tag{4.1}$$

Suppose that $\mathfrak{A}$ is a set of annulus structures $\mathcal{A}$ such that each annulus $A\in\mathcal{A}$ is contained in $\mathcal{Q}$, and $\mathfrak{A}$ has the property that each $S\in\mathcal{S}(r,k)$ must be compatible with at least one $\mathcal{A}\in\mathfrak{A}$. We claim that any such set $\mathfrak{A}$ satisfies

$$\mathbb{P}[\mathscr{S}\in\mathcal{S}(r,k)]\leqslant\sum_{\mathcal{A}\in\mathfrak{A}}h(\mathcal{A})^2. \tag{4.2}$$

For this, it is clearly sufficient to verify the following lemma, which is a generalization of (3.5) from Lemma 3.2.

LEMMA 4.3. *For any annulus structure $\mathcal{A}$ with $\bigcup\mathcal{A}\subset\mathcal{Q}$, we have*

$$\mathbb{P}[\mathscr{S}\text{ is compatible with }\mathcal{A}]\leqslant h(\mathcal{A})^2.$$

Proof. We write $f(\theta,\eta)$ for the value of f, where θ is the configuration inside $\bigcup\mathcal{A}$ and η is the configuration outside. For each θ, let F_θ be the function of η defined by $F_\theta(\eta):=f(\theta,\eta)$. If θ is such that there is an annulus $A\in\mathcal{A}$ without the 4-arm event, then the connectivity of points outside the outer boundary of A does not depend on the configuration inside the inner disk of A, and therefore F_θ does not depend on any of the variables in the inner disk of A. Thus, if a subset of variables S is disjoint from $\bigcup\mathcal{A}$ (so that we can talk about $\hat{f}_\theta(S)$) and intersects this inner disk, then the corresponding Fourier coefficient vanishes: $\hat{f}_\theta(S)=\mathbb{E}[F_\theta\chi_S]=0$. Therefore, if we let W be the linear space of functions spanned by $\{\chi_S : S\text{ is compatible with }\mathcal{A}\}$, and P_W denotes the orthogonal projection onto W, then $P_WF_\theta\neq0$ implies the 4-arm event of θ in every $A\in\mathcal{A}$.

Now, observe that $P_W f = P_W \mathbb{E}[f\,|\,\eta]$, because for all $g \in W$ we have

$$\mathbb{E}[\mathbb{E}[f\,|\,\eta]g] = \mathbb{E}[\mathbb{E}[fg\,|\,\eta]] = \mathbb{E}[fg].$$

Next, by the independence of η and θ from each other, we have $\mathbb{E}[f\,|\,\eta] = \mathbb{E}^\theta[F_\theta]$, where the right-hand side is an expectation with respect to θ, and hence is still a function of η. Thus, $P_W f = P_W \mathbb{E}^\theta[F_\theta] = \mathbb{E}^\theta[P_W F_\theta]$. Consequently,

$$\mathbb{P}[\mathscr{S} \text{ is compatible with } \mathcal{A}] = \|P_W f\|^2 = \|\mathbb{E}^\theta[P_W F_\theta]\|^2 \leqslant \mathbb{E}^\theta[\|P_W F_\theta\|]^2,$$

where the last step follows from the triangle inequality for $\|\cdot\|$. For every θ, the function F_θ is bounded from above by 1 in absolute value. Therefore, $\|F_\theta\| \leqslant 1$ for every θ. This implies that $\|P_W F_\theta\| \leqslant 1$ for every θ (the norm is the L^2 norm and P_W is an orthogonal projection). Hence, we get

$$\mathbb{P}[\mathscr{S} \text{ is compatible with } \mathcal{A}] \leqslant \mathbb{P}^\theta\big[P_W F_\theta \neq 0\big]^2 \leqslant h(\mathcal{A})^2.$$

This proves the lemma. □

A trivial but important instance of (4.2) is when $\mathfrak{A} = \{\varnothing\}$: the empty annulus structure $\varnothing$ is compatible with any S, while $h(\varnothing)^2 = 1$.

Now, for each set $S \in \mathcal{S}(r,k)$ we construct a compatible annulus structure $\mathcal{A}(S)$, such that the set $\mathfrak{A} = \mathfrak{A}(r,k) = \{\mathcal{A}(S) : S \in \mathcal{S}(r,k)\}$ will be small and effective enough for (4.2) to imply Proposition 4.2. The main idea for this construction is that the spectral sample tends to be clustered together, which can already be foreseen in Lemma 4.3: each additional thick annulus (corresponding to some part of the spectral sample far from all other parts) decreases the weight $h(\mathcal{A})^2$ by quite a lot—more than what can be balanced by the number of essentially different ways that this can happen. (We will make this vague description of clustering more precise after the end of the proof, in Remark 4.5.)

We now prepare to define the annulus structure $\mathcal{A}(S)$ corresponding to an $S \in \mathcal{S}(r,k)$, based on the geometry of S. For this definition, we need to first define what we call *clusters* of S. Set $V = V_S := (S \cap U)_r$, the set of squares in the $r\mathbb{Z}^2$ lattice which meet $S \cap U$. For $j \in \mathbb{N}_+$ let G_j be the graph on V, where two squares are joined if their Euclidean distance is at most $2^j r$, and let G_0 be the graph on V with no edges. If $j \in \mathbb{N}_+$ and $C \subseteq V$ is a connected component of G_j, but is not connected in G_{j-1}, then C is called a level-j cluster of S. The level-0 clusters are the connected components of G_0, that is, the singletons contained in V. The level of a cluster C is denoted by $j(C) = j_S(C)$. The Euclidean diameter of a cluster C is clearly at most $2^{j(C)+2} r|C|$.

We will now associate with each cluster C of S an "inner" and an "outer bounding square", whose difference will be the annulus of C. We will have to make sure that

the annuli we are constructing are all disjoint from the set S, hence the inner bounding square should be large enough to contain the points of C, while the outer square should be small enough not to intersect other clusters. However, we cannot simply take the smallest and largest such squares for each C, because we do not want to get too many different annulus structures, i.e., the bounding squares should not depend too sensitively on C. One consequence of this crudeness is that some of the clusters may have an empty annulus associated with them: this will happen when the other clusters are too close.

Let C be a cluster in S at some level $j=j(C)$. We choose a point of the plane, $z\in[C]$, in an arbitrary but fixed way (say the lowest among the leftmost points of $[C]$), where $[C]$ denotes the set of points of the plane covered by the r-squares in C. Let z' be a point with both coordinates divisible by $2^j r$, which is closest to z, with ties broken in some arbitrary but fixed manner. Define the *inner bounding square* $B(C)$ as the square with edge-length $|C|2^{j+4}r$ centered at z' (with edges parallel to the coordinate axes). Note that $[C]\subseteq B(C)$ and the distance from $[C]$ to $\partial B(C)$ is at least $2^{j+3}|C|r-2^j r-2^{j+2}|C|r\geqslant 2^j r$.

If $C\neq V$ is a cluster, then the smallest cluster in V that properly contains C will be called the *parent* of C and denoted by C^p. In this case, let the *outer bounding square* $\overline{B}(C)=\overline{B}_S(C)$ of C be the square concentric with $B(C)$ having side length $(2^{j(C^p)-4}r)\wedge\left(\frac{1}{4}aR\right)$. For the case $C=V$, we let $\overline{B}(V)$ be the square concentric with $B(V)$ having side length $\frac{1}{4}aR$.

It is not necessarily the case that $\overline{B}(C)\supset B(C)$, nor even that $\overline{B}(C)\supset[C]$. Also, it may happen that for two disjoint clusters C and C' we have $B(C)\cap B(C')\neq\varnothing$. However, we do have the following important properties of these squares, which are easy to verify:

(1) if $C'\subsetneq C$ is a subcluster of C, then $\overline{B}(C')\subseteq B(C)$;
(2) if C and C' are disjoint clusters, then $\overline{B}(C)\cap\overline{B}(C')=\varnothing$;
(3) $B(C)$ depends only on C;
(4) $\overline{B}(C)$ depends only on $B(C)$ and $j(C^p)$.

Define $A(C)=A_S(C):=\overline{B}(C)\setminus B(C)$. Note that if $A(C)\neq\varnothing$, then $|C|2^{j(C)+4}r<\frac{1}{4}aR$; therefore, using that $C\subseteq U'_r$, that the distance between z and z' is at most $2^j r$, that $\overline{B}(C)$ has side length at most $\frac{1}{4}aR$, and that the distance of U' from ∂U is at least aR, we get $A(C)\subseteq U$. This means that the annulus $A(C)=A_S(C)$ we have constructed is disjoint from S, both inside and outside U.

Define an annulus structure $\mathcal{A}_1(S)$ associated with S by

$$\mathcal{A}_1(S):=\{A_S(C):C\text{ is a cluster in }S\text{ and }A_S(C)\neq\varnothing\}.$$

By properties (1) and (2) above, the different annuli in $\mathcal{A}_1(S)$ are disjoint. See Figure 4.1. It is also clear that $\mathcal{A}_1(S)$ is compatible with S. However, we still need to modify $\mathcal{A}_1(S)$ for it to be useful.

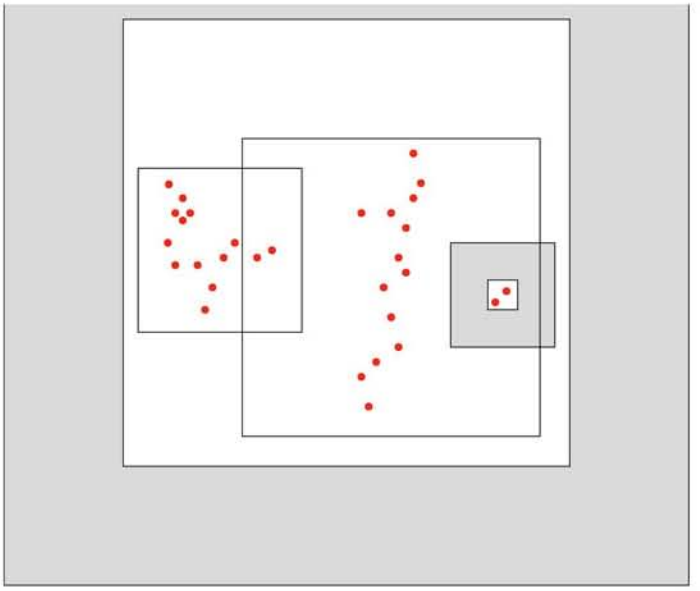


Figure 4.1. A cluster with three child clusters, two of which have empty annuli; only the inner bounding squares are shown for those two.

It follows from (2.6) that the function $\gamma_r(2^j r)$ is decaying exponentially, in the sense that there are absolute constants $c_0, c_1 > 0$ such that

$$\gamma_r(2^{j'} r) \leqslant c_0 2^{-c_1(j'-j)} \gamma_r(2^j r) \quad \text{if } j' > j \geqslant 0. \tag{4.3}$$

It would be rather convenient in the proof below to have $\gamma_r(2^{j+1} r) \leqslant \gamma_r(2^j r)$. However, we do not want to try to prove this. Instead, we use the function $\bar{\gamma}_r(\varrho) := \inf_{\varrho' \in [r,\varrho]} \gamma_r(\varrho')$ in place of γ. Clearly, $\bar{\gamma}$ also satisfies (4.3) and $\bar{\gamma}_r(\varrho) \leqslant \gamma_r(\varrho) \leqslant O(1)\bar{\gamma}_r(\varrho)$.

A cluster C will be called *overcrowded* if

$$g(|C|)\bar{\gamma}_r(2^{j(C)} r) > 1,$$

with the constant $\vartheta > 0$ in the definition of $g(k)$ to be determined later. In particular, clusters at level 0 are overcrowded. For each overcrowded cluster, we remove from $\mathcal{A}_1(S)$ all the annuli corresponding to its proper subclusters. The resulting annulus structure will be denoted $\mathcal{A}(S)$, and is still compatible with S. Finally, we set

$$\mathfrak{A} = \mathfrak{A}(r, k, U, U') := \{\mathcal{A}(S) : S \in \mathcal{S}(r, k)\}.$$

We will show that, for some constant $c_0 > 0$,

$$\sum_{\mathcal{A} \in \mathfrak{A}} h(\mathcal{A})^2 \leqslant O(1) \left(\frac{k}{a}\right)^{c_0} g(k) \bar{\gamma}_r(R). \tag{4.4}$$

(Note that there are usually a lot of sets S with the same $\mathcal{A}(S)$ in $\mathfrak{A}$, or even with the same V_S. Indeed, a main point of our construction of $\mathfrak{A}$ is to use as few annulus structures as possible. In particular, the above sum is really over different annulus structures, without multiplicities coming from the different sets S.) This and (4.2) together imply Proposition 4.2, with possibly a different choice for the constant ϑ.

For each S, there is a natural tree structure on the clusters that correspond to the annuli of $\mathcal{A}(S)$. The root is V itself, and the parent C^p of each cluster $C \neq V$ is also its parent in the tree. The leaves are the overcrowded clusters. We will use this tree structure to build the bound (4.4) inductively: we will write each term $h(\mathcal{A})^2$ as a product of weights corresponding to smaller annulus structures, with the trivial annulus structures of overcrowded clusters as the base step. (The effect of doing this induction on the notation is that the letter k (or k_i) will be used not only for the size of V_S, but of subclusters, too.) The reason for introducing the annulus structures $\mathcal{A}(S)$ instead of $\mathcal{A}_1(S)$ is that inside overcrowded clusters, the factors we would get from extra annuli would not balance the number of possible ways they can be added, so it seems better to use no annuli for them at all.

For a cluster C of S let $\mathcal{A}'(C)$ denote the subset of $\mathcal{A}(S)$ corresponding to the proper subclusters of C. Note that $\mathcal{A}'(C)$ depends only on C; that is, it is not affected by a modification of S, as long as C remains a cluster of S and it does not acquire or lose an overcrowded ancestor. Also note that $\mathcal{A}'(V) = \mathcal{A}(S) \setminus \{A_S(V)\}$, where $V = V_S$.

Fix some $j, k \in \mathbb{N}_+$. If $B = B(C)$, where $|C| = k$ and $j(C) = j$, then the coordinates of the four corners of B are divisible by $2^j r$, and its side length is $k2^{j+4}r$. For such B (which might correspond to more that one pair of k and j), define

$$\mathfrak{A}'(B,k,j) := \{\mathcal{A}'(V_S) : S \in \mathcal{S}(r,k),\ B(V_S) = B \text{ and } j(V_S) = j\},$$

and

$$H(j,k) := \sup_B \sum_{\mathcal{A} \in \mathfrak{A}'(B,k,j)} h(\mathcal{A})^2.$$

(Note that the sum on the right may depend on B, since it may happen, for example, that $B \not\subseteq U'$.)

Define $J(k) := \max\{j \in \mathbb{N} : g(k)\bar{\gamma}_r(2^j r) > 1\} = O(\vartheta)\log_2^2(k+2)$. If $j \leqslant J(k)$, then every $S \in \mathcal{S}(r,k)$ with $j(V_S) = j$ and $B(V_S) = B$ has V_S overcrowded, and hence $\mathcal{A}'(V_S) = \varnothing$. Therefore, if there exist such sets S, then we have $\mathfrak{A}'(B,k,j) = \{\varnothing\}$ and thus $H(j,k) = 1$; otherwise, we have $\mathfrak{A}'(B,k,j) = \varnothing$ and thus $H(j,k) = 0$. That is,

$$H(j,k) \leqslant 1 \quad \text{for all } j \leqslant J(k). \tag{4.5}$$

This is the bound we will use for the overcrowded clusters at the base steps of the induction.

For general j and k, we will show by induction on j (for all $j\in\mathbb{N}$) that

$$H(j,k)\leqslant(g(k)\bar\gamma_r(2^jr))^{1.99}\quad\text{for all } k\geqslant 1. \tag{4.6}$$

Before proving (4.6), let us demonstrate that it implies (4.4). If $j(V_S)=j$ (and $S\in\mathcal{S}(r,k)$), then the number of possible choices for $B(V_S)$ is at most $(2R/2^jr)^2$. If $A(V_S)\neq\varnothing$, then the probability of the 4-arm event in $A(V_S)$ is at most $O(1)\alpha_4\big(k2^jr,aR\big)$, by (2.3). If $A(V_S)=\varnothing$, then the probability of the 4-arm event is 1, while $k2^{j+4}r\geqslant\frac14 aR$, hence the previous bound is $O(1)$, and thus remains true. Therefore,

$$\sum_{\mathcal{A}\in\mathfrak{A}}h(\mathcal{A})^2\leqslant O(1)\sum_{j=0}^{\lceil\log_2(2R/r)\rceil}\left(\frac{R}{2^jr}\right)^2\alpha_4(k2^jr,aR)^2H(j,k).$$

Now, we use the quasi-multiplicativity of the 4-arm event, (2.3), (4.5) and (4.6) to rewrite this as

$$\begin{aligned}\sum_{\mathcal{A}\in\mathfrak{A}}h(\mathcal{A})^2&\leqslant\sum_{j=0}^{\lceil\log_2(2R/r)\rceil}O(1)\left(\frac{k}{a}\right)^{c_0}\frac{\bar\gamma_r(R)}{\bar\gamma_r(2^jr)}H(j,k)\\&\leqslant O(1)\left(\frac{k}{a}\right)^{c_0}\bar\gamma_r(R)\Bigg(\sum_{j\leqslant J(k)}\frac{1}{\bar\gamma_r(2^jr)}+g(k)^{1.99}\sum_{j>J(k)}\bar\gamma_r(2^jr)^{0.99}\Bigg)\end{aligned} \tag{4.7}$$

for some finite constant c_0. Because (4.3) holds for $\bar\gamma$, the first sum is dominated by a constant times the summand corresponding to $j=J(k)$ and the second sum is dominated by a constant times the summand corresponding to $j=J(k)+1$. Since $\bar\gamma_r(2^{J(k)}r)>1/g(k)$ and $\bar\gamma_r(2^{J(k)+1}r)\leqslant1/g(k)$, we get the bound claimed in (4.4).

In order to complete the proof of Proposition 4.2, it remains to establish (4.6). The proof of this inequality will be inductive and somewhat similar to the proof of (4.4) from (4.6), but there are some important differences. In passing from (4.6) to (4.4), we "used up" some of the exponent 1.99—it has become 1. Such a loss would not be sustainable if it had to be repeated in every inductive step. What saves us in the following proof of (4.6) is that clusters that are not singletons have more than one child cluster. (In other words, any non-leaf vertex of the cluster tree has at least two children.) This results in almost squaring the estimate at each inductive step, which makes the proof work.

We now proceed to prove (4.6) by induction. Since $\bar\gamma$ is non-increasing, the claim holds for all $j\leqslant J(k)$, because of (4.5). (In particular, for $j=0$ and any $k\geqslant1$.) We now fix some j and k with $j>J(k)$, and assume that (4.6) holds for all smaller values of j. Fix some square B such that $\mathfrak{A}'(B,k,j)\neq\varnothing$. Observe that if $B(V_S)=B$, then S has some $d=d(S)\geqslant2$ children in its tree of clusters (we know that $d\neq0$ because V_S

is not overcrowded, since $j>J(k)$). Fix some $d\in\{2,3,...\}$, $k_1,...,k_d$, $j_1,...,j_d$ and subsquares $B_1,...,B_d$ such that there is some $S\in\mathcal{S}(r,k)$ with $j(V_S)=j$ and $B(V_S)=B$, and having cluster children $C^1,...,C^d$ with $|C^i|=k_i$, $j(C^i)=j_i$ and $B(C^i)=B_i$. Note that $A_i:=A_S(C^i)$ does not depend on the choice of S satisfying the above (by property (4) of squares noted previously and by having a fixed j).

Let

$$\mathfrak{A}''=\mathfrak{A}''(d,k_1,...,k_d,j_1,...,j_d,B_1,...,B_d)$$

denote the set of all elements of $\mathfrak{A}'(B,k,j)$ that arise from such an S. Note that every $\mathcal{A}\in\mathfrak{A}''$ is of the form $\{A_1,...,A_d\}\cup\bigcup_{i=1}^d\mathcal{A}_i$, where the $\mathcal{A}_i\in\mathfrak{A}'(B_i,k_i,j_i)$ are d disjoint annulus substructures, and the A_i's are fixed by $\mathfrak{A}''$. (It is possible that some of the A_i's here are empty annuli, see, e.g., Figure 4.1.) For such an $\mathcal{A}$, we have, by (4.1),

$$h(\mathcal{A})=\prod_{i=1}^d h(A_i)h(\mathcal{A}_i).$$

Hence,

$$\begin{aligned}\sum_{\mathcal{A}\in\mathfrak{A}''}h(\mathcal{A})^2&\leqslant\sum_{\mathcal{A}_1\in\mathfrak{A}'(B_1,k_1,j_1)}\sum_{\mathcal{A}_2\in\mathfrak{A}'(B_2,k_2,j_2)}...\sum_{\mathcal{A}_d\in\mathfrak{A}'(B_d,k_d,j_d)}\prod_{i=1}^d(h(A_i)h(\mathcal{A}_i))^2\\&=\prod_{i=1}^d h(A_i)^2\sum_{\mathcal{A}\in\mathfrak{A}'(B_i,k_i,j_i)}h(\mathcal{A})^2\\&\leqslant\prod_{i=1}^d h(A_i)^2H(j_i,k_i).\end{aligned}$$

Now, $h(A_i)=O(1)\alpha_4(k_i2^{j_i}r,2^jr)$, where we use the convention that $\alpha_4(\varrho,\varrho')=1$ if $\varrho\geqslant\varrho'$. Furthermore, given B, d, $j_1,...,j_d$ and $k_1,...,k_d$, there are no more than $O(1)(k2^{j-j_i})^2$ possible choices for B_i. Hence, summing the above over all such choices, we get

$$H(j,k)\leqslant\sum_{d=2}^k\sum_{\substack{(j_1,...,j_d)\\(k_1,...,k_d)}}\prod_{i=1}^d O(k2^{j-j_i})^2\alpha_4(k_i2^{j_i}r,2^jr)^2H(j_i,k_i).$$

The inductive hypothesis (4.6) for each of the pairs (j_i,k_i) implies $H(j_i,k_i)\leqslant\bar{\gamma}_r(2^{j_i}r)g(k_i)$ when $j_i>J(k_i)$, since we decreased the exponent from 1.99 to 1, with a base less than 1. This upper bound also holds when $j_i\leqslant J(k_i)$, because of (4.5). We now use this, together

with the quasi-multiplicativity and the RSW estimate (2.3), as before, to obtain

$$\begin{aligned} H(j,k) &\leqslant \sum_{d=2}^{k} \sum_{\substack{(j_1,\dots,j_d)\\(k_1,\dots,k_d)}} \prod_{i=1}^{d} O(k)^2 k_i^{O(1)} \frac{\bar\gamma_r(2^j r)}{\bar\gamma_r(2^{j_i} r)} \bar\gamma_r(2^{j_i} r) g(k_i) \\ &\leqslant \sum_{d=2}^{k} \sum_{\substack{(j_1,\dots,j_d)\\(k_1,\dots,k_d)}} \prod_{i=1}^{d} O(k)^{O(1)} \bar\gamma_r(2^j r) g(k_i) \qquad (4.8) \\ &\leqslant \sum_{d=2}^{k} (O(k)^{O(1)} j \bar\gamma_r(2^j r))^d \sum_{(k_1,\dots,k_d)} \prod_{i=1}^{d} g(k_i). \end{aligned}$$

Since $\log_2^2(x+2)$ is concave on $[0,\infty)$, it follows that when $k_1+\dots+k_d=k$, we have $\prod_{i=1}^d g(k_i) \leqslant g(k/d)^d$. Since for a fixed d the number of possible d-tuples $(k_1,\dots,k_d)$ is clearly bounded by k^d, the above gives

$$H(j,k) \leqslant \sum_{d=2}^{k} (c k^c j \bar\gamma_r(2^j r) g(k/d))^d$$

for some constant c. Noting that $g(k/d) \leqslant g(k/2)$, and setting

$$\Phi = \Phi(j,k) := c k^c j \bar\gamma_r(2^j r) g(k/2),$$

we then get $H(j,k) \leqslant \Phi^2+\Phi^3+\Phi^4+\dots = \Phi^2/(1-\Phi)$, provided $\Phi<1$. Consequently, the proof of (4.6) and of Proposition 4.2 are completed by the following lemma. □

LEMMA 4.4. *For all $\varepsilon \in (0,\frac12)$, if ϑ in the definition of g is chosen sufficently large (depending only on ε), then for all $k \in \mathbb{N}_+$ and all $j > J(k)$, we have*

$$\Phi(j,k) < \frac{(g(k)\bar\gamma_r(2^j r))^{1-\varepsilon}}{2} \qquad (4.9)$$

and $\Phi(j,k) < \frac12$.

Proof. The estimate $\Phi<\frac12$ follows from (4.9), since $g(k)\bar\gamma_r(2^j r) \leqslant 1$. Write

$$\frac{\Phi}{(g(k)\bar\gamma_r(2^j r))^{1-\varepsilon}} = c k^c \bar\gamma_r(2^j r)^{\varepsilon} j g(k/2)^{\varepsilon} \left(\frac{g(k/2)}{g(k)}\right)^{1-\varepsilon}. \qquad (4.10)$$

Since (4.3) holds for $\bar\gamma$ and $\bar\gamma_r(2^{J(k)+1} r) g(k) \leqslant 1$, we have

$$\bar\gamma_r(2^j r)^{\varepsilon} g(k/2)^{\varepsilon} \leqslant \bar\gamma_r(2^j r)^{\varepsilon} g(k)^{\varepsilon} \leqslant O(1) 2^{-\varepsilon c_1 (j - J(k))}.$$

Hence, $j\bar{\gamma}_r(2^j r)^\varepsilon g(k/2)^\varepsilon \leqslant O_\varepsilon(J(k)+1)$. As noted before, we have $J(k)=O(\vartheta)\log_2^2(k+2)$. Thus, (4.10) gives

$$\frac{\Phi}{(g(k)\bar{\gamma}_r(2^j r))^{1-\varepsilon}} \leqslant O_\varepsilon(\vartheta)\log_2^2(k+2)k^c\left(\frac{g(k/2)}{g(k)}\right)^{1/2}.$$

It is easy to verify that the right-hand side tends to 0 as $\vartheta\to\infty$, uniformly in $k\in\mathbb{N}_+$. This completes the proof. □

For later use, let us point out that having an exponent larger than 1 in (4.6) was important when we derived the bound (4.4), but not for the induction. In (4.8), we used only the bound (4.6) with the exponent 1. This was sufficient because $d\geqslant 2$.

Remark 4.5. Our proof shows a *clustering effect* for spectral samples of very small size. Firstly, a positive proportion of our main upper bound (4.4) comes from annulus structures with a single overcrowded cluster, as is clear from (4.7). Moreover, the contribution from sets with large diameter is small: for any $d\leqslant R/r$, the calculation in (4.7) implies immediately that if $k\leqslant O(1)\log_2 d$, then

$$\mathbb{P}[\mathscr{S}\in\mathcal{S}(r,k),\operatorname{diam}(\mathscr{S})>rd]\leqslant\gamma_r(R)\gamma_r(rd)^{1+o(1)}d^{o(1)}$$

(of course, the exponent 0.99 in (4.7) can be modified to $1+o(1)$). Recall that (4.3) says that $\gamma_r(rd)<O(1)d^{-c_1}$ for some $c_1>0$. Since we will see in §5 that Lemma 3.2 is sharp, and will handle the boundary issues in §4.3, we easily obtain that

$$\mathbb{P}[1\leqslant|\mathscr{S}_r|\leqslant k]\geqslant\mathbb{P}[|\mathscr{S}_r|=1]\geqslant O(1)\gamma_r(R).$$

Therefore the above bound gives for $k\leqslant O(1)\log_2 d$ that

$$\mathbb{P}\big[\operatorname{diam}(\mathscr{S})>rd\,\big|\,1\leqslant|\mathscr{S}_r|\leqslant k\big]\leqslant\gamma_r(rd)^{1+o(1)}.$$

To illustrate this formula (with $r=1$), if one is looking for spectral samples of size less than $\log R$, then they have small diameter:

$$\mathbb{P}\big[\operatorname{diam}(\mathscr{S})>R^\alpha\,\big|\,1\leqslant|\mathscr{S}|\leqslant\log R\big]\leqslant\gamma(R^\alpha)^{1+o(1)},$$

which for the triangular lattice gives $R^{-\alpha/2+o(1)}$.

Remark 4.6. Our strategy proving Proposition 4.1 also works for pivotals, showing that

$$\mathbb{P}[|\mathscr{P}_r|=k]\leqslant g(k)\alpha_6(r,R)\left(\frac{R}{r}\right)^2.$$

The only difference is that we need to replace the factor $\alpha_4(r,R)^2$, coming from Lemma 3.2 and its generalization Lemma 4.3, with $\alpha_6(r,R)$. The reason for this factor is that having pivotals in the inner disk of an annulus but no pivotals in the annulus itself corresponds to the 6-arm event in the annulus. On $\mathbb{Z}^2$ it is known that

$$\alpha_6(r,R)\Big(\frac{R}{r}\Big)^2 \leqslant O(1)\Big(\frac{r}{R}\Big)^{\varepsilon}$$

for some $\varepsilon>0$ [SSt, Corollary A.8]; on the triangular lattice,

$$\alpha_6(r,R)\Big(\frac{R}{r}\Big)^2 = \Big(\frac{r}{R}\Big)^{11/12+o(1)}$$

[SW]. Thus the clustering effect for pivotals is expressed here in the following way:

$$\mathbb{P}\big[\operatorname{diam}(\mathscr{P}) > R^\alpha \;\big|\; 1 \leqslant |\mathscr{P}| \leqslant \log R\big] \leqslant R^{2\alpha}\alpha_6(R^\alpha)^{1+o(1)},$$

which for the triangular lattice gives $R^{-11\alpha/12+o(1)}$.

4.3. Handling boundary issues

Proof of Proposition 4.1. Since the proof is rather similar to that of Proposition 4.2, we will just indicate the necessary modifications.

First of all, we need the *half-plane* and *quarter-plane j-arm events*. So let $\alpha_j^+(r,R)$ be the probability of having j disjoint arms of alternating colors connecting $\partial B(0,r)$ to $\partial B(0,R)$ inside the half-annulus $(B(0,R)\setminus B(0,r))\cap(\mathbb{R}\times\mathbb{R}_+)$. Similarly, $\alpha_j^{++}(r,R)$ is the probability of having j arms of alternating colors connecting $\partial B(0,r)$ to $\partial B(0,R)$ inside the quarter-annulus $(B(0,R)\setminus B(0,r))\cap(\mathbb{R}_+\times\mathbb{R}_+)$. As before, we let $\alpha_j^+(R):=\alpha_j^+(2j,R)$ and similarly for α_j^{++}. The RSW and quasi-multiplicativity bounds (2.1) and (2.2) hold for these quantities, as well.

The reason for these definitions is that if x is a point on one of the sides of $[0,R]^2$ such that its distance from the other sides is at least r', then the percolation configuration inside $B(x,r)$ has an effect on the crossing event only if we have the 3-arm event in the half-annulus $(B(x,r')\setminus B(x,r))\cap[0,R]^2$. Similarly, if x is one of the corners of $[0,R]^2$, and $r'\leqslant R$, then we need the 2-arm event in the quarter-annulus $(B(x,r')\setminus B(x,r))\cap[0,R]^2$ in order for the configuration inside $B(x,r)$ to have any effect.

In the annulus structures we are going to build, we will have to consider clusters that are close to a side or even to a corner of $[0,R]^2$. These will be called side and corner clusters (defined precisely below). To understand the contribution of such clusters, we define

$$\gamma_r^+(\varrho) := \frac{\varrho}{r}\alpha_3^+(r,\varrho)^2 \quad\text{and}\quad \gamma_r^{++}(\varrho) := \alpha_2^{++}(r,\varrho)^2.$$

The function γ_r^+ plays a role similar to γ_r, but in relation to the side clusters. Similarly, γ_r^{++} relates to the corner clusters. The motivation for the linear factor ϱ/r in the definition of γ_r^+ is that (up to a constant factor) the number of different ways to choose a square of some fixed size whose center is on a line segment of length ϱ and whose position on the line segment is divisible by ϱ' is ϱ/ϱ', when $\varrho>\varrho'>0$. Such a factor is not necessary in the case of γ_r^{++}, because it corresponds to a corner, and there is no choice in placing a square of a fixed size into a fixed corner.

As we will see, the key reason for the boundary $\partial[0,R]^2$ to play no significant role in the behavior of the spectral sample is that there is a $\delta>0$ and a constant c such that when $r\leqslant\varrho'\leqslant\varrho$,

$$\frac{\gamma_r^+(\varrho)}{\gamma_r^+(\varrho')}\leqslant c\biggl(\frac{\gamma_r(\varrho)}{\gamma_r(\varrho')}\biggr)^{1+\delta}\quad\text{and}\quad\frac{\gamma_r^{++}(\varrho)}{\gamma_r^{++}(\varrho')}\leqslant c\biggl(\frac{\gamma_r(\varrho)}{\gamma_r(\varrho')}\biggr)^{1+\delta}.\tag{4.11}$$

We now prove these inequalities. Firstly, it is known that

$$\alpha_3^+(\varrho',\varrho)\asymp\biggl(\frac{\varrho'}{\varrho}\biggr)^2,\tag{4.12}$$

see [W, first exercise sheet]. Hence $\gamma_r^+(\varrho)/\gamma_r^+(\varrho')\asymp(\varrho'/\varrho)^3$. Secondly, observe that

$$\alpha_2^{++}(\varrho',\varrho)^2\leqslant\alpha_4^+(\varrho',\varrho)\leqslant\alpha_3^+(\varrho',\varrho)\alpha_1^+(\varrho',\varrho)\leqslant O(1)\alpha_3^+(\varrho',\varrho)\biggl(\frac{\varrho'}{\varrho}\biggr)^{\varepsilon},\tag{4.13}$$

with some $\varepsilon>0$, where the second step used Reimer's inequality [R] (or color-switching and the van den Berg–Kesten (BK) inequality), and the third step used the RSW estimate for the 1-arm half-plane event. Therefore, $\gamma_r^{++}(\varrho)/\gamma_r^{++}(\varrho')\leqslant O(1)(\varrho'/\varrho)^{2+\varepsilon}$. On the other hand, (2.6) implies that

$$\frac{\gamma_r(\varrho)}{\gamma_r(\varrho')}\geqslant\biggl(\frac{\varrho'}{\varrho}\biggr)^{2-\varepsilon'}$$

for some $\varepsilon'>0$. Combining these upper and lower bounds we get (4.11).

Let us note that for the triangular lattice we actually know that

$$\alpha_2^{++}(\varrho',\varrho)=\biggl(\frac{\varrho'}{\varrho}\biggr)^{2+o(1)},\tag{4.14}$$

since $\alpha_2^{++}(\varrho',\varrho)=\alpha_2^+(\varrho',\varrho)^{2+o(1)}$ by the conformal invariance of the scaling limit, while $\alpha_2^+(\varrho',\varrho)\asymp\varrho'/\varrho$ is known for both lattices by RSW arguments, see again [W, first exercise sheet].

After these preparations, we define $\mathcal{S}(r,k)$ as the set of all $S\subseteq\mathcal{I}$ such that $|S_r|=k$. The set $V=V_S$ is defined as S_r. As it turns out, we will need to limit the diameters of

the clusters. For that purpose, set $\bar{\jmath}=\bar{\jmath}_k:=\lfloor \log_2(R/kr)\rfloor -5$ and $J:=\{0,1,...,\bar{\jmath}\}$. Clearly, we may assume without loss of generality that $\bar{\jmath}>0$. The clusters at level 0 are once again the sets of the type $C=\{x\}$, where $x\in V$. If $j\in J$, then an *interior cluster* at level j is defined as a connected component $C\subseteq V$ of G_j such that the distance from C to $\partial([0,R]^2)$ is larger than $2^j r$ and C is not connected in G_{j-1}. The interior clusters at level 0 are just the connected components of G_0; that is, the singletons in V. Inductively, we define the *side clusters*: a connected component $C\subseteq V$ of G_j is a side cluster at level $j\in J$ if it is within distance $2^j r$ of precisely one of the four boundary edges of $[0,R]^2$ and it is not a side cluster at any level $j'\in\{1,2,...,j-1\}$. Likewise, a *corner cluster* at level $j\in J$ is a connected component $C\subseteq V$ of G_j that is within distance $2^j r$ of precisely two adjacent boundary edges of $[0,R]^2$, but is not a corner cluster at any level $j'\in\{1,2,...,j-1\}$. Finally, the unique *top cluster* has level $\bar{\jmath}+1$ (by definition) and consists of all of V. With these choices, when $j\in J$ every connected component of G_j is either an interior, side, or corner cluster, and this is the reason for setting the upper bound $\bar{\jmath}$.

Note that the top cluster contains at least one cluster (which could be a side cluster or a corner cluster or an interior cluster), a corner cluster contains at least one side or interior cluster, and possibly also a corner cluster, a side cluster contains an interior cluster or at least two side clusters (but no corner clusters), and an interior cluster at level $j>0$ contains at least two interior clusters (and no side or corner clusters).

The inner and outer bounding squares associated with the clusters are defined as before, except that the squares associated with side clusters are centered at points on the corresponding edge and squares associated with corner clusters are centered at the corresponding corner, which is the meeting point of the two sides of $[0,R]^2$ closest to the cluster. There are no squares associated with the top cluster. The annulus associated with each cluster (other than the top cluster, which does not have its own annulus) is the annulus between its outer square and its inner square, provided that the outer square strictly contains the inner square.

We define $\bar{\gamma}_r^+(\varrho):=\inf_{\varrho'\in[r,\varrho]}\gamma_r^+(\varrho)$ and similarly for $\bar{\gamma}_r^{++}$. The exponential decay (4.3) holds for these functions as well. As before, clusters (even side or corner clusters) are considered *overcrowded* if they satisfy $g(|C|)\gamma_r(2^j r)>1$, and the annulus structure $\mathcal{A}(S)$ is defined as above.

There are some modifications necessary in the definition of $h(\mathcal{A})$ in order for (4.2) to still hold in the present setting. For a side annulus A define $h(A)$ as the probability of the 3-arm crossing event within $A\cap[0,R]^2$ between the two boundary components of A, and for a corner annulus define $h(A)$ as the probability of the 2-arm crossing event within $A\cap[0,R]^2$ between the two boundary components of A. With these modifications, (4.2) still holds, and the proof is essentially the same.

The definition of $H(j,k)$ is similar to the one in the proof of Proposition 4.2, but now the supremum only refers to interior squares. (In the definition of $\mathfrak{A}'(B,k,j)$, we presently restrict to such S so that V_S is an interior cluster at level j, in addition to being the top cluster, and when we write $B(V_S)$, it is understood as the inner square defined for an interior cluster.) We also define $H^+(j,k)$ and $H^{++}(j,k)$, which refer to the supremum over side and corner squares, respectively. We also set

$$H^{\square}(R,k) := \sum_{S\in\mathcal{S}(r,k)} h(\mathcal{A}(S))^2,$$

and our goal is to show that this quantity is at most $g(k)\gamma_r(R)$.

We first prove a similar bound on $H^+(j,k)$. It is convenient to separate the annulus structures $\mathcal{A}'(S)$ where $B(V_S)=B$ and B is a side square into those where V_S has a single child cluster and into those where V_S has at least two child clusters. In the case of a single child cluster, that child has to be an interior cluster, and using (4.6), the argument giving (4.4) now gives the bound

$$\sum_{\mathcal{A}} h(\mathcal{A})^2 \leqslant O(1)k^{O(1)}g(k)\bar{\gamma}_r(2^j r),$$

where the sum extends over such (single interior cluster child) annulus structures. By increasing the constant ϑ in the definition of g, we may then incorporate the factor $O(1)k^{O(1)}$ into g. The bound on the sum over the annulus structures with at least two child clusters can now be established by induction on j, in almost the same way that (4.6) was proved by induction. The main difference here is that the children can fall into two types, which slightly complicates the calculations but adds no significant difficulties. (Indeed, since each square among B_i at level j_i, $i=1,2,...,d$, can either correspond to a side cluster or an interior cluster, each factor in the first row of (4.8) presently needs to be replaced by

$$O(k)^2 k_i^{O(1)}\left(\frac{\bar{\gamma}_r(2^j r)}{\bar{\gamma}_r(2^{j_i} r)}+\frac{\bar{\gamma}_r^+(2^j r)}{\bar{\gamma}_r^+(2^{j_i} r)}\right)\bar{\gamma}_r(2^{j_i} r)g(k_i).$$

Due to (4.11), the fraction featuring $\bar{\gamma}^+$ is dominated by a constant times the other fraction and is therefore certainly inconsequential.) A point worth noting is that in the inductive proof of (4.6) we only used the inductive hypothesis with exponent 1 in place of 1.99. In our present situation, this is what we have at our disposal in the base step of the induction, since the induction now has to be started from overcrowded clusters or side clusters with a single child cluster.

Summing up the two types, we conclude (by changing ϑ again, if necessary) that $H^+(j,k)\leqslant g(k)\bar{\gamma}_r(2^j r)$. Of course, in this type of argument we should not change ϑ at

every induction step, for then it may end up depending on R. But we have not committed any such offence.

We can show that $H^{++}(j,k)\leqslant g(k)\bar{\gamma}_r(2^j r)$ similarly. We separate the annulus structures into those with a single child at the top level which is an interior cluster, those with a single child which is a side cluster, and those with several children at the top level. The first type is handled as in the bound for $H^{+}(j,k)$. The second type is handled similarly, but now we do not have the exponent 1.99 as in (4.6), but only the exponent 1 that we showed for $H^{+}(j,k)$. So, we use instead the fact that $\delta>0$ in (4.11). The argument bounding the third type uses induction as in the multi-child case of $H^{+}(j,k)$.

Finally, the bound for $H^{\square}(R,k)$ follows in the same way, using our previous bounds for $H^{+}(j,k)$ and $H^{++}(j,k)$ together with (4.11). The last small difference is that the child clusters of the top cluster (at some levels j_i) have outer bounding squares of size $\asymp r2^{\bar{\jmath}}$, but the number of ways to place each of these clusters is $\asymp(R/r2^{j_i})^2$, instead of $\asymp(2^{\bar{\jmath}}/2^{j_i})^2$. But $R/r2^{\bar{\jmath}}\asymp k$, so this discrepancy gives only an $O(k^2)$ factor for each child cluster, which can be absorbed into $g(k)$ in the usual way. This completes the proof. □

4.4. The radial case

In this subsection, we will consider the spectral sample $\mathscr{S}=\mathscr{S}_f$, where f is the indicator function (not the ±1-indicator function) of the ℓ-arm event in the annulus $[-R,R]^2\setminus[-\ell,\ell]^2$ and $\ell=1$ or $\ell\in2\mathbb{N}_+$. Thus, $\mathbb{E}[f]=\mathbb{E}[f^2]\asymp\alpha_\ell(R)$. Instead of the probability measure for $\mathscr{S}$ that we have worked with so far, it will be easier notationally to use the unnormalized measure $\mathbb{Q}[\mathscr{S}=S]:=\hat{f}(S)^2$.

For any $S\subset\mathbb{R}^2$, we let $S^*:=S\cup\{0\}$, and define S_r as before. In particular, $\mathscr{S}_r^*$ is the set of r-squares whose intersection with $\mathscr{S}^*$ is non-empty. We are going to prove the following analogue of Proposition 4.1.

PROPOSITION 4.7. *Let $\ell=1$ or $\ell\in2\mathbb{N}_+$, and let $\mathscr{S}$ denote the spectral sample of the indicator function of the ℓ-arm event in the annulus $[-R,R]^2\setminus[-\ell,\ell]^2$. Then there is some $\vartheta^*=\vartheta^*_\ell>0$ such that*

$$\mathbb{Q}[|\mathscr{S}_r|=k]\leqslant g^*(k)\alpha_\ell(r,R)^2\alpha_\ell(r)$$

holds with $g^*(k):=2^{\vartheta^*\log_2^2(k+2)}$ for all $k\in\mathbb{N}_+$ and all $R\geqslant r\geqslant\ell$.

It might be surprising at first glance that no 4-arm probabilities show up in this upper bound. See (4.15) below for a rough explanation.

Proof. The main difference from the square crossing case is that we will use *centered annulus structures*, which have two kinds of annuli: annuli centered at the origin (0),

for which we are interested in the ℓ-arm event, and annuli with outer square disjoint from 0, for which we are interested in the 4-arm event. Each centered annulus structure is required to have an annulus centered at 0 whose inner square does not contain any other annuli. The *inner radius* of the annulus structure is defined as the inner radius of this innermost centered annulus. For a centered annulus structure $\mathcal{A}$, we define $h^*(\mathcal{A})$ to be the probability of having the 4-arm event in the annuli with outer square disjoint from 0 and the ℓ-arm event in the annuli centered at 0. Now, we have the following analogue of Lemma 4.3.

LEMMA 4.8. *For any centered annulus structure $\mathcal{A}$ with inner radius $r_{\mathcal{A}}$,*

$$\mathbb{Q}[\mathscr{S}^* \text{ is compatible with } \mathcal{A}] \leqslant \alpha_\ell(r_{\mathcal{A}}) h^*(\mathcal{A})^2.$$

Proof. Similarly to the proof of Lemma 4.3, we divide the set of relevant bits into parts: θ is the configuration inside $\bigcup \mathcal{A}$, while η_0 is the configuration inside the inner disk of the smallest centered annulus, and η_1 is the configuration neither in θ nor in η_0. As before, F_θ is the function defined by $F_\theta(\eta_0,\eta_1) := f(\theta,\eta_0,\eta_1)$; furthermore, W is the linear space of functions spanned by $\{\chi_S : S^* \text{ compatible with } \mathcal{A}\}$, and P_W denotes the orthogonal projection onto W. Now, $P_W F_\theta \neq 0$ implies the 4-arm event in every interior non-centered $A \in \mathcal{A}$, the ℓ-arm event in every centered $A \in \mathcal{A}$, and the 3-arm event in every boundary (or corner) annulus. (Note that this uses the fact that when $\ell \neq 1$ we are considering the alternating arms event. In particular, we are restricted to $\ell \in \{1\} \cup 2\mathbb{N}_+$.) Moreover, for any θ, we have $F_\theta(\eta_0,\eta_1) = 0$ if η_0 does not have the ℓ-arm event. Thus, $\|P_W F_\theta\|^2 \leqslant \|F_\theta\|^2 = \mathbb{E}[F_\theta^2] \leqslant \alpha_\ell(r_{\mathcal{A}})$. Altogether, similarly to the proof of Lemma 4.3,

$$\begin{aligned}\mathbb{Q}[\mathscr{S}^* \text{ is compatible with } \mathcal{A}] &= \|P_W f\|^2 \leqslant \mathbb{E}^\theta[\|P_W F_\theta\|]^2 \\ &\leqslant \mathbb{P}^\theta[P_W F_\theta \neq 0]^2 \alpha_\ell(r_{\mathcal{A}}) \leqslant h^*(\mathcal{A})^2 \alpha_\ell(r_{\mathcal{A}}),\end{aligned}$$

which proves the lemma. □

Note that a centered annulus structure compatible with $\mathscr{S}$ is also compatible with $\mathscr{S}^*$, but not necessarily vice versa, hence the lemma is stronger with $\mathscr{S}^*$ than it would be with $\mathscr{S}$. This strengthening is crucial, as shown by the following example: for an r-square B at distance $t \in (r,R)$ from 0, by the remark after Lemma 2.2, we have $\mathbb{Q}[\varnothing \neq \mathscr{S} \subseteq B] \leqslant O(1)\mathbb{P}[\Lambda_B]^2 \asymp (\alpha_\ell(1,R)\alpha_4(r,t))^2$, a bound that we would not be able to reproduce from the weaker (starless) version of the above lemma. On the other hand, when B is centered at 0, then the bound $O(1)\alpha_\ell(r)\alpha_\ell(r,R)^2$ given by Lemma 4.8 is stronger than the $O(1)\alpha_\ell(r,R)^2$ bound of Lemma 2.2. (Unlike Lemma 4.3, which was only a generalization of Lemmas 2.2 and 3.2.)

These bounds provide a back-of-the-envelope explanation for how the result of Proposition 4.7 arises, at least for $k=1$:

$$\begin{aligned}\mathbb{Q}[|\mathscr{S}_r|=1] &\leqslant O(1)\alpha_\ell(r)\alpha_\ell(r,R)^2+O(1)\sum_{s=1}^{O(R/r)} s(\alpha_\ell(1,R)\alpha_4(r,sr))^2 \\ &\leqslant O(1)\alpha_\ell(r)\alpha_\ell(r,R)^2+O(1)\alpha_\ell(1,R)^2,\end{aligned} \tag{4.15}$$

where we used that $s\alpha_4(r,sr)^2\leqslant O(1)s^{-1-\varepsilon}$, by (2.6). The first term being dominant also shows that a small $\mathscr{S}_r$ should typically be close to 0. (Which is another manifestation of the 4-arm events playing a small role here.)

Back to the actual proof, analogously to §4.2, for each $S\subset[-R,R]^2$ with $|S_r|=k$ we will build a centered annulus structure $\mathcal{A}(S)$ that is compatible with S^* (but not necessarily with S itself!) and that has $r_{\mathcal{A}(S)}\geqslant r$. Furthermore, the collection $\mathfrak{A}^*(r,k)$ of all these centered annulus structures will be small enough to have

$$\sum_{\mathcal{A}\in\mathfrak{A}^*(r,k)} h^*(\mathcal{A})^2 \leqslant g^*(k)\alpha_\ell(r,R)^2. \tag{4.16}$$

The combination of (4.16) and Lemma 4.8 proves Proposition 4.7.

To construct the annulus structure $\mathcal{A}(S)$, we take $V=V_S$ to be S_r^*, set

$$\bar{\jmath}:=\left\lfloor \log_2\frac{R}{kr}\right\rfloor -5,$$

and define the clusters exactly as in §4.3. A cluster is called *centered* if it contains 0. In constructing the inner and outer bounding squares, we use the additional rule that for centered clusters C the inner bounding square must be centered at 0. (Note that this is just a special case of the "arbitrary but fixed way" of choosing a vertex $z\in[C]$.)

The centered analogue of $\gamma_r(\varrho)$ is now $\gamma_r^*(\varrho):=\alpha_\ell(r,\varrho)^2$. We again let

$$\bar{\gamma}_r^*(\varrho):=\inf_{\varrho'\in[r,\varrho]}\gamma_r^*(\varrho'),$$

and we note the exponential decay (4.3) for $\bar{\gamma}_r^*(r2^j)$ in j. A non-centered cluster C is called *overcrowded* if $g(|C|)\bar{\gamma}_r(r2^{j(C)})>1$, with the function $g(k)$ of Proposition 4.1. A centered cluster is *overcrowded* if $g^*(|C|)\bar{\gamma}_r^*(r2^{j(C)})>1$. We define $J(k)$ as before, using $g(k)$, and similarly define $J^*(k)$, using $g^*(k)$. In particular, $\bar{\gamma}_r^*(r2^{J^*(k)})g^*(k)\asymp 1$.

Using these notions of overcrowdedness, we get our annulus structure $\mathcal{A}(S)$. The collection of them for all $S\subseteq\mathcal{I}$ with $|S_r|=k$ is $\mathfrak{A}^*(r,k)$.

We define $H(j,k)$ similarly as before, but now only for interior inner bounding squares that do not contain 0. We similarly define the quantities $H^+(j,k)$ and $H^{++}(j,k)$

for side and corner squares not containing 0. Finally, we let $H^*(j,k)$ be the analogous quantity where the inner bounding square is required to be centered. We will show that there is some constant $\delta>0$, depending only on ℓ, such that, for $j\in\{J^*(k),\dots,\bar{\jmath}\}$,

$$H^*(j,k)\leqslant(g^*(k)\bar\gamma_r^*(r2^j))^{1+\delta}. \tag{4.17}$$

This implies (4.16) (with a possibly larger constant ϑ^*) in exactly the same way as in §4.2 the bound (4.6) implied (4.4) (with the small additional care regarding the cutoff $\bar{\jmath}$ that we have seen in §4.3).

As usual, we prove (4.17) by induction on j. We may assume that $j>J^*(k)$. Recall that $H^*(j,k)$ is defined as a sum, where each summand corresponds to a centered annulus structure. Suppose that $\mathcal{A}$ is a centered annulus structure contributing to the sum, where $\mathcal{A}=\mathcal{A}(S)$ for some $S\subseteq\mathcal{I}$ with $|S_r|=k$. Let j_* be $j(C^*)$, where C^* is the largest proper centered subcluster of S_r^*, and let $k_*=|C^*\cap S_r|$. Every such $\mathcal{A}$ can be formed as a union of the topmost (centered) annulus in $\mathcal{A}$, a centered annulus structure $\mathcal{A}^*$ for (j_*,k_*), and the annulus structure $\mathcal{A}'$ formed by dropping all the centered annuli from $\mathcal{A}$. Moreover,

$$h^*(\mathcal{A})^2\asymp(k+2)^{O(1)}h^*(\mathcal{A}^*)^2h(\mathcal{A}')^2\alpha_\ell(r2^{j_*},r2^j)^2.$$

The sum over such $\mathcal{A}$ with j_* and k_* fixed is bounded by

$$(k+2)^{O(1)}\alpha_\ell(r2^{j_*},r2^j)^2\biggl(\sum_{\mathcal{A}^*}h^*(\mathcal{A}^*)^2\biggr)\biggl(\sum_{\mathcal{A}'}h(\mathcal{A}')^2\biggr),$$

where the sums run over the appropriate collections of annulus structures. The first sum is bounded by $H^*(j_*,k_*)$, and the proof of (4.4), possibly incorporating boundary clusters, shows that the second factor is bounded by $g(k-k_*)\bar\gamma_r(r2^j)$, with possibly a different choice of the constant ϑ implicit in g. (Note that the annulus structure $\mathcal{A}'$ may have just one annulus whose outer square is roughly at the scale corresponding to j. This case is handled by the computation in (4.15), and this is the reason for the estimate being of the type (4.4), rather than the type estimated in (4.6).) The induction hypothesis therefore gives

$$\begin{aligned}H^*(j,k)&\leqslant(k+2)^{O(1)}\sum_{k_*,j_*}\alpha_\ell(r2^{j_*},r2^j)^2g^*(k_*)\bar\gamma_r^*(r2^{j_*})g(k-k_*)\bar\gamma_r(r2^j)\\&\leqslant(k+2)^{O(1)}g(k)g^*(k)j\bar\gamma_r(r2^j)\bar\gamma_r^*(r2^j).\end{aligned}$$

If $\delta>0$ is small enough, then $j\bar\gamma_r(r2^j)\leqslant O(1)\bar\gamma_r^*(r2^j)^\delta$. Given this δ, if ϑ^* is large enough, we also have $O(1)(k+2)^{O(1)}g(k)\leqslant g^*(k)^\delta$. Thus, our last upper bound is at most $(g^*(k)\bar\gamma_r^*(r2^j))^{1+\delta}$, so (4.17) is proved, and our proof of Proposition 4.7 is complete. □

5. Partial independence in the spectral sample

5.1. Setup and main statement

Let $\mathscr{S}$ denote the spectral sample of the ±1-indicator function of having a percolation left-right crossing in $[0,R]^2$ (in either of our two favorite lattices). In order to prove that $|\mathscr{S}|$ is rarely much smaller than its mean it would be useful to have some independence of the following kind: if B_1 and B_2 are two distant squares, then we would expect that

$$\mathbb{P}[\mathscr{S}\cap B_1 = S_1 \,|\, \mathscr{S}\cap B_1\neq\varnothing, \mathscr{S}\cap B_2 = S_2] \asymp \mathbb{P}[\mathscr{S}\cap B_1 = S_1 \,|\, \mathscr{S}\cap B_1\neq\varnothing].$$

It turns out that it is hard to control such correlations. Nevertheless, we will prove a weaker independence result that will be enough for our purposes.

Consider some box B of radius r inside $[0,R]^2$. (Recall from §2.1 that a box $B(x,r)$ of radius r is the union of tiles whose centers are in $x+[-r,r)^2$.) We want to understand the behavior of $\mathscr{S}$ in B. Because of boundary issues, we will actually look at $\mathscr{S}$ in a smaller concentric box B', of radius $\frac{1}{3}r$.

We saw in (3.7) that $O(1)\mathbb{E}\big[|\mathscr{S}\cap B'|\,\big|\,\mathscr{S}\cap B\neq\varnothing\big]\geqslant r^2\alpha_4(r)$. In this section, we will strengthen this by proving that $|\mathscr{S}\cap B'|$ is at least of this size with a uniform positive probability, moreover, this remains true when we add $\mathscr{S}\cap W=\varnothing$ to the conditioning, where W is an arbitrary set in the complement of B:

$$\mathbb{P}[|\mathscr{S}\cap B'|\geqslant cr^2\alpha_4(r)\,|\,\mathscr{S}\cap B\neq\varnothing, \mathscr{S}\cap W=\varnothing] > a, \tag{5.1}$$

with some fixed constants $c,a>0$. However, the following stronger statement is closer to what we actually need.

Proposition 5.1. *Let $\mathscr{S}$ be the spectral sample of the ±1-indicator function of the left-right crossing event in $\mathcal{Q}=[0,R]^2$. Let B be a box of some radius r. Let B' be the concentric box with radius $\frac{1}{3}r$, and assume that $B'\subset\mathcal{Q}$ (note that B does not need to be in $\mathcal{Q}$). We also assume that $r\geqslant\bar{r}$, where $\bar{r}>0$ is some universal constant. Fix any set $W\subset\mathbb{R}^2\setminus B$, and let $\mathcal{Z}$ be a random subset of $\mathcal{I}$ that is independent of $\mathscr{S}$, where each element of $\mathcal{I}$ is in $\mathcal{Z}$ with probability $1/\alpha_4(r)r^2$ independently. (By (2.6), $\alpha_4(r)r^2\geqslant1$ if $\bar{r}$ is sufficiently large.) Then*

$$\mathbb{P}[\mathscr{S}\cap B'\cap\mathcal{Z}\neq\varnothing\,|\,\mathscr{S}\cap B\neq\varnothing, \mathscr{S}\cap W=\varnothing] > a,$$

where $a>0$ is a universal constant.

The estimate (5.1) follows immediately from the proposition, since

$$\mathbb{P}[\mathscr{S}\cap B'\cap\mathcal{Z}\neq\varnothing\,|\,\mathscr{S}\cap B', \mathscr{S}\cap B\neq\varnothing, \mathscr{S}\cap W=\varnothing] = 1-(1-r^{-2}\alpha_4(r)^{-1})^{|\mathscr{S}\cap B'|}.$$

It is important to note that in the proposition the constant $a>0$ is independent of the position of the box B relative to the square $[0,R]^2$. Such a uniform control over the domain would be harder to achieve in the case of general quads. Instead, after proving this uniform result for the square, we will prove a local version (Proposition 5.11) for general quads. We will also prove a radial version (Proposition 5.12), which will be important for the application to exceptional times of dynamical percolation.

The proof of the proposition is straightforward once we have the following bounds on the first and second moments. Recall the definition of $\lambda_{B,W}$ right before Lemma 2.2.

PROPOSITION 5.2. (First moment) *Assume the setup of Proposition* 5.1. *There is an absolute constant* $c_1>0$ *such that for any* $x\in B'\cap\mathcal{I}$,

$$\mathbb{P}[x\in\mathcal{S},\mathcal{S}\cap W=\varnothing]\geqslant c_1\mathbb{E}[\lambda_{B,W}^2]\alpha_4(r). \tag{5.2}$$

PROPOSITION 5.3. (Second moment) *Let* $\mathcal{S}$ *be the spectral sample of* $f=f_{\mathcal{Q}}$, *where* $\mathcal{Q}\subset\mathbb{R}^2$ *is some arbitrary quad. Let* $z\in\mathcal{Q}$ *and* $r>0$. *Set* $B:=B(z,r)$ *and* $B':=B\bigl(z,\frac{1}{3}r\bigr)$. *Suppose that* $B\bigl(z,\frac{1}{2}r\bigr)\subset\mathcal{Q}$ *and that* B *and* W *are disjoint. Then for every* $x,y\in B'\cap\mathcal{I}$, *we have*

$$\mathbb{P}[x,y\in\mathcal{S},\mathcal{S}\cap W=\varnothing]\leqslant c_2\mathbb{E}[\lambda_{B,W}^2]\alpha_4(|x-y|)\alpha_4(r), \tag{5.3}$$

where $c_2<\infty$ *is an absolute constant.*

Proof of Proposition 5.1. (Assuming the first- and second-moment estimates.) Consider the random variable

$$Y:=|\mathcal{S}\cap B'\cap\mathcal{Z}|1_{\{\mathcal{S}\cap W=\varnothing\}}.$$

Since $\mathcal{Z}$ is independent of $\mathcal{S}$ and $\mathbb{P}[x\in\mathcal{Z}]=1/\alpha_4(r)r^2$, we obtain, by summing (5.2) over all $x\in B'\cap\mathcal{I}$, that $O(1)\mathbb{E}[Y]\geqslant\mathbb{E}[\lambda_{B,W}^2]$. On the other hand, summing (5.3) over all $x,y\in B'\cap\mathcal{I}$, similarly to the second-moment estimate in Lemma 3.1, gives

$$\begin{aligned}\mathbb{E}[Y^2]&\leqslant\overbrace{O(1)\mathbb{E}[\lambda_{B,W}^2]\alpha_4(r)r^2\mathbb{P}[x\in\mathcal{Z}]}^{\text{diagonal term}}+\overbrace{O(1)\mathbb{E}[\lambda_{B,W}^2]\alpha_4(r)^2r^4\mathbb{P}[x\in\mathcal{Z}]^2}^{\text{off-diagonal term}}\\&\leqslant O(1)\mathbb{E}[\lambda_{B,W}^2],\end{aligned}$$

by our choice of $\mathbb{P}[x\in\mathcal{Z}]$. Note that this choice for $\mathbb{P}[x\in\mathcal{Z}]$ is of the smallest possible order that does not make the diagonal term the leading contribution. Now, by Cauchy–Schwarz' inequality,

$$\mathbb{P}[Y>0]\geqslant\frac{\mathbb{E}[Y]^2}{\mathbb{E}[Y^2]}\geqslant\frac{\mathbb{E}[\lambda_{B,W}^2]^2}{O(1)\mathbb{E}[\lambda_{B,W}^2]}=\frac{\mathbb{E}[\lambda_{B,W}^2]}{O(1)}. \tag{5.4}$$

The proposition now follows from Lemma 2.2. □

Remark 5.4. $\mathbb{P}[\mathscr{S}\cap B\neq\varnothing, \mathscr{S}\cap W=\varnothing]$ is obviously not smaller than the left-hand side of (5.4). Therefore, (5.4) and Lemma 2.2 imply that in the present setting

$$\mathbb{P}[\mathscr{S}\cap B\neq\varnothing, \mathscr{S}\cap W=\varnothing]\asymp\mathbb{E}[\lambda_{B,W}^2]. \tag{5.5}$$

The definition of $\lambda_{B,W}$ easily gives

$$\mathbb{E}[\lambda_{B,\varnothing}^2]=\alpha_\square(B,\mathcal{Q}) \quad\text{and}\quad \lambda_{B,B^c}=\alpha_\square(B,\mathcal{Q}). \tag{5.6}$$

Combining these with (5.5), we get that for B as above, approximate equalities hold in Lemma 3.2, i.e.,

$$\mathbb{P}[\mathscr{S}\cap B\neq\varnothing]\asymp\alpha_\square(B,\mathcal{Q}) \quad\text{and}\quad \mathbb{P}[\varnothing\neq\mathscr{S}\subseteq B]\asymp\alpha_\square(B,\mathcal{Q})^2. \tag{5.7}$$

5.2. Bounding the second moment

Due to the way in which $\lambda_{B,W}$ was defined, it is generally easier to obtain $\lambda_{B,W}$ as an upper bound up to constants, than as a lower bound up to constants. Consequently, the second-moment estimate is easier to prove, and for this reason we start with that.

Proof of Proposition 5.3. Let θ denote the restriction of ω to the complement of $W\cup\{x,y\}$. Then Lemma 2.1 gives

$$\mathbb{P}[x,y\in\mathscr{S}, \mathscr{S}\cap W=\varnothing]=\mathbb{P}[\mathscr{S}\cap(W\cup\{x,y\})=\{x,y\}]=\mathbb{E}[\mathbb{E}[\chi_{\{x,y\}}(\omega)f(\omega)\,|\,\theta]^2]. \tag{5.8}$$

Set

$$g(\theta):=\mathbb{E}[\chi_{\{x,y\}}(\omega)f(\omega)\,|\,\theta].$$

Then $\mathbb{E}[g^2]$ is the quantity that we need to estimate. Since $B\cap W=\varnothing$, the information in θ includes the configuration in $B\setminus\{x,y\}$. If θ does not have the 4-arm event from the tile of x to distance $\frac{1}{4}|x-y|$, then flipping ω_x does not effect $f(\omega)$, and hence $g(\theta)=0$. A similar statement holds for y. Also, if the box $\widetilde{B}$ of radius $2|x-y|$ centered at $\frac{1}{2}(x+y)$ does not intersect ∂B, then $g(\theta)=0$ unless θ has the 4-arm event in the corresponding annulus $B\setminus\widetilde{B}$. Let A_x, A_y and $A_{x,y}$ denote the indicator functions for the 4-arm event in the corresponding three annuli, where we take $A_{x,y}=1$ if $\widetilde{B}\cap\partial B\neq\varnothing$. Then we have $g(\theta)=0$ if $A_xA_yA_{x,y}=0$.

We now argue that $|g(\theta)|\leqslant\lambda_{B,W}$. For this purpose, write

$$g=\mathbb{E}[\chi_{\{x,y\}}f\,|\,\mathcal{F}_{(W\cup\{x,y\})^c}]=\mathbb{E}[\mathbb{E}[\chi_{\{x,y\}}f\,|\,\mathcal{F}_{\{x,y\}^c}]\,|\,\mathcal{F}_{W^c}],$$

where we used that our measure is i.i.d. Clearly, $|\mathbb{E}[\chi_{\{x,y\}} f | \mathcal{F}_{\{x,y\}^c}]| \leqslant 1_{\Lambda_{\{x,y\}}} \leqslant 1_{\Lambda_B}$, where $\Lambda_\cdot$ is as defined above Lemma 2.2. Taking conditional expectation given $\mathcal{F}_{W^c}$ then gives

$$\big|\mathbb{E}\big[\mathbb{E}[\chi_{\{x,y\}} f \,|\, \mathcal{F}_{\{x,y\}^c}] \,\big|\, \mathcal{F}_{W^c}\big]\big| \leqslant \mathbb{E}\big[|\mathbb{E}[\chi_{\{x,y\}} f \,|\, \mathcal{F}_{\{x,y\}^c}]| \,\big|\, \mathcal{F}_{W^c}\big] \leqslant \lambda_{B,W}.$$

Since the left-hand side is $|g|$, we get $|g| \leqslant \lambda_{B,W}$.

Putting together the above, we arrive at $|g(\theta)| \leqslant A_x A_y A_{x,y} \lambda_{B,W}$. Thus,

$$g(\theta)^2 \leqslant A_x A_y A_{x,y} \lambda_{B,W}^2.$$

Independence on disjoint sets then gives

$$\mathbb{E}[g^2] \leqslant \alpha_4\big(\tfrac{1}{4}|x-y|\big)^2 \alpha_4\big(2|x-y|, \tfrac{1}{3}r\big)\mathbb{E}[\lambda_{B,W}^2].$$

The proposition now follows from the familiar properties of α_4. □

5.3. Reformulation of the first-moment estimate

Before proving the first-moment estimate (Proposition 5.2), we explain how it can be reformulated as a quasi-multiplicativity property analogous to the quasi-multiplicativity property of the j-arm events (2.2). Recall that

$$\mathbb{E}[\lambda_{B,W}^2] = \mathbb{E}[\mathbb{P}[\Lambda_B \,|\, \mathcal{F}_{W^c}]^2].$$

It is not a priori clear how to work with $\mathbb{E}[\lambda_{B,W}^2]$, but here is a useful observation about this quantity. Let ω' and ω'' be two critical percolation configurations which coincide on W^c but are independent on W. Let $\mathscr{A}_\square(B, \mathcal{Q})$ denote the set of percolation configurations ω for which the 4-arm event occurs in the annulus $\mathcal{Q} \setminus B$ with the appropriately colored arms terminating on the correct boundary arcs of $\mathcal{Q}$; that is, the primal (white) arms terminating on the two distinguished arcs of $\partial\mathcal{Q}$ and the dual (black) arms terminating on the two complementary arcs. Then

$$\mathbb{E}[\lambda_{B,W}^2] = \mathbb{P}[\omega', \omega'' \in \Lambda_B] = \mathbb{P}[\omega', \omega'' \in \mathscr{A}_\square(B, \mathcal{Q})];$$

that is, $\mathbb{E}[\lambda_{B,W}^2]$ is just the probability that the corresponding 4-arm event occurs in both ω' and ω''. Lemma 2.1 gives

$$\mathbb{P}[x \in \mathscr{S}, \mathscr{S} \cap W = \varnothing] = \mathbb{E}[\mathbb{E}[f\chi_x \,|\, \mathcal{F}_{(W \cup \{x\})^c}]^2].$$

Now, if f is an increasing function taking values in $\{-1,1\}$, then

$$\mathbb{E}[f\chi_x \,|\, \mathcal{F}_{\{x\}^c}] = 1_{\Lambda_{\{x\}}}, \tag{5.9}$$

and

$$\mathbb{E}[f\chi_x \,|\, \mathcal{F}_{(W\cup\{x\})^c}] = \mathbb{E}\big[\mathbb{E}[f\chi_x \,|\, \mathcal{F}_{\{x\}^c}] \,\big|\, \mathcal{F}_{W^c}\big] = \mathbb{E}[1_{\Lambda_{\{x\}}} \,|\, \mathcal{F}_{W^c}] = \lambda_{x,W}.$$

Hence,

$$\mathbb{P}[x \in \mathscr{S}, \mathscr{S} \cap W = \varnothing] = \mathbb{E}[\lambda_{x,W}^2] = \mathbb{P}[\omega', \omega'' \in \mathscr{A}_\square(x, \mathcal{Q})],$$

where $\mathscr{A}_\square(x,\mathcal{Q})$ has the obvious meaning. Likewise, since $W\cap B=\varnothing$, we have $\omega'=\omega''$ in B, and so

$$\alpha_4(r) \asymp \mathbb{P}[\omega', \omega'' \in \mathscr{A}_4(x, B)],$$

where $\mathscr{A}_4(x,B)$ is the 4-arm event (which does not pay attention to any distinguished arcs on ∂B). Hence (5.2) can be rewritten as

$$\mathbb{P}[\omega', \omega'' \in \mathscr{A}_\square(x, \mathcal{Q})] \geqslant c_1 \mathbb{P}[\omega', \omega'' \in \mathscr{A}_4(x, B)]\mathbb{P}[\omega', \omega'' \in \mathscr{A}_\square(B, \mathcal{Q})]. \tag{5.10}$$

To see that this is indeed a quasi-multiplicativity property, observe that if we take $W=\varnothing$ and replace the events with $\mathscr{A}_\square$ by the corresponding events with $\mathscr{A}_4$, then this is essentially the same as the case $j=4$ in the left inequality of (2.2).

It turns out that, with a few extra twists, a proof which gives the quasi-multiplicativity estimates (2.2) generalizes to give (5.10). This will be explained in the next subsections.

Remark 5.5. Proposition 5.1 generalizes to the radial setting, in which we consider the event of a crossing from the origin to a large distance away. However, at present it does not generalize to the radial 2-arm event where a vacant crossing and an occupied crossing occur simultaneously. The only argument in the proof that does not generalize to the 2-arm event is (5.9), which is not true for non-monotone functions. Instead, we have

$$\mathbb{E}[f\chi_x \,|\, \mathcal{F}_{\{x\}^c}] = 1_{M_x^+} - 1_{M_x^-}, \tag{5.11}$$

where M_x^+ is the event that x is monotonically pivotal (i.e., $f\big(\omega_{\{x\}}^+\big)=1=-f\big(\omega_{\{x\}}^-\big)$) and M_x^- is the event that x is anti-monotonically pivotal. The problem with such functions is that for the first moment we would need to bound $\mathbb{E}[\mathbb{E}[1_{M_x^+}-1_{M_x^-}\,|\mathcal{F}_{W^c}]^2]$ from below. This expression is easily controlled from above by $\mathbb{E}[\lambda_{x,W}^2]$, but not from below due to cancellations between M_x^+ and M_x^-. These cancellations are far from being negligible, and thus there is no hope to get $O(1)\mathbb{E}[\mathbb{E}[1_{M_x^+}-1_{M_x^-}\,|\mathcal{F}_{W^c}]^2]\geqslant\mathbb{E}[\lambda_{x,W}^2]$ for general W. For

instance, if $W=\{x\}^c$ and f is an even function ($f(-\omega)=f(\omega)$) like the 2-arm indicator function for site percolation on the triangular grid, then

$$\mathbb{P}[\mathscr{S}=\{x\}]=\mathbb{E}[\mathbb{E}[1_{M_x^+}-1_{M_x^-}]^2]=0.$$

This "unfortunate" cancellation between the events M_x^+ and M_x^- is the reason for the breakdown of our methods for such events.

5.4. Quasi-multiplicativity for coupled configurations

Rather than proving inequality (5.10) specifically, we first address a related statement which is somewhat cleaner. In the following, W is any fixed subset of $\mathcal{I}$, and ω' and ω'' are the above coupled configurations, which are independent in W and agree on $\mathcal{I}\setminus W$. The annulus $B(0,R)\setminus B(0,r)$ will be denoted by $A(r,R)$. Let $j\in\mathbb{N}_+$ be either 1 or an even number and let $\mathscr{A}_j(r,R)$ denote the set of configurations ω that satisfy the alternating j-arm event in the annulus $A(r,R)$. Set

$$\beta_j^W(r,R):=\mathbb{P}[\omega',\omega''\in\mathscr{A}_j(r,R)].$$

We will prove the following quasi-multiplicativity result.

PROPOSITION 5.6. (Quasi-multiplicativity) *Let $j\in\mathbb{N}_+$ be either 1 or an even integer, and let $W\subseteq\mathcal{I}$. Then*

$$\beta_j^W(r_1,r_2)\beta_j^W(r_2,r_3)\leqslant C_j\beta_j^W(r_1,r_3)$$

holds for every $0<r_1<r_2<r_3$ satisfying $r_2\geqslant\bar{r}_j$, where C_j and $\bar{r}_j$ are finite constants depending only on j (and, in particular, not on W).

Note that the opposite inequality with $C_j=1$ holds by independence on disjoint sets.

To prepare for the proof of the proposition, we first need to prove a few lemmas. The first observation is the following monotonicity property:

$$\beta_j^{W_2}(r,R)\leqslant\beta_j^{W_1}(r,R)\quad\text{if } W_1\subseteq W_2. \tag{5.12}$$

Indeed, since

$$\beta_j^W(r,R)=\mathbb{E}[\mathbb{P}[\omega\in\mathscr{A}_j(r,R)\,|\,\mathcal{F}_{W^c}]^2],$$

the claimed monotonicity follows by the orthogonality property of martingale increments.

The case $j=1$ in Proposition 5.6 easily follows from the Russo–Seymour–Welsh theorem and from the Harris–FKG inequality. In the following, we will restrict ourselves to the case $j=4$, since the other even values of j are essentially the same.

Let δ be some small positive constant, and let $r_0>0$. We say that $r\geqslant r_0$ is *δ-good* if $\beta_4^W(r_0,2r)\geqslant\delta\beta_4^W(r_0,r)$. Of course, this notion of goodness depends on W, δ and r_0.

LEMMA 5.7. *For any $\delta>0$, there exist $\bar{r}=\bar{r}(\delta)>0$ and $c=c(\delta)>0$ (both depending only on δ), such that for any $W\subseteq\mathcal{I}$ and any $r_0>0$ the following holds: if one assumes that $r\geqslant r_0\vee\bar{r}$ is δ-good, then, for every $r'>r$,*

$$\beta_4^{W\cup A(r,r')}(r_0,r')\geqslant c\beta_4^W(r_0,r)\left(\frac{r}{r'}\right)^d,$$

where d is a universal constant.

The proof of this lemma will rely on Lemmas A.2 and A.3 from [SSt].

Proof. Assume that r is δ-good. Then $\beta_4^W(r_0,2r)\geqslant\delta\beta_4^W(r_0,r)$. Set

$$X':=\mathbb{P}[\omega'\in\mathscr{A}_4(r_0,2r)\,|\,\omega'_{B(0,r)}],$$
$$X'':=\mathbb{P}[\omega''\in\mathscr{A}_4(r_0,2r)\,|\,\omega''_{B(0,r)}].$$

Then

$$\beta_4^W(r_0,2r)=\mathbb{P}[\omega',\omega''\in\mathscr{A}_4(r_0,2r)]=\mathbb{E}[\mathbb{P}[\omega',\omega''\in\mathscr{A}_4(r_0,2r)\,|\,\omega'_{B(0,r)},\omega''_{B(0,r)}]].\quad(5.13)$$

Now, since

$$\mathbb{P}[\omega',\omega''\in\mathscr{A}_4(r_0,2r)\,|\,\omega'_{B(0,r)},\omega''_{B(0,r)}]\leqslant\mathbb{P}[\omega'\in\mathscr{A}_4(r_0,2r)\,|\,\omega'_{B(0,r)},\omega''_{B(0,r)}]=X',$$

and a similar relation holds with X'', we have

$$\mathbb{P}[\omega',\omega''\in\mathscr{A}_4(r_0,2r)\,|\,\omega'_{B(0,r)},\omega''_{B(0,r)}]\leqslant X'\wedge X''=:\widetilde{X},$$

where $\widetilde{X}=X'\wedge X''$ denotes the minimum of X' and X''. Because r is δ-good, (5.13) now gives $\mathbb{E}[\widetilde{X}]\geqslant\delta\beta_4^W(r_0,r)$. Since $\{\widetilde{X}>0\}\subseteq\{\omega',\omega''\in\mathscr{A}_4(r_0,r)\}$, and the latter event has probability $\beta_4^W(r_0,r)$, this gives

$$\mathbb{E}[\widetilde{X}\,|\,\omega',\omega''\in\mathscr{A}_4(r_0,r)]\geqslant\delta.\quad(5.14)$$

Now let $\widetilde{\omega}'$ and $\widetilde{\omega}''$ be two percolation configurations that have the same law as ω, that are independent of each other outside of $B(0,r)$, and inside $B(0,r)$ satisfy $\widetilde{\omega}'=\omega'$ and $\widetilde{\omega}''=\omega''$. Let s' be the least distance between the endpoints on $\partial B(0,2r)$ of any pair of disjoint interfaces of $\widetilde{\omega}'$ that cross the annulus $A(r,2r)$. (Take $s'=\infty$ if there is at most one such interface.) We claim that r/s' is tight, in the following sense: for every $\varepsilon>0$ there is a constant $M=M_\varepsilon$, depending only on ε, such that $\mathbb{P}[r/s'>M]<\varepsilon$. This is proved, for example, in [SSt, Lemma A.2]. We use this with $\varepsilon=\frac{1}{2}\delta$. Thus, we have

$$\mathbb{P}\left[s'<\frac{r}{M}\right]<\frac{\delta}{2}.\quad(5.15)$$

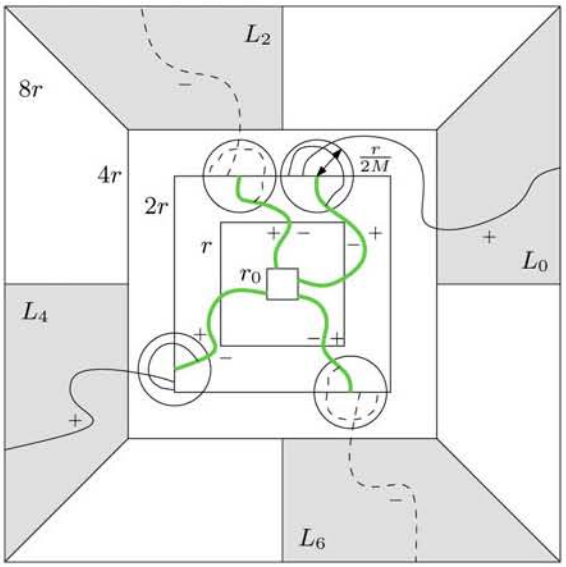


Figure 5.1. How to use "separation of arms" in order to get equation (5.16).

This property will be referred to below as the "separation of arms" phenomenon.

Assume now that $r \geqslant 100M =: \bar{r}$. Then when $s' \geqslant r/M$, we know that s' is substantially larger than the lattice mesh. Observe that the distance between the endpoints on $\partial B(0,2r)$ of any two disjoint interfaces of $\widetilde{\omega}'$ that cross $A(r_0,2r)$ is at least s' (since every such interface also crosses $A(r,2r)$), and if $\widetilde{\omega}' \in \mathscr{A}_4(r_0,2r)$ then there exist at least four such interfaces. Let L_k denote the sector $\{\varrho e^{i\theta} : \varrho>0 \text{ and } \theta \in [\frac{1}{4}\pi k, \frac{1}{4}\pi(k+1)]\}$. Let $\mathcal{Z}'$ denote the event that in $\widetilde{\omega}'$ for each $k \in \{0,2,4,6\}$ there is a crossing from $\partial B(0,r_0)$ to $\partial B(0,8r)$ in $L_k \cup A(r_0,4r)$, which is white when $k \in \{0,4\}$ and black when $k \in \{2,6\}$. By the proof of [SSt, Lemma A.3], we know (see Figure 5.1) that there is a constant $c_0 = c_0(M) > 0$ such that

$$\mathbb{P}\Big[\mathcal{Z}' \,\Big|\, \widetilde{\omega}'_{B(0,r)}, \widetilde{\omega}' \in \mathscr{A}_4(r_0,2r), s' \geqslant \frac{r}{M}\Big] \geqslant c_0. \tag{5.16}$$

Note that s' is independent of $\widetilde{\omega}'_{B(0,r)} = \omega'_{B(0,r)}$. Therefore, (5.15) gives

$$\begin{aligned}\mathbb{P}\Big[s' \geqslant \frac{r}{M}, \widetilde{\omega}' \in \mathscr{A}_4(r_0,2r) \,\Big|\, \omega'_{B(0,r)}\Big] &\geqslant \mathbb{P}[\widetilde{\omega}' \in \mathscr{A}_4(r_0,2r) \,|\, \omega'_{B(0,r)}] - \mathbb{P}\Big[s' < \frac{r}{M} \,\Big|\, \omega'_{B(0,r)}\Big] \\ &\geqslant \widetilde{X} - \tfrac{1}{2}\delta.\end{aligned}$$

Together with (5.16), this shows that

$$\mathbb{P}[\mathcal{Z}' \,|\, \omega'_{B(0,r)}] \geqslant c_0\big(\widetilde{X} - \tfrac{1}{2}\delta\big).$$

Now let $\mathcal{Z}''$ be defined as $\mathcal{Z}'$, but with $\widetilde{\omega}''$ replacing $\widetilde{\omega}'$. Since $\widetilde{\omega}'$ and $\widetilde{\omega}''$ are conditionally independent given $(\omega'_{B(0,r)}, \omega''_{B(0,r)})$, we get

$$\mathbb{P}\big[\mathcal{Z}', \mathcal{Z}'' \,\big|\, \omega'_{B(0,r)}, \omega''_{B(0,r)}\big] \geqslant c_0^2\big(\widetilde{X} - \tfrac{1}{2}\delta\big)_+^2, \tag{5.17}$$

where $(x)_+$ denotes $x \vee 0$. Since $\big(\widetilde{X} - \frac{1}{2}\delta\big)_+^2$ is a convex function of $\widetilde{X}$, we get from Jensen's inequality and (5.14) that

$$\mathbb{E}\big[\big(\widetilde{X} - \tfrac{1}{2}\delta\big)_+^2 \,\big|\, \omega', \omega'' \in \mathscr{A}_4(r_0, r)\big] \geqslant \tfrac{1}{4}\delta^2.$$

Thus, taking the expectation of both sides of (5.17) gives

$$\mathbb{P}[\mathcal{Z}', \mathcal{Z}''] \geqslant c_0^2\big(\tfrac{1}{4}\delta^2\big)\mathbb{P}[\omega', \omega'' \in \mathscr{A}_4(r_0, r)] = c_0^2\big(\tfrac{1}{4}\delta^2\big)\beta_4^W(r_0, r). \tag{5.18}$$

This clearly implies the statement of the lemma in the case where $r' \leqslant 8r$. Assume therefore that $r' > 8r$. Note that $\mathcal{Z}' \cap \mathcal{Z}''$ is increasing inside $(L_0 \cup L_4) \setminus B(0, 6r)$ and decreasing in $(L_2 \cup L_6) \setminus B(0, 6r)$. Hence, it is positively correlated with the event $\widetilde{\mathcal{Z}}$ that for each of $\widetilde{\omega}'$ and $\widetilde{\omega}''$ there are white paths separating $\partial B(0, 6r)$ from $\partial B(0, 8r)$ in each of L_0 and L_4, and similar black paths in L_2 and L_6, and moreover, there are black paths in each of L_2 and L_6 joining $\partial B(0, 6r)$ and $\partial B(0, r')$ and white paths in each of L_0 and L_4 joining $\partial B(0, 6r)$ and $\partial B(0, r')$. By the Russo–Seymour–Welsh theorem (see Figure 5.2), $\mathbb{P}[\widetilde{\mathcal{Z}}] \geqslant c_1(r/r')^d$ for some absolute constants $c_1 > 0$ and $d < \infty$.

Taking $c := \frac{1}{4}c_1 c_0^2 \delta^2$, we obtain that $\mathbb{P}[\mathcal{Z}', \mathcal{Z}'', \widetilde{\mathcal{Z}}] \geqslant c\beta_4^W(r_0, r)(r/r')^d$. The lemma follows, since $\mathcal{Z}' \cap \mathcal{Z}'' \cap \widetilde{\mathcal{Z}} \subseteq \{\widetilde{\omega}', \widetilde{\omega}'' \in \mathscr{A}_4(r_0, r')\}$. □

Lemma 5.8. *There are absolute constants $\delta_0 > 0$ and $\bar{R} > 0$ such that*

$$\beta_4^W(r_0, 2\varrho) \geqslant \delta_0 \beta_4^W(r_0, \varrho) \tag{5.19}$$

holds for any $r_0 > 0$ and any $\varrho \geqslant r_0 \vee \bar{R}$. Furthermore,

$$\beta_4^W\big(\tfrac{1}{2}\varrho, r_0\big) \geqslant \delta_0 \beta_4^W(\varrho, r_0) \tag{5.20}$$

holds for any $\bar{R} \leqslant \varrho \leqslant r_0$.

Proof. We start by proving the first claim. Let us consider some $\delta \in \big(0, 2^{-(d+1)}\big)$ (where d is the universal constant from Lemma 5.7), whose value will be determined later. Let $\bar{r} = \bar{r}(\delta)$ and $c(\delta)$ be as in Lemma 5.7. Let $r_0 > 0$ be some radius. Let $r \geqslant \bar{r} \vee r_0$, and assume for now that r is δ-good. Then, by Lemma 5.7 and the monotonicity property (5.12), we have

$$\beta_4^W(r_0, r') \geqslant c(\delta)\beta_4^W(r_0, r)\Big(\frac{r}{r'}\Big)^d \tag{5.21}$$

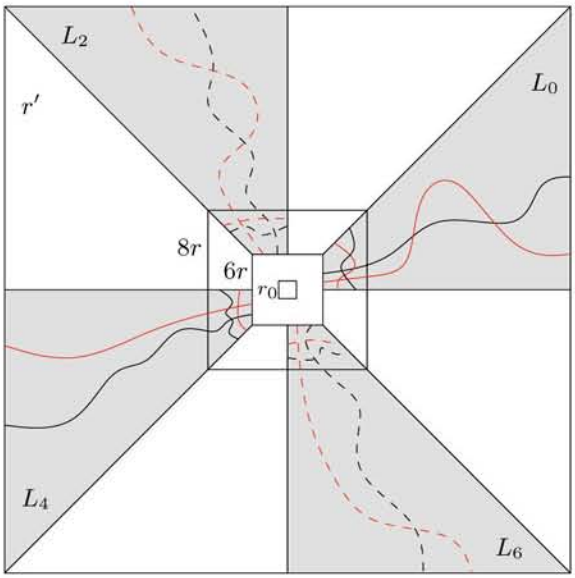


Figure 5.2. A realization of the event $\widetilde{\mathcal{Z}}$. In each sector L_{2i}, there is one arm in $\widetilde{\omega}'$ and one in $\widetilde{\omega}''$.

for every $r'>r$. Set $\varrho_k:=2^k r$, and let $\bar{k}:=\inf\{k\in\mathbb{N}_+:\varrho_k \text{ is } \delta\text{-good}\}$, with $\bar{k}:=\infty$ if this set is empty. If $m\in\mathbb{N}_+$ and $m\leqslant\bar{k}$, then by the definition of $\bar{k}$ and by (5.21) with $r':=\varrho_m$, we have

$$\beta_4^W(r_0,\varrho_1)\delta^{m-1}\geqslant\beta_4^W(r_0,\varrho_m)\geqslant c(\delta)\beta_4^W(r_0,r)2^{-dm}\geqslant c(\delta)\beta_4^W(r_0,\varrho_1)2^{-dm}.$$

Now, since $\delta<2^{-d-1}$, the above gives $2^{-(d+1)(m-1)}\geqslant c(\delta)2^{-dm}$, which implies that $2^{d+1}\geqslant c(\delta)2^m$. Hence, $\bar{k}$ is bounded from above by some finite constant depending only on δ (recall that d is a universal constant). We may conclude that δ-good radii appear in scales with bounded gaps, since the same argument may be applied with r replaced by $\varrho_{\bar{k}}$. If ϱ is in the range $(r,\varrho_{\bar{k}})$, then we have the estimate

$$\frac{\beta_4^W(r_0,2\varrho)}{\beta_4^W(r_0,\varrho)}\geqslant\frac{\beta_4^W(r_0,2\varrho_{\bar{k}})}{\beta_4^W(r_0,r)}\overset{(5.21)}{\geqslant}c(\delta)2^{-d(\bar{k}+1)},$$

which means that ϱ is δ_1-good with $\delta_1:=\delta_1(\delta)=c(\delta)2^{-d(\bar{k}+1)}$. The same statement applies to any $\varrho\geqslant r$, since, above r, the δ-good radii appear in scales with bounded gaps.

The proof of (5.19) is nearly complete, since we only have to initialize the above recursion. Given $r_0>0$, we want to find $\delta>0$ small enough and some $\overline{R}\geqslant\bar{r}=\bar{r}(\delta)$ that does not depend on r_0, such that $\overline{R}\vee r_0$ will be δ-good (recall that the definition of δ-good in Lemma 5.7 above does depend on r_0). Once this is established, we can start the induction with $r:=\overline{R}\vee r_0$, and (5.19) will be satisfied for all $\varrho\geqslant r$, with $\delta_0:=\delta_1$.

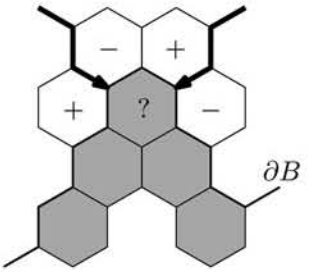


Figure 5.3. If we had chosen boxes with such boundaries, then in this case whatever the color of the ?-hexagon is, we would not be able to continue both interfaces.

Clearly, by the RSW estimate, $\beta_4^W(r,2r)$ is bounded from below once $r\geqslant R_0$, where R_0 is some absolute constant. We now fix $\delta:=\delta(R_0)\in(0,2^{-(d+1)})$ such that for any $r\geqslant R_0$,

$$\beta_4^W(r,2r)\geqslant\delta.$$

Now that δ is fixed, define $\bar{R}:=\bar{r}(\delta)\vee R_0$. This certainly ensures that the requirement $\bar{R}\geqslant\bar{r}$ is fulfilled.

We now distinguish between two cases. If $r_0\geqslant\bar{R}$, then we want to start the induction with $r:=r_0$, which is fine, because $r_0\geqslant\bar{R}\geqslant R_0$, hence r_0 is δ-good (for r_0 and W) by the choice of δ. We thus obtain the desired inequality (5.19) for any $\varrho\geqslant r=r_0$, with $\delta_0:=\delta_1(\delta)$.

Let us now deal with the case where $0<r_0<\bar{R}$. Let $\varrho\geqslant\bar{R}$. Note that for any coupled configurations ω',ω'' satisfying $\{\omega',\omega''\in\mathscr{A}_4(\bar{R},2\varrho)\}$, there always exists a pair of configurations $\widetilde{\omega}',\widetilde{\omega}''$ in the ball $B(\bar{R})$ such that the concatenations $(\widetilde{\omega}',\omega')$ and $(\widetilde{\omega}'',\omega'')$ both satisfy $\mathscr{A}_4(r_0,2\varrho)$. Note that, for this to be entirely rigorous, one would need to choose the boxes $B(r)$ in a proper way to exclude discrete effects; see Figure 5.3. On the triangular lattice, choosing the balls according to the triangular graph distance does it, for instance.

This implies that, for any $1<r_0<\bar{R}\leqslant\varrho$,

$$\beta_4^W(r_0,2\varrho)\geqslant\left(\tfrac{1}{2}\right)^{2|B(\bar{R})|}\beta_4^W(\bar{R},2\varrho)\geqslant\left(\tfrac{1}{2}\right)^{2|B(\bar{R})|}\delta_1(\delta)\beta_4^W(\bar{R},\varrho)\geqslant\left(\tfrac{1}{2}\right)^{2|B(\bar{R})|}\delta_1(\delta)\beta_4^W(r_0,\varrho),$$

where the second inequality follows from $r_0=\bar{R}$ in the above analysis. Hence the first claim of the lemma is now proved with $\delta_0:=2^{-2|B(\bar{R})|}\delta_1(\delta)$.

To prove (5.20), we follow a similar argument but with annuli growing towards 0 rather than towards infinity. For this, a corresponding analogue of Lemma 5.7 is needed. Since the proofs in this case are essentially the same, they are omitted. □

Proof of Proposition 5.6. As remarked above, we only prove the case $j=4$, since $j=1$ is very easy and the proof for the case $j=4$ applies to all even j, with no essential changes.

If $r_2 \leqslant 4r_1$ or $r_3 \leqslant 4r_2$, then the claim follows from Lemma 5.8. Hence, assume that $r_1 < \frac{1}{4}r_2$ and $4r_2 < r_3$. By the monotonicity property (5.12), it suffices to prove that

$$O(1)\beta_4^{W \cup A(r_2/4, 4r_2)}(r_1, r_3) \geqslant \beta_4^W\big(r_1, \tfrac{1}{4}r_2\big)\beta_4^W(4r_2, r_3).$$

This follows from the proof of Lemma 5.7: we just need to apply the same argument twice, once going outwards from 0 and using (5.19) with $r_0 := r_1$ and $\varrho := \frac{1}{4}r_2$ to verify that $\frac{1}{4}r_2$ is δ_0-good, and once going inwards towards 0 and using (5.20) with $r_0 := r_3$ and $\varrho := 4r_2$. The easy details are left to the reader. □

Although this will not be needed in this paper, we note that the following generalization of Proposition 5.6 to arbitrary sequences of crossings holds. (This can be proved by combining the above arguments with the proof of [SSt, Proposition A.5].)

PROPOSITION 5.9. *Let $j \in \mathbb{N}_+$ and fix a color sequence $X \in \{\text{black}, \text{white}\}^j$. For any set $W \subseteq \mathcal{I}$, the probabilities for the existence in both coupled configurations ω' and ω'', of j crossings whose colors match this sequence in counterclockwise order satisfy the inequalities in Proposition* 5.6.

5.5. Proof of the first-moment estimate

Proof of Proposition 5.2. As remarked in §5.3, the proof of the first-moment estimate reduces to proving (5.10). We will now explain how the proof of Proposition 5.6 needs to be adapted to give (5.10). As in the case of Proposition 5.6, one needs to show that arms "tend to separate" when one conditions on the event we are interested in (i.e. $\mathscr{A}_\square(x, \mathcal{Q})$ here). In Proposition 5.6, the conditioning was very symmetric around x, while for Proposition 5.2, if the point x happens to be close to the boundary $\partial\mathcal{Q}$, then, under the conditioning, boundary effects will have a strong influence. In order to deal with this influence of the boundary, it is convenient to locate the point x first with respect to its closest edge (basically inducing a 3-arm event) and then with respect to its closest corner (inducing a 2-arm event).

Fix some $x \in \mathcal{I}$ that is relevant for the left-right crossing in $\mathcal{Q} = [0, R]^2$. Let x^+ be the closest point to x on $\partial\mathcal{Q}$ and let x^{++} be the closest point to x among the four corners of $\mathcal{Q}$. Set $R^+ := \|x - x^+\|_\infty$ and $R^{++} := \|x - x^{++}\|$. We now define

$$B_r := \begin{cases} B(x, r), & \text{if } \bar{r} \leqslant r \leqslant \frac{1}{8}R^+, \\ B(x^+, r), & \text{if } 8R^+ \leqslant r \leqslant \frac{1}{8}R^{++}, \\ B(x^{++}, r), & \text{if } 8R^{++} \leqslant r \leqslant \frac{1}{8}R, \end{cases}$$

where $\bar{r}>1$ is some fixed constant, and let $\mathcal{R}$ denote the set of r for which B_r is defined; that is, $\mathcal{R}:=\left[\bar{r}, \frac{1}{8}R^{+}\right]\cup\left[8R^{+}, \frac{1}{8}R^{++}\right]\cup\left[8R^{++}, \frac{1}{8}R\right]$. Given any $r\in\mathcal{R}$, define

$$\tilde{r} := \inf(\mathcal{R}\cap[2r, R]).$$

Then we say that r is *δ-good* if $\mathbb{P}[\omega', \omega''\in\mathscr{A}_4(x, B_{\tilde{r}})]\geqslant\delta\mathbb{P}[\omega', \omega''\in\mathscr{A}_4(x, B_r)]$. The proof that there is a universal constant δ such that every $r\in\mathcal{R}$ satisfying $2r\leqslant\sup\mathcal{R}$ is δ-good proceeds like the proof of (5.19) with a few minor changes. The fact that some of the boxes considered are not concentric with each other is of no consequence. The only significant modification needed is that in the argument corresponding to Lemma 5.7, if $B_r\cap\partial\mathcal{Q}\neq\varnothing$, then the interfaces considered are in the intersection of the corresponding annulus and $\mathcal{Q}$ and the definition of s' needs to be modified. In the adapted proof, s' is defined as the least distance between any two distinct points that are either endpoints on $\partial B_{\tilde{r}}$ of the interfaces or points in the intersection $\partial B_{\tilde{r}}\cap\partial\mathcal{Q}$. The remaining details are left to the reader, as is the similar proof that $\mathbb{P}[\omega', \omega''\in\mathscr{A}_{\square}(B_{\check{r}}, \mathcal{Q})]\geqslant\delta\mathbb{P}[\omega', \omega''\in\mathscr{A}_{\square}(B_r, \mathcal{Q})]$ when $\check{r}=\sup\bigl(\mathcal{R}\cap\left[\bar{r}, \frac{1}{2}r\right]\bigr)\geqslant\bar{r}$. The proof of (5.10) then follows as in the proof of Proposition 5.6. □

Remark 5.10. In §7 and §8, we will need to apply Proposition 5.1 to a set of boxes B that form a grid covering $\mathcal{Q}$. Since we need each B' to be contained in $\mathcal{Q}$, there is some care needed in placing the grid of boxes. In fact, for some radii r, this is actually impossible. There are several alternative solutions to this problem. The easiest solution is to restrict r to the set of radii that admit grids of boxes that cover $\mathcal{Q}$ well. This happens, for example, when r divides R. However, this solution has the drawback of not being easily adaptable to other settings, for example, to the setting in which $\mathcal{Q}$ is a rectangle or some smooth perturbation of a rectangle. For this reason, we now describe a somewhat different solution. Let V be a maximal set of points in $\mathcal{Q}$ such that the distance between any pair of distinct points in V is at least r and the distance between any $v\in V$ to the closest point on $\partial\mathcal{Q}$ is at least r. Consider the intrinsic metric $d_{\mathcal{Q}}$ on $\mathcal{Q}$, where $d_{\mathcal{Q}}(x, x')$ is the infimum length of any curve in $\mathcal{Q}$ connecting x and x'. Let $(T_v : v\in V)$ denote the Voronoi tiling associated with V and with this metric, and let B_v denote the union of the lattice tiles meeting T_v. If we assume that $\mathcal{Q}$ is "reasonably nice", then the maximal $d_{\mathcal{Q}}$-diameter of any B_v is $O(r)$. (Of course, we assume that $r>1$.) This will be the case, for example, when $\mathcal{Q}$ is a rectangle whose smaller side length is larger than $2r$, or more generally, if $\mathcal{Q}=R\mathcal{Q}_0$ for a piecewise smooth quad $\mathcal{Q}_0$ and $R>c(\mathcal{Q}_0)r$. Now observe that the disk of radius $\frac{1}{4}r$ around each $v\in V$ is contained in the interior of T_v and is bounded away from $\partial\mathcal{Q}$. We may define B'_v as the union of the lattice tiles that meet this disk. The statement and proof of Proposition 5.1 hold with B_v and B'_v

replacing B and B', though the constants will depend on the upper bound we have for $\operatorname{diam}(B_v)/r$.

5.6. A local result for general quads

In this subsection, we prove the following local result, which is a key step in estimates for noise sensitivity in the case of general quads.

PROPOSITION 5.11. *Let $\mathcal{Q}\subset\mathbb{R}^2$ be some quad, and let U be an open set whose closure is contained in the interior of $\mathcal{Q}$. For $R>0$, let $\mathcal{S}:=\mathcal{S}_{f_{R\mathcal{Q}}}$ be the spectral sample of $f_{R\mathcal{Q}}$, the ±1 indicator function for the crossing event in $R\mathcal{Q}$. Then, there is a constant $\bar{r}=\bar{r}(U,\mathcal{Q})$ such that for any box $B\subset RU$ of radius $r\in[\bar{r},R\operatorname{diam}(U)]$ and any set W with $W\cap B=\varnothing$, we have*

$$\mathbb{P}[\mathcal{S}_{f_{R\mathcal{Q}}}\cap B'\cap\mathcal{Z}\neq\varnothing\,|\,\mathcal{S}_{f_{R\mathcal{Q}}}\cap B\neq\varnothing,\mathcal{S}_{f_{R\mathcal{Q}}}\cap W=\varnothing]\geqslant a(U,\mathcal{Q}),$$

where B' is concentric with B and has radius $\frac{1}{3}r$, the random set $\mathcal{Z}$ is defined as in Proposition 5.1, and $a(U,\mathcal{Q})>0$ is a constant that depends only on U and $\mathcal{Q}$.

Proof. Here, the main new issue to deal with is that the quad $\mathcal{Q}$ is general; but, in contrast to the situation in Proposition 5.2, the box B is bounded away from $\partial\mathcal{Q}$, which simplifies parts of the proof.

The second-moment estimate (Proposition 5.3) applies in the present setup. We now prove the corresponding first-moment estimate.

In the following discussion, the constants are allowed to depend on U and $\mathcal{Q}$. We start by proving the analogue of (5.20). Let K be a compact set contained in the interior of $\mathcal{Q}$ and containing the closure of U in its interior. Let M denote the set of all squares $S\subseteq K$ that intersect $\overline{U}$ while the concentric square of twice the radius is not contained in the interior of K. Then M is a compact set of squares in the natural topology, and the radius of the squares in M is bounded away from zero. Fix some $S\in M$. Let $B(RS)$ denote the union of the lattice tiles that meet RS. A simple application of the Russo–Seymour–Welsh theorem shows that

$$\liminf_{R\to\infty}\mathbb{P}[\omega',\omega''\in\mathscr{A}_{\square}(B(RS),R\mathcal{Q})]>0.$$

Moreover, the same estimate holds in a neighborhood of S; that is, there is a set $V\subset M$ that contains S and is open in the topology of M, and there is a constant $R_0=R_0(V,U,\mathcal{Q})>0$ such that

$$\inf_{R\geqslant R_0}\inf_{S'\in V}\mathbb{P}[\omega',\omega''\in\mathscr{A}_{\square}(B(RS'),R\mathcal{Q})]>0.$$

Since M is compact, this cover of M by open subsets V has a finite subcover, and therefore there is some constant $R_1=R_1(U,\mathcal{Q})$ such that

$$\inf_{R\geqslant R_1}\inf_{S\in M}\mathbb{P}[\omega',\omega''\in\mathscr{A}_\square(B(RS),R\mathcal{Q})]>0.$$

It is clear that there is some constant $b>1$ such that for every $S\in M$ the concentric square whose radius is b times the radius of S is still contained in $\mathcal{Q}$. The above then shows that there is a constant $\delta>0$ such that for all $R\geqslant R_1$ and all $S\in M$,

$$\mathbb{P}[\omega',\omega''\in\mathscr{A}_\square(B(RS),R\mathcal{Q})]\geqslant\delta\mathbb{P}[\omega',\omega''\in\mathscr{A}_\square(B(RS_b),R\mathcal{Q})],\tag{5.22}$$

where S_b denotes the square concentric with S whose radius is b times the radius of S. Let $\widehat{M}$ denote the set of squares that are contained in and concentric with some square in M. Once we have (5.22) for all $S\in M$, we can conclude, as in the proof of (5.20) in Lemma 5.8, that the same holds with possibly a different constant δ for every $S\in\widehat{M}$ such that $\operatorname{diam}(RS)\geqslant\bar{r}$, for some constant $\bar{r}>0$. For this, the powers of 2 that were used in the proof of (5.19) and (5.20) (for example, for the definition of the notion of "good") need to be replaced by powers of b, but this is of little consequence. We also need here a version of Lemma 5.7 for the events $\mathscr{A}_\square(B(RS_b),R\mathcal{Q})$, but that can be proved in the same way as the original version, using [SSt, Lemmas A.2 and A.3] and (5.22). Finally, the restriction that $R\geqslant R_1$ may be avoided by taking $\bar{r}$ sufficiently large. Thus, the analogue of (5.20) is established.

Based on (5.19) and the above analogue of (5.20), we obtain the analogue of (5.10) for the current setup, yielding the first-moment estimate. (Note that we do not need to adapt the outward bound (5.19) to this local result.) The proof of the current proposition from the first and second moment estimates follows as in the proof of Proposition 5.1. □

5.7. The radial case

For the study of the set of exceptional times for dynamical percolation, we will need some tightness for the spectral samples of the "radial" indicator function. For this purpose, the following analogue of Proposition 5.1 for the radial setting will be useful.

PROPOSITION 5.12. *Let $f=f_R$ be the* 0-1 *indicator function of the existence of a white crossing between the two boundary components of the annulus $[-R,R]^2\setminus[-1,1]^2$, and let $\mathcal{S}=\mathcal{S}_f$ be its spectral sample with law $\mathbb{P}[\mathcal{S}=S]=\hat{f}(S)^2/\|f\|^2$. Also let $W\subseteq\mathcal{I}$. Let B be a box of some radius r that does not intersect W and let B' be the concentric box with radius $\frac{1}{3}r$. Suppose that $B'\subset[-R,R]^2$ and $B\cap[-4r,4r]^2=\varnothing$. We also assume that $r\geqslant\bar{r}$, where $\bar{r}>0$ is some universal constant. Then*

$$\mathbb{P}[\mathcal{S}\cap B'\cap\mathcal{Z}\neq\varnothing\,|\,\mathcal{S}\cap B\neq\varnothing,\mathcal{S}\cap W=\varnothing]>a,$$

where $\mathcal{Z}$ is as in Proposition 5.1 *and $a>0$ is a universal constant.*

Proof. The proof is similar to the above proofs. For this reason, we will be brief and leave many details to the reader. Let z be the center of the box B and set $r_1:=|z|>4r$. Assume first that $r_1<\frac{1}{3}R$. We then consider the three annuli

$$B\big(0,\tfrac{1}{3}r_1\big)\setminus B(0,1), \quad B(0,R)\setminus B(0,3r_1) \quad \text{and} \quad B\big(z,\tfrac{1}{3}r_1\big)\setminus B.$$

In order for Λ_B to hold, it is necessary that the 1-arm event occurs in the first two annuli and that the 4-arm event occurs in the third annulus. Thus, by Lemma 2.2,

$$\begin{aligned}\mathbb{Q}[B\cap\mathcal{S}\neq\varnothing=\mathcal{S}\cap W] &\leqslant 4\mathbb{E}[\lambda_{B,W}^2]\\ &\leqslant 4\mathbb{P}\big[\omega',\omega''\in\mathcal{A}_1\big(B(0,1),B\big(0,\tfrac{1}{3}r_1\big)\big)\big]\\ &\quad\times\mathbb{P}[\omega',\omega''\in\mathcal{A}_1(B(0,3r_1),B(0,R))]\\ &\quad\times\mathbb{P}\big[\omega',\omega''\in\mathcal{A}_4\big(B,B\big(z,\tfrac{1}{3}r_1\big)\big)\big].\end{aligned} \tag{5.23}$$

Now, using the same techniques of quasi-multiplicativity for coupled configurations and the separation of arms as we did before (starting with the argument of §5.3), one can prove the following first-moment estimate for any $x\in B'$:

$$\begin{aligned}\mathbb{Q}[x\in\mathcal{S},\mathcal{S}\cap W=\varnothing] &\geqslant c_1\beta_1^W\big(1,\tfrac{1}{3}r_1\big)\beta_4^W\big(x,\tfrac{1}{3}r_1\big)\beta_1^W(3r_1,R)\\ &\geqslant c_2\beta_1^W\big(1,\tfrac{1}{3}r_1\big)\alpha_4(r)\beta_4^W\big(r,\tfrac{1}{3}r_1\big)\beta_1^W(3r_1,R)\\ &\geqslant c_3\alpha_4(r)\mathbb{E}[\lambda_{B,W}^2],\end{aligned} \tag{5.24}$$

where (5.23) is used for the last step. One can easily prove the analogous second-moment estimate, and the claim now follows as in the proof of Proposition 5.1 (it is the same second-moment type of argument, except that one has to renormalize the measure $\mathbb{Q}$ to get the probability measure $\mathbb{P}=\mathbb{P}_{f_R}$).

Suppose now that $r_1>\frac{2}{3}R$. In this case, we need to consider a different system of annuli. Let d denote the distance from B to $\partial B(0,R)$ and let z' denote a closest point to B on $\partial B(0,R)$. In the annulus $B\big(0,\frac{1}{3}R\big)\setminus B(0,1)$ we consider the 1-arm event, in the annulus $B\big(z,r+\frac{1}{2}d\big)\setminus B$ we consider the 4-arm event, and in the intersection of $B(0,R)$ with $B\big(z',\frac{1}{3}R\big)\setminus B(z',5r+d)$ (assuming that this is non-empty), we consider the 3-arm event between $\partial B\big(z',\frac{1}{3}R\big)$ and $\partial B(z',5r+d)$. Again, the claim follows.

In the intermediate case $\frac{1}{3}R\leqslant r_1\leqslant\frac{2}{3}R$, we need to consider the 1-arm event in $B\big(0,\frac{1}{6}R\big)\setminus B(0,1)$ and the 4-arm event in $B\big(z,\frac{1}{6}R\big)\setminus B$, and the claim likewise follows. □

6. A large deviation result

In order to deduce Theorem 1.1 from the results of §5 and §4, we will need the following general result. Let us note that not only the statement bears a vague resemblance to [LSS], as explained in §1.6, but also the proof method of averaging using an independent random sample $I \subseteq [n]$ does.

PROPOSITION 6.1. *Let $n \in \mathbb{N}_+$, let x and y be random variables in $\{0,1\}^n$ and set $X := \sum_{j=1}^n x_j$ and $Y := \sum_{j=1}^n y_j$. Suppose that a.s. $y_i \leqslant x_i$ for each $i \in [n]$ and that there is a constant $a \in (0,1]$ such that for each $j \in [n]$ and every $I \subset [n] \setminus \{j\}$ we have*

$$\mathbb{P}[y_j = 1 \mid y_i = 0 \text{ for all } i \in I] \geqslant a \mathbb{P}[x_j = 1 \mid y_i = 0 \text{ for all } i \in I]. \tag{6.1}$$

Then

$$\mathbb{P}[Y = 0 \mid X > 0] \leqslant a^{-1} \mathbb{E}[e^{-aX/e} \mid X > 0]. \tag{6.2}$$

For completeness, we will also show below that

$$\mathbb{P}[Y \leqslant t] \leqslant \mathbb{P}\Big[X < \frac{e}{a}(t+s)\Big] + \frac{e^{t-1}}{s} \mathbb{E}[e^{-aX/e}] \tag{6.3}$$

holds for every $t \geqslant 0$ and $s > 0$. However, we do not have an application for this inequality.

Proof. We may write our assumption (6.1) as follows:

$$\mathbb{P}[y_j = 1, y_i = 0 \text{ for all } i \in I] \geqslant a \mathbb{P}[x_j = 1, y_i = 0 \text{ for all } i \in I] 1_{\{j \notin I\}}, \tag{6.4}$$

which is now true even when $j \in I$: it simply says $0 \geqslant a0$. This gives us many inequalities, which we will average out in a useful manner. Fix $\lambda \in (0,1)$. Now multiply (6.4) by $\lambda^{n-|I|}(1-\lambda)^{|I|}$; we now think of I as a random subset of $[n]$, where each $i \in [n]$ is in I independently with probability $1-\lambda$. We will use $\mathbb{P}^I$ for this independent extra randomness. Summing (6.4) over all choices of $j \in [n]$ and $I \subset [n]$, one gets, for the left-hand side,

$$\sum_{j,I} \lambda^{n-|I|}(1-\lambda)^{|I|} \mathbb{P}[y_j = 1, y_i = 0 \text{ for all } i \in I] = \mathbb{E}[Y \mathbb{P}^I[y_i = 0 \text{ for all } i \in I]] = \mathbb{E}[Y \lambda^Y].$$

The right-hand side of (6.4) after summing over all choices of $j \in [n]$ and $I \subset [n]$ gives

$$\sum_{j,I} \lambda^{n-|I|}(1-\lambda)^{|I|} 1_{\{j \notin I\}} \mathbb{P}[x_j = 1, y_i = 0 \text{ for all } i \in I] = \mathbb{E}\Bigg[\sum_{j : x_j = 1} \mathbb{P}^I[j \notin I, y_i = 0 \text{ for all } i \in I]\Bigg].$$

Here, the random set I not only has to avoid the points i with $y_i=1$, but also the point j; hence, depending on whether $y_j=1$ or not, I has to avoid Y or $Y+1$ points. Therefore, $\mathbb{E}[X\lambda^{Y+1}]$ is a lower bound for the last displayed quantity. Summarizing these computations, one ends up with

$$\mathbb{E}[Y\lambda^Y] \geqslant a\mathbb{E}[X\lambda^{Y+1}].$$

This may be rewritten as $\mathbb{E}[Z]\geqslant 0$, where $Z:=(Y-a\lambda X)\lambda^Y$. At this point, we choose $\lambda:=e^{-1}$. In order to bound Z from above by a function of X only, we maximize Z over Y, and get the bound $Z\leqslant\exp(-1-aX/e)$. On $X=0$, we also have $Y=0$ and $Z=0$, while on $Y=0<X$, we have $Z\leqslant -a/e$. Therefore, $\mathbb{E}[Z]\geqslant 0$ gives

$$\frac{a}{e}\mathbb{P}[Y=0<X] \leqslant \mathbb{E}\left[1_{X>0}\exp\left(-1-\frac{aX}{e}\right)\right].$$

Dividing by $(a/e)\mathbb{P}[X>0]$, we obtain (6.2). □

We now prove (6.3). Set $r:=(e/a)(t+s)$, $Z_+:=\max\{Z,0\}$ and $Z_-:=Z-Z_+$. Note that on the event $\{X\geqslant r, Y\leqslant t\}$ we have $Z\leqslant -se^{-t}$. Hence,

$$\mathbb{E}[Z_-] \leqslant -se^{-t}\mathbb{P}[X\geqslant r, Y\leqslant t] \leqslant -se^{-t}(\mathbb{P}[Y\leqslant t]-\mathbb{P}[X<r]).$$

On the other hand, $\mathbb{E}[Z_+]\leqslant\mathbb{E}[\exp(-1-aX/e)]$. Since $0\leqslant\mathbb{E}[Z]=\mathbb{E}[Z_+]+\mathbb{E}[Z_-]$, (6.3) follows.

7. The lower tail of the spectrum

In this section, we prove Theorem 1.1 and a few related results.

7.1. Local version

We start with a version which avoids the issues involving the boundary, as we did in §4.2 and §5.6. The bound we give here is sharp up to a constant factor, but we will not need the lower bound, so will prove sharpness only in the square case (with boundary), in §7.2.

THEOREM 7.1. *Consider some quad $\mathcal{Q}$ and let $\mathcal{S}=\mathcal{S}_{f_{R\mathcal{Q}}}$ be the spectral sample of $f_{R\mathcal{Q}}$, the ±1-indicator function for the crossing event in $R\mathcal{Q}$. Let $U\subset\mathcal{Q}$ be open, and let $U'\subset\overline{U}'\subset U$. Then, for some constants $\bar{r}=\bar{r}(U',U,\mathcal{Q})>0$ and $q(U',U,\mathcal{Q})>0$, for any $r\in[\bar{r},R\,\mathrm{diam}(U)]$,*

$$\mathbb{P}[0<|\mathcal{S}_{f_{R\mathcal{Q}}}\cap RU|\leqslant r^2\alpha_4(r), \mathcal{S}_{f_{R\mathcal{Q}}}\cap RU\subset RU'] \leqslant q(U',U,\mathcal{Q})\frac{R^2\alpha_4(R)^2}{r^2\alpha_4(r)^2}. \tag{7.1}$$

Proof. Let the distance between U' and the complement of U be $\delta>0$. Then, with no loss of generality, we may assume that $r\leqslant\delta\frac{1}{10}R$. (Otherwise, r/R remains bounded away from 0, so, by choosing $q(U',U,\mathcal{Q})$ large enough compared to δ, the upper bound in (7.1) becomes larger than 1, and we are done.) Consider the tiling of the plane by $r\times r$ squares given by the grid $r\mathbb{Z}^2$, recall from the end of §2.1 that each square gives a box that together form a tiling of the plane, and let $\{B_1,B_2,...,B_n\}$ be the set of those boxes that intersect RU'. Let $U''\subset U$ be such that $RU''\supset\bigcup_{j=1}^n B_j$ but the distance of U'' to the complement of U is at least $\frac{1}{2}\delta$; one can choose a U'' that works for all $r\leqslant\frac{1}{10}\delta R$ at the same time. Let $\mathcal{Z}$ be a subset of $\mathcal{I}\cap RU''$, where each bit $i\in RU''\cap\mathcal{I}$ is in $\mathcal{Z}$ with probability $1/r^2\alpha_4(r)$, independently from each other and from $\mathcal{S}$. Let y_j be the indicator function of the event

$$\mathcal{S}\cap B_j\cap\mathcal{Z}\neq\varnothing,$$

and let x_j be the indicator function of the event $\mathcal{S}\cap B_j\neq\varnothing$. As in Propositions 5.1 and 5.11, let $\mathbb{P}$ denote the law of $\mathcal{S}$ coupled with the independent point process $\mathcal{Z}$. Let $\widetilde{\mathbb{P}}$ denote the law $\mathbb{P}$ (on $(\mathcal{S},\mathcal{Z})$) conditioned on the event $\mathcal{S}\cap(RU\setminus RU'')=\varnothing$, and let $\widetilde{\mathbb{E}}$ denote the corresponding expectation operator. For every $I\subseteq\{1,...,n\}$ and every $j\in\{1,...,n\}\setminus I$, by applying Proposition 5.11 to U'' (in place of U there) and $\mathcal{Q}$, we get

$$\widetilde{\mathbb{P}}[y_j=1\,|\,y_i=0\text{ for all }i\in I]\geqslant a\widetilde{\mathbb{P}}[x_j=1\,|\,y_i=0\text{ for all }i\in I]$$

for some constant $a=a(U'',\mathcal{Q})>0$. (Technically, in order to apply Proposition 5.11 here, one has to first condition on the values of $\mathcal{Z}$ on the boxes in I. Furthermore, in order to obtain $\widetilde{\mathbb{P}}$ instead of $\mathbb{P}$, before applying Proposition 5.11, one needs to add the set $RU\setminus RU''$ to the set W.) Therefore, the large deviation result Proposition 6.1 gives (using that $RU'\subseteq\bigcup_i B_i$) that

$$\widetilde{\mathbb{P}}[\mathcal{S}\cap\mathcal{Z}=\varnothing\neq\mathcal{S}\cap RU']\leqslant a^{-1}\widetilde{\mathbb{E}}[e^{-aX/e}1_{X>0}],$$

where $X:=|\{j:\mathcal{S}\cap B_j\neq\varnothing\}|$. This yields

$$\mathbb{P}[\mathcal{S}\cap\mathcal{Z}=\varnothing\neq\mathcal{S}\cap RU\subset RU']\leqslant a^{-1}\sum_{k=1}^{\infty}e^{-ak/e}\mathbb{P}[X=k,\mathcal{S}\cap RU\subset RU''].$$

We estimate the terms $\mathbb{P}[X=k,\mathcal{S}\cap RU\subset RU'']$ using Proposition 4.2, and get the bound

$$\mathbb{P}[\mathcal{S}\cap\mathcal{Z}=\varnothing\neq\mathcal{S}\cap RU\subset RU']\leqslant O(1)\sum_{k=1}^{\infty}e^{-ak/e}g(k)\gamma_r(R)=O(1)\gamma_r(R),\tag{7.2}$$

where g is as defined in the proposition and the constants implied by the $O(1)$ terms may depend on U', U and $\mathcal{Q}$.

Now, by the choice of $\mathcal{Z}$, for $|\mathcal{S}\cap RU'|\leqslant r^2\alpha_4(r)$ we have

$$\mathbb{P}[\mathcal{S}\cap\mathcal{Z}\cap RU'=\varnothing\,|\,\mathcal{S}]=\left(1-\frac{1}{r^2\alpha_4(r)}\right)^{|\mathcal{S}\cap RU'|}\geqslant c$$

for some absolute constant $c>0$, and hence

$$\begin{aligned}c\mathbb{P}[|\mathcal{S}\cap RU'|\leqslant r^2\alpha_4(r),\varnothing\neq\mathcal{S}\cap RU\subset RU']&\leqslant\mathbb{P}[\mathcal{S}\cap\mathcal{Z}=\varnothing\neq\mathcal{S}\cap RU\subset RU']\\&\leqslant O(1)\gamma_r(R),\end{aligned}$$

by (7.2). This proves (7.1). □

7.2. Square version

We prepare for the proof of Theorem 1.1 by first showing that

$$\mathbb{E}|\mathcal{S}_{f_R}|\asymp R^2\alpha_4(R),\tag{7.3}$$

which will be needed for the (easier) lower bound. Note that this also implies (1.6) for the triangular lattice, by quasi-multiplicativity and [SW].

First, the lower bound on $\mathbb{E}|\mathcal{S}_{f_R}|$ follows immediately from Lemma 3.1. For the upper bound, we will need to consider the half-plane 3-arm events and the quarter-plane 2-arm events that were discussed in §4.3. Let $\mathcal{Q}=[0,R]^2$, $f=f_{\mathcal{Q}}$ and $\mathcal{S}=\mathcal{S}_f$. Let $x\in\mathcal{I}$ be an input bit of f. If x is at distance r_0 from the closest edge of $[0,R]^2$, and at distance r_1 from the closest corner, then by (3.6) and quasi-multiplicativity, we have

$$\mathbb{P}[x\in\mathcal{S}]=\alpha_\square(x,\mathcal{Q})\asymp\alpha_4(r_0)\alpha_3^+(r_0,r_1)\alpha_2^{++}(r_1,R).$$

Now observe that $\alpha_3^+(r_0,r_1)\leqslant O(1)\alpha_4(r_0,r_1)$ follows from (4.12), and (2.6). Thus,

$$\alpha_4(r_0)\alpha_3^+(r_0,r_1)\leqslant O(1)\alpha_4(r_1).$$

Moreover, $\alpha_2^{++}(r_1,R)\leqslant O(r_1/R)$, by (4.13) and (4.12). (For the triangular lattice, we have $\alpha_2^{++}(r_1,R)\leqslant\alpha_4(r_1,R)$ from (4.14) and (2.5), and hence we get $\mathbb{P}[x\in\mathcal{S}]\leqslant O(1)\alpha_4(R)$, which proves (7.3) immediately.)

Since the number of $x\in\mathcal{I}$ with $r_1\in[2^j,2^{j+1})$ is $O(2^{2j})$, we get

$$\begin{aligned}\mathbb{E}|\mathcal{S}|&=\sum_{x\in\mathcal{I}}\mathbb{P}[x\in\mathcal{S}]\leqslant\sum_{j=0}^{\lceil\log_2 R\rceil}O(2^{2j})\alpha_4(2^j)\frac{2^j}{R}\\&=\frac{\alpha_4(R)}{R}\sum_{j=0}^{\lceil\log_2 R\rceil}\frac{O(2^{3j})}{\alpha_4(2^j,R)}\overset{(2.6)}{\leqslant}\frac{\alpha_4(R)}{R}\sum_{j=0}^{\lceil\log_2 R\rceil}\frac{O(2^{3j})}{(2^j/R)^2}=O(R^2)\alpha_4(R).\end{aligned}$$

Thus, we get (7.3).

Proof of Theorem 1.1. The proof of

$$\mathbb{P}[0<|\mathscr{S}|<r^2\alpha_4(r)]\leqslant O(1)\gamma_r(R), \tag{7.4}$$

for $1\leqslant r\leqslant R$, is very similar to the proof of Theorem 7.1, with some small modifications, which we now discuss. Note that the set of $x\in\mathcal{I}$ that are relevant for f are all within distance at most 2 from $[0,R]^2$. Note that we may assume, without loss of generality, that $r\geqslant\bar{r}$ for some absolute constant $\bar{r}$, and that $(R+4)/r\in\mathbb{N}_+$. Let $\{B_1,B_2,...,B_n\}$ be the set of boxes corresponding to the tiling of $[-2,R+2]^2$ by $r\times r$ squares. In the proof of Theorem 7.1 we are now allowed to take $RU=RU'=RU''=[-2,R+2]^2$, by replacing the appeals to Propositions 5.11 and 4.2 with appeals to Propositions 5.1 and 4.1, respectively. (In particular, we do not need to introduce the measure $\widetilde{\mathbb{P}}$ now.) This gives (7.4).

We now show that the inequality in (7.4) is actually an equality up to constants. Let $\mathcal{N}$ be the set of indices i such that the r-box B_i is at least at distance $\frac{1}{10}R$ from the boundary $\partial[0,R]^2$. Consider the events

$$V_i:=\{|\mathscr{S}\cap B_i|\geqslant Cr^2\alpha_4(r)\}\quad\text{and}\quad W_i:=\{\mathscr{S}\cap B_i\neq\varnothing,\mathscr{S}\subseteq B_i\}$$

for $i\in\mathcal{N}$. We claim that we may take the constant C large enough so that $\mathbb{P}[V_i|W_i]\leqslant\frac{1}{2}$. This will follow from Markov's inequality, once we know that

$$\mathbb{E}\big[|\mathscr{S}\cap B_i|\,\big|\,W_i\big]\leqslant O(1)r^2\alpha_4(r). \tag{7.5}$$

To prove (7.5), first observe that for each $i\in\mathcal{N}$,

$$\mathbb{P}[W_i]\overset{(5.7)}{\asymp}\alpha_\square(B_i,[0,R]^2)^2\overset{(2.4)}{\asymp}\alpha_4(r,R)^2.$$

Then, we need a good upper bound on $\mathbb{P}[x\in\mathscr{S},\mathscr{S}\subseteq B_i]$. We know that this equals $\mathbb{E}[\lambda^2_{x,B_i^c}]$, see, e.g., §5.3. Similarly to the proof of Lemma 3.2, or following §5.3, one can easily show that this is at most $\alpha_4(x,B_i)\alpha_\square(B_i,[0,R]^2)^2$. Summing up over all $x\in B_i$, and using the above estimate on $\mathbb{P}[W_i]$ and a computation similar to (7.3), we get (7.5).

So, we have $\mathbb{P}[V_i^c|W_i]\geqslant\frac{1}{2}$. Note that the events $V_i^c\cap W_i$ for different i's are disjoint, hence

$$\mathbb{P}[0<|\mathscr{S}|\leqslant Cr^2\alpha_4(r)]\geqslant\sum_{i\in\mathcal{N}}\mathbb{P}[V_i^c\cap W_i]\geqslant c\left(\frac{R}{r}\right)^2\alpha_4(r,R)^2$$

for some $c>0$, and the lower bound is proved. □

Remark 7.2. For the triangular lattice, the following variant of (1.7) may also be established:

$$\begin{aligned}\limsup_{R\to\infty}\mathbb{P}[0<|\mathscr{S}_{f_R}|\leqslant\lambda\mathbb{E}|\mathscr{S}_{f_R}|]&\asymp\lambda^{2/3},\\ \liminf_{R\to\infty}\mathbb{P}[0<|\mathscr{S}_{f_R}|\leqslant\lambda\mathbb{E}|\mathscr{S}_{f_R}|]&\asymp\lambda^{2/3},\end{aligned}\tag{7.6}$$

holds for every $\lambda\in(0,1]$, where the implied constants do not depend on λ. In view of Theorem 1.1, this follows from the fact that

$$\lim_{R\to\infty}\alpha_4(tR,R)\asymp t^{5/4},\quad t\in(0,1],$$

which holds since the limit of critical percolation is described by SLE_6 (this is explained in [SW]), and the probabilities for the corresponding events for SLE are determined up to constant factors [LSW1] and have no lower order corrections to the power law.

7.3. Radial version

We also have the following radial version, where $\mathscr{S}$ is the spectral sample of the 0-1 indicator function f of the crossing event from $\partial[-1,1]^2$ to $\partial[-R,R]^2$, so that $\mathbb{E}[f^2]\asymp\alpha_1(R)$. Recall that we have the measures $\mathbb{P}[\mathscr{S}=S]=\mathbb{Q}\big[\mathscr{S}=S\big]/\mathbb{E}[f^2]=\hat{f}(S)^2/\mathbb{E}[f^2]$.

THEOREM 7.3. *Let $\mathscr{S}$ be as above, and let $r\in[1,R]$. Then*

$$\mathbb{Q}[|\mathscr{S}|<\alpha_4(r)r^2]\leqslant O(1)\frac{\alpha_1(R)^2}{\alpha_1(r)}\quad\text{and}\quad\mathbb{P}[|\mathscr{S}|<\alpha_4(r)r^2]\leqslant O(1)\frac{\alpha_1(R)}{\alpha_1(r)}.$$

Proof. Again, the bits relevant for f are contained in $[-R',R']^2$, where $R'=R+2$. We may assume that r is such that $R'/r\in\mathbb{N}_+$ and $r\in\big[\bar{r},\frac{1}{8}R\big]$ for some fixed constant $\bar{r}>0$, which guarantees that $k:=\alpha_4(r)r^2>1$. Take a subdivision of $[-R',R']^2$ into boxes B_j of side length r, and let K denote the union of the boxes that intersect $[-4r,4r]^2$. We now let $\mathscr{Z}$ be the random set in $\mathcal{I}\cap[-R',R']^2\setminus K$ where each bit is in $\mathscr{Z}$ with probability $1/k$, and $\mathscr{Z}$ is independent of $\mathscr{S}$. We also let $X:=|\{j:\mathscr{S}\cap B_j\neq\varnothing,B_j\not\subset K\}|=|(\mathscr{S}\setminus K)_r|$. Note that $|\mathscr{S}_r|-X$ is bounded from above by the number of boxes in K, which is bounded by a constant.

Exactly as before, Propositions 5.12 and 6.1 give that

$$\mathbb{P}[\mathscr{S}\cap\mathscr{Z}=\varnothing\neq\mathscr{S}\setminus K]\leqslant a^{-1}\mathbb{E}[e^{-aX/e}1_{X>0}],$$

with some absolute constant $a>0$. We can use Proposition 4.7 to bound each $\mathbb{P}[X=n]$, $n\in\mathbb{N}_+$, and the argument that finished the proof of Theorem 7.1 above now gives

$$\mathbb{P}[0<|\mathscr{S}\setminus K|<k]\leqslant O(1)\mathbb{P}[\mathscr{S}\cap\mathscr{Z}=\varnothing\neq\mathscr{S}\setminus K]\leqslant O(1)\alpha_1(r,R).$$

(The situation is more similar to §7.2 than to §7.1 in that the boundary issues are already dealt with in Propositions 5.12 and 4.7, hence we do not need U'' and the measure $\widetilde{\mathbb{P}}$.) Finally, observe that

$$\begin{aligned}\mathbb{P}[|\mathscr{S}|<k] &\leqslant \mathbb{P}[\mathscr{S}\subset K]+\mathbb{P}[0<|\mathscr{S}\setminus K|<k]\\ &\leqslant \mathbb{P}[\mathscr{S}^* \text{ is compatible with } [-R',R']^2\setminus K]+O(1)\alpha_1(r,R)\\ &\leqslant O(1)\alpha_1(r,R)+O(1)\alpha_1(r,R),\end{aligned}$$

by Lemma 4.8, and the theorem is proved. □

7.4. Tightness of the quad spectral sample

This section will be devoted to the following analog of (1.3) in the setting of more general quads, showing that the appropriately normalized spectral sample is tight.

THEOREM 7.4. *Let $\mathcal{Q}\subset\mathbb{R}^2$ be a quad and for $R>0$ let $\mathscr{S}_{f_{R\mathcal{Q}}}$ denote the spectral sample of $f_{R\mathcal{Q}}$, the ±1-indicator function of the crossing event of $R\mathcal{Q}$. Then*

$$\lim_{t\to\infty}\inf_{R>1}\mathbb{P}[|\mathscr{S}_{f_{R\mathcal{Q}}}|\in[t^{-1}R^2\alpha_4(R),tR^2\alpha_4(R)]\cup\{0\}]=1.$$

We do not presently prove that $\mathbb{E}|\mathscr{S}_{f_{R\mathcal{Q}}}|\asymp R^2\alpha_4(R)$ as $R\to\infty$, though we tend to believe that this holds.

The main technical difficulty in the case of a general quad $\mathcal{Q}$ compared to a square is the boundary: our explicit computations in §4.3 do not apply to a general quad (even if it has piecewise smooth boundary). Thus, the proof will begin by showing that even in a general quad the spectral sample is unlikely to be very close to $\partial\mathcal{Q}$.

Proof. For every fixed $\delta>0$ we can find a quad $\mathcal{Q}'$ that is contained in the interior of $\mathcal{Q}$ and such that

$$\limsup_{R\to\infty}\mathbb{P}[f_{R\mathcal{Q}}\neq f_{R\mathcal{Q}'}]<\delta. \tag{7.7}$$

This is easy to see, and also worked out in detail in [SS]. Let $U',U\subset\mathcal{Q}$ be open sets satisfying $\mathcal{Q}'\subset U'\subset\overline{U}'\subset U$.

Now, (7.7) and (2.7) imply that for all large enough R, the laws of the spectral samples $\mathscr{S}_{f_{R\mathcal{Q}}}$ and $\mathscr{S}_{f_{R\mathcal{Q}'}}$ have a total variation distance at most $4\sqrt{\delta}$, and

$$\mathbb{P}[\mathscr{S}_{f_{R\mathcal{Q}}}\subseteq RU']\geqslant 1-4\sqrt{\delta}. \tag{7.8}$$

Theorem 7.1 can now be invoked to get

$$\lim_{t\to\infty}\limsup_{R\to\infty}\mathbb{P}[\mathscr{S}_{f_{R\mathcal{Q}}}\subseteq RU',0<|\mathscr{S}_{f_{R\mathcal{Q}}}|<t^{-1}R^2\alpha_4(R)]=0.$$

In conjunction with (7.8), this gives

$$\limsup_{t\to\infty}\limsup_{R\to\infty}\mathbb{P}[0<|\mathscr{S}_{f_{R\mathcal{Q}}}|<t^{-1}R^2\alpha_4(R)]\leqslant 4\sqrt{\delta},$$

and since δ was an arbitrary positive number,

$$\lim_{t\to\infty}\limsup_{R\to\infty}\mathbb{P}[0<|\mathscr{S}_{f_{R\mathcal{Q}}}|<t^{-1}R^2\alpha_4(R)]=0. \tag{7.9}$$

In the other direction, it is easy to see that $\mathbb{E}|\mathscr{S}_{f_{R\mathcal{Q}}}\cap RU'|=O(R^2)\alpha_4(R)$, as $R\to\infty$. Therefore, Markov's inequality and (7.8) imply that for all sufficiently large R,

$$\mathbb{P}[|\mathscr{S}_{f_{R\mathcal{Q}}}|>tR^2\alpha_4(R)]\leqslant 4\sqrt{\delta}+O\biggl(\frac{1}{t}\biggr),$$

where the implied constant may depend on δ but not on R. Thus,

$$\lim_{t\to\infty}\limsup_{R\to\infty}\mathbb{P}[|\mathscr{S}_{f_{R\mathcal{Q}}}|>tR^2\alpha_4(R)]=0. \tag{7.10}$$

Since for every $R_0\in(1,\infty)$ we obviously have

$$\lim_{t\to\infty}\sup_{R\in[1,R_0]}\mathbb{P}[|\mathscr{S}_{f_{R\mathcal{Q}}}|\notin[t^{-1}R^2\alpha_4(R),tR^2\alpha_4(R)]\cup\{0\}]=0,$$

the theorem follows immediately from (7.9) and (7.10). □

8. Applications to noise sensitivity

We are ready to prove Corollary 1.2 and Theorem 1.3, together with some generalizations.

8.1. Noise sensitivity in a square and a quad

We will now prove Corollary 1.2 and its generalization Corollary 8.1 to quad crossings. The proof of the latter will be quite simple using Theorem 7.4, and, actually, the same argument could be used to prove Corollary 1.2 from (1.3). Nevertheless, we are giving a longer proof of Corollary 1.2 that has the advantage of being more quantitative. In particular, it implies (8.7) below, and a variation on it will also be used in §9 to prove our quantitative dynamical sensitivity results.

Proof of Corollary 1.2. Let $\mathcal{I}_R$ denote the set of bits on which f_R depends, and write $\varepsilon=\varepsilon_R$. Recall from (1.2) that if y is an ε-noisy version of $x\in\{-1,+1\}^{\mathcal{I}_R}$, then

$$\Psi_R:=\mathbb{E}[f_R(y)f_R(x)]-\mathbb{E}[f_R(x)]^2=\sum_{k=1}^{|\mathcal{I}_R|}(1-\varepsilon)^k\mathbb{P}[|\mathscr{S}_{f_R}|=k]. \tag{8.1}$$

Breaking the sum over k in (8.1) into parts $(j-1)/\varepsilon<k\leqslant j/\varepsilon$, with $j=1,2,\dots$, we get

$$\Psi_R\leqslant\sum_{j=1}^{\infty}(1-\varepsilon)^{(j-1)/\varepsilon}\mathbb{P}\biggl[\frac{j-1}{\varepsilon}<|\mathscr{S}_{f_R}|\leqslant\frac{j}{\varepsilon}\biggr]\leqslant\sum_{j=1}^{\infty}e^{1-j}\mathbb{P}\biggl[0<|\mathscr{S}_{f_R}|\leqslant\frac{j}{\varepsilon}\biggr]. \tag{8.2}$$

Recall that $r^2\alpha_4(r)\to\infty$ as $r\to\infty$, by (2.6). For $s\geqslant1$, let $\varrho(s)$ be the least $r\in\mathbb{N}_+$ such that $r^2\alpha_4(r)\geqslant s$, and let $\gamma(r):=r^2\alpha_4(r)^2$. Since $\alpha_4(r+1)\leqslant\alpha_4(r)$, we have

$$(r+1)^2\alpha_4(r+1)\leqslant O(1)r^2\alpha_4(r),$$

and thus

$$s\leqslant\varrho(s)^2\alpha_4(\varrho(s))\leqslant O(s)\quad\text{for } s\geqslant1. \tag{8.3}$$

By (2.6), there exist $c_1,c_2>0$ such that $s^{c_1}/O(1)\leqslant\varrho(s)\leqslant O(1)s^{c_2}$. Substituting this and (2.3) into (8.3) gives

$$\frac{1}{O(1)}\biggl(\frac{s'}{s}\biggr)^{c_3}\leqslant\frac{\varrho(s')}{\varrho(s)}\leqslant O(1)\biggl(\frac{s'}{s}\biggr)^{c_4}\quad\text{for } s'\geqslant s\geqslant1, \tag{8.4}$$

with some $c_3,c_4>0$. This and (2.3) imply that

$$O(1)\frac{\gamma(\varrho(s'))}{\gamma(\varrho(s))}\geqslant\biggl(\frac{s}{s'}\biggr)^{O(1)}\quad\text{for } s'\geqslant s\geqslant1. \tag{8.5}$$

Now set $\varrho_j:=\varrho(j/\varepsilon)$. Then, for $j\in\mathbb{N}_+$,

$$\begin{aligned}\mathbb{P}\biggl[0<|\mathscr{S}_{f_R}|\leqslant\frac{j}{\varepsilon}\biggr]&\leqslant\mathbb{P}[0<|\mathscr{S}_{f_R}|\leqslant\varrho_j^2\alpha_4(\varrho_j)]\\&\leqslant O(1)\frac{\gamma(R)}{\gamma(\varrho_j)}&&\text{by (7.4)}\\&\leqslant O(1)j^{O(1)}\frac{\gamma(R)}{\gamma(\varrho_1)}&&\text{by (8.5).}\end{aligned}$$

Therefore (8.2) gives

$$\Psi_R\leqslant O(1)\frac{\gamma(R)}{\gamma(\varrho_1)}. \tag{8.6}$$

If $\lim_{R\to\infty}\mathbb{E}|\mathscr{S}_{f_R}|\varepsilon_R=\infty$, then by (7.3) and the usual properties of α_4 we have

$$\frac{R}{\varrho(1/\varepsilon)}\to\infty$$

as well as $\gamma(R)/\gamma(\varrho_1)=\gamma(R)/\gamma(\varrho(1/\varepsilon))\to0$, which together with (8.6) proves (1.8).

Now assume that $\varepsilon_R\mathbb{E}|\mathscr{S}_{f_R}|\to0$. Applying (1.2) and Jensen's inequality, we get

$$\mathbb{E}[f_R(x)f_R(y)]=\mathbb{E}\bigl[(1-\varepsilon)^{|\mathscr{S}_{f_R}|}\bigr]\geqslant(1-\varepsilon)^{\mathbb{E}|\mathscr{S}_{f_R}|}\to1,$$

as $R\to\infty$. Since $f_R(x)f_R(y)\leqslant1=f_R(x)^2$, (1.9) follows. □

Suppose that we are in the setting of the triangular grid, and $\varepsilon = t/\mathbb{E}|\mathscr{S}_{f_R}|$, where $t>1$. Then with the above notation, we have by (7.6) and (8.2) that

$$\limsup_{R\to\infty} \Psi_R \leqslant O(1)t^{-2/3}.$$

Using the fact that we also have lower bounds in Theorem 1.1, it is easy to see that

$$\limsup_{R\to\infty} \Psi_R \asymp t^{-2/3} \asymp \liminf_{R\to\infty} \Psi_R. \tag{8.7}$$

In a forthcoming paper we plan to use this to show that in the appropriate scaling limit of critical dynamical percolation (where both space and time are rescaled), the crossing events in the unit square at time 0 and at time t have correlations that decay like $t^{-2/3}$ as $t\to\infty$.

We also have the following generalization for the ±1-indicator function of the left-right crossing in scaled versions of an arbitrary fixed quad $\mathcal{Q}$.

COROLLARY 8.1. *Assume that $\varepsilon_R \in (0,1)$ is such that $\varepsilon_R R^2 \alpha_4(R)\to\infty$ as $R\to\infty$, and that y is an ε_R-noisy version of x. Then*

$$\mathbb{E}[f_{R\mathcal{Q}}(y) f_{R\mathcal{Q}}(x)] - \mathbb{E}[f_{R\mathcal{Q}}(x)]\mathbb{E}[f_{R\mathcal{Q}}(y)] \to 0.$$

On the other hand, if $\varepsilon_R R^2 \alpha_4(R)\to 0$, then

$$\mathbb{E}[f_{R\mathcal{Q}}(y) f_{R\mathcal{Q}}(x)] - \mathbb{E}[f_{R\mathcal{Q}}(x)^2] \to 0.$$

Proof. First assume that $\varepsilon_R R^2 \alpha_4(R)\to\infty$. Given $\delta>0$, by Theorem 7.4 we have

$$\sup_{R>1} \mathbb{P}[0 < |\mathscr{S}_{f_{R\mathcal{Q}}}| < t^{-1} R^2 \alpha_4(R)] < \delta$$

if $t \geqslant t_1(\delta)$ is large enough. Now let R be large enough so that $\varepsilon_R R^2 \alpha_4(R) > t^2$. Then, by (8.1),

$$\Psi_R \leqslant \delta + \sum_{|S| \geqslant R^2\alpha_4(R)/t} \hat{f}(S)^2 (1-\varepsilon_R)^{|S|} \leqslant \delta + \left(1 - \frac{t^2}{R^2\alpha_4(R)}\right)^{R^2\alpha_4(R)/t} \leqslant \delta + O(1)\exp(-t),$$

which is at most 2δ if $t \geqslant t_2(\delta)$. We can choose R large enough with respect to this new t, and hence $\Psi_R \to 0$ is proved.

Now assume that $\varepsilon_R R^2 \alpha_4(R)\to 0$. Given $\delta>0$, by Theorem 7.4 we have

$$\sup_{R>1} \mathbb{P}[|\mathscr{S}_{f_{R\mathcal{Q}}}| > t R^2 \alpha_4(R)] < \delta$$

if $t \geqslant t_1(\delta)$ is large enough. Now let R be so large that $\varepsilon_R R^2 \alpha_4(R) < t^{-2}$. Then, by (1.2), for $\mathscr{S} = \mathscr{S}_{f_{R\mathcal{Q}}}$,

$$\begin{aligned}\mathbb{E}[f_{R\mathcal{Q}}(y) f_{R\mathcal{Q}}(x)] &\geqslant \mathbb{E}\big[(1-\varepsilon_R)^{|\mathscr{S}|} \,\big|\, |\mathscr{S}| < tR^2\alpha_4(R)\big]\mathbb{P}[|\mathscr{S}| < tR^2\alpha_4(R)] \\ &\geqslant \left(1 - \frac{t^{-2}}{R^2\alpha_4(R)}\right)^{tR^2\alpha_4(R)} (1-\delta) \\ &\geqslant \exp\left(-\frac{O(1)}{t}\right)(1-\delta).\end{aligned}$$

This is arbitrarily close to $\mathbb{E}[f_{R\mathcal{Q}}(x)^2] = 1$ for t large; hence we are done. □

8.2. Resampling a fixed set of bits

We now prove a general version of Theorem 1.3. If $y \in \Omega = \{-1, 1\}^{\mathcal{I}}$ is a noisy version of x, then $y_j = x_j$, except on a small random set of $j \in \mathcal{I}$. We may consider a variation of this situation, where we have some fixed deterministic set $\mathcal{U} \subseteq \mathcal{I}$, and we take $y_j = x_j$ for $j \in \mathcal{U}$ and take the restriction of y to $\mathcal{U}^c := \mathcal{I} \setminus \mathcal{U}$ to be independent of x (and y is uniform in Ω). Although this setup was mentioned in [BKS], the techniques developed there and in [SSt] fell short of being able to handle this variation. Now, we can analyse this variation without difficulty, and prove the following proposition.

PROPOSITION 8.2. *Let $\mathcal{Q} \subset \mathbb{R}^2$ be some quad and for $R > 1$ let $f_R = f_{R\mathcal{Q}}$ be the ± 1-indicator function of the crossing event in $R\mathcal{Q}$ (either in $\mathbb{Z}^2$ or in the triangular lattice). For every $R > 1$, let $\mathcal{U}_R \subseteq \mathcal{I}$ be some set of bits, and let r_R be the maximal radius of any disk contained in $R\mathcal{Q}$ that is disjoint from $\mathcal{U}_R^c := \mathcal{I} \setminus \mathcal{U}_R$. If*

$$\lim_{R\to\infty} \frac{r_R}{\sqrt{R^2\alpha_4(R)}} = 0, \tag{8.8}$$

then the family $(\mathcal{U}_R)_{R>1}$ is asymptotically clueless *in the sense that*

$$\lim_{R\to\infty} \|\mathbb{E}[f_R \,|\, \mathcal{F}_{\mathcal{U}_R}] - \mathbb{E}[f_R]\| = 0.$$

On the other hand, if

$$\lim_{R\to\infty} |\mathcal{U}_R^c| \alpha_4(R) = 0, \tag{8.9}$$

then $(\mathcal{U}_R)_{R>1}$ is asymptotically decisive *in the sense that*

$$\lim_{R\to\infty} \|\mathbb{E}[f_R \,|\, \mathcal{F}_{\mathcal{U}_R}] - f_R\| = 0,$$

which means that there is asymptotically no loss of information about the crossing f_R.

Notice that even though the convergence of $\mathbb{E}[f_{R\mathcal{Q}}]$ is not known in $\mathbb{Z}^2$ for general quads or even rectangles other than squares, our definitions of being asymptotically clueless or decisive still make perfect sense.

Note that $O(1)|\mathcal{U}_R^c| \geqslant (R/r_R)^2$ and there are examples where $|\mathcal{U}_R^c| \asymp (R/r_R)^2$. Thus, in some sense the conditions (8.8) and (8.9) are nearly complementary. However, the following two examples are not covered. Suppose that $\mathcal{Q}$ is the unit square, and for each R we take $\mathcal{U}_R$ to be the set of bits contained in the left half of the square $R\mathcal{Q}$. It is left to the reader to verify that in this case $\mathcal{U}_R$ is neither asymptotically decisive, nor asymptotically clueless.

In the second example, we take $\mathcal{Q}$ to be the unit square again, and let $\mathcal{U}_R$ be the set of bits outside of the disk of radius ϱ_R centered at the center of the square $R\mathcal{Q}$. Then $\mathcal{U}_R$ is asymptotically decisive as long as $\varrho_R/R \to 0$, but this does not follow from the proposition (unless ϱ_R is small enough). However, Remark 8.5 below does give a general statement which covers this example.

Remark 8.3. When $(\mathcal{U}_R)_{R>1}$ is asymptotically clueless, it is immediate to see that if x_R and y_R are two coupled percolation configurations which coincide on $\mathcal{U}_R$, but are independent elsewhere, then

$$\lim_{R\to\infty} (\mathbb{E}[f_R(x_R) f_R(y_R)] - \mathbb{E}[f_R]^2) = 0.$$

On the other hand, if $(\mathcal{U}_R)_{R>1}$ is asymptotically decisive, then

$$\lim_{R\to\infty} (\mathbb{E}[f_R(x_R) f_R(y_R)] - \mathbb{E}[f_R^2]) = 0.$$

Proof of Proposition 8.2. By (2.9) and orthogonality of martingale differences, we have

$$\mathbb{P}[\varnothing \neq \mathscr{S}_{f_R} \subseteq \mathcal{U}_R] = \mathbb{E}[\mathbb{E}[f_R \,|\, \mathcal{F}_{\mathcal{U}_R}]^2 - \mathbb{E}[f_R]^2] = \mathbb{E}[\big(\mathbb{E}[f_R \,|\, \mathcal{F}_{\mathcal{U}_R}] - \mathbb{E}[f_R]\big)^2].$$

Thus, $(\mathcal{U}_R)_{R>1}$ is asymptotically clueless if and only if the spectral sample of f_R satisfies $\mathbb{P}[\varnothing \neq \mathscr{S}_{f_R} \subseteq \mathcal{U}_R] \to 0$. Similarly,

$$\mathbb{P}[\mathscr{S}_{f_R} \not\subseteq \mathcal{U}_R] = \mathbb{E}[f_R^2 - \mathbb{E}[f_R \,|\, \mathcal{F}_{\mathcal{U}_R}]^2] = \mathbb{E}[(\mathbb{E}[f_R \,|\, \mathcal{F}_{\mathcal{U}_R}] - f_R)^2].$$

Hence, a necessary and sufficient condition for $(\mathcal{U}_R)_{R>1}$ to be asymptotically decisive is that $\mathbb{P}[\mathscr{S}_{f_R} \subseteq \mathcal{U}_R] \to 1$.

We now consider the simpler case in which $\mathcal{Q} = [0,1]^2$, and assume (8.8). Since the proof is rather similar to the proof of Theorem 1.1, we will be brief here, and only

indicate some of the essential points and the arguments where a more substantial modification is necessary. As in §7.2, subdivide $[-2,R+2]^2$ into boxes $B_1,B_2,...,B_{m^2}$ of radius $(R+4)/2m$, where $m=m_R\in\mathbb{N}$ tends to infinity as $R\to\infty$, but very slowly. As above let B_j' denote the box concentric with B_j whose radius is a third of the radius of B_j. Let $H_R\subseteq\mathcal{U}_R^c$ be a maximal subset of $\mathcal{U}_R^c$ with the property that the distance between any two distinct elements in H_R is at least r_R. Then for some constant C every disk of radius Cr_R in $R\mathcal{Q}$ contains a point of H_R, but a disk of radius smaller than $\frac{1}{2}r_R$ contains at most one point of H_R. Let x_j be the indicator function of the event $\mathscr{S}_{f_R}\cap B_j\neq\varnothing$ and let y_j be the indicator function of the event

$$\mathscr{S}_{f_R}\cap B_j'\cap H_R\neq\varnothing.$$

Our goal is to prove that for each $I\subseteq\{1,...,m^2\}$ and every $j\in\{1,...,m^2\}\setminus I$,

$$\mathbb{P}[y_j=1\,|\,y_i=0\text{ for all }i\in I]\geqslant a\mathbb{P}[x_j=1\,|\,y_i=0\text{ for all }i\in I],\tag{8.10}$$

holds with some absolute constant $a>0$. We mimic the proof of Proposition 5.1. Fix such j and I, and set $n:=|B_j'\cap H_R|$, $W:=H_R\cap\bigcup_{i\in I}B_i'$ and $Y:=|\mathscr{S}_{f_R}\cap B_j'\cap H_R|$. Using Proposition 5.2, we get

$$O(1)\mathbb{E}[Y1_{\mathscr{S}_{f_R}\cap W=\varnothing}]\geqslant\mathbb{E}[\lambda_{B,W}^2]\alpha_4\left(\frac{R}{m}\right)n,$$

and Proposition 5.3 can be used to obtain

$$\mathbb{E}[Y^2 1_{\mathscr{S}_{f_R}\cap W=\varnothing}]\leqslant O(1)\mathbb{E}[\lambda_{B,W}^2]\alpha_4\left(\frac{R}{m}\right)^2 n^2,$$

provided that $\liminf_{R\to\infty}\alpha_4(R/m)n>0$ and hence the diagonal term is dominated by a constant times the off-diagonal term. (Intuitively, we need that the "density" of the set H_R inside the box B_j of radius R/m is good enough to see pivotals once the box has any of them.) Since $n\asymp R^2(mr_R)^{-2}$, this follows from (8.8), provided that m_R tends to ∞ sufficiently slowly. Then, (8.10) follows from the Cauchy–Schwarz second-moment bound.

Using Propositions 4.1 and 6.1, and following the proof of the tightness Theorem 1.1, we have for R large enough that

$$\mathbb{P}[\varnothing\neq\mathscr{S}_{f_R}\subseteq\mathcal{U}_R]\leqslant O(1)\gamma_{R/m}(R)\xrightarrow{(2.6)}0,\quad\text{as }R\to\infty.$$

By the above, it follows that $(\mathcal{U}_R)_{R>1}$ is asymptotically clueless.

For the case of a general quad, we can do the same trick as in §7.4 and in the proof of Corollary 8.1. For any $\delta>0$, there is a quad $\mathcal{Q}'$ contained in the interior of $\mathcal{Q}$ such that for R large enough

$$\mathbb{P}[\mathscr{S}_{f_R} \subseteq R\mathcal{Q}'] > 1-\delta. \tag{8.11}$$

We may assume that $\mathcal{Q}'$ is smooth, though this is not really needed, since we are going to use Proposition 5.11, with $\mathcal{Q}'$ in place of $\mathcal{Q}$. So, the above arguments (for the square) can easily be adapted to show that

$$\limsup_{R\to\infty} \mathbb{P}[\varnothing \neq \mathscr{S}_{f_R} \subseteq \mathcal{U}_R \cap R\mathcal{Q}'] = 0,$$

because the distance from $R\mathcal{Q}'$ to the complement of $R\mathcal{Q}$ is bounded from below by a positive constant times R. In combination with (8.11), this gives

$$\limsup_{R\to\infty} \mathbb{P}[\varnothing \neq \mathscr{S}_{f_R} \subseteq \mathcal{U}_R] \leqslant \delta.$$

Since the left-hand side does not depend on δ, we may let δ tend to 0 and deduce the first claim of the proposition.

For the other direction, we do a first-moment argument. Let $\mathcal{Q}'$ and δ satisfy (8.11), as before. Let $\varrho_0>0$ denote the distance from $\partial\mathcal{Q}'$ to $\partial\mathcal{Q}$. Then for $x\in\mathcal{I}\cap R\mathcal{Q}'$, we have $\mathbb{P}[x\in\mathscr{S}_{f_R}]\leqslant O(1)\alpha_4(\varrho_0 R)$. Thus,

$$\mathbb{E}[|\mathscr{S}_{f_R}\cap\mathcal{U}^c\cap R\mathcal{Q}'|] \leqslant |\mathcal{U}^c|\alpha_4(\varrho_0 R).$$

If we assume (8.9), then this tends to zero as $R\to\infty$. Thus,

$$\lim_{R\to\infty} \mathbb{P}[\mathscr{S}_{f_R}\cap\mathcal{U}^c\cap R\mathcal{Q}' \neq \varnothing] = 0,$$

and (8.11) gives

$$\limsup_{R\to\infty} \mathbb{P}[\mathscr{S}_{f_R}\cap\mathcal{U}^c \neq \varnothing] \leqslant \delta.$$

Once again, since $\delta>0$ is arbitrary, this completes the proof. □

Proof of Theorem 1.3. The theorem follows immediately from Proposition 8.2 and Remark 8.3. □

Remark 8.4. Recall that we used the random set $\mathcal{Z}$ in §5 and §7, with

$$\mathbb{P}[x\in\mathcal{Z}] = \frac{1}{\alpha_4(r)r^2},$$

just as a tool to measure the size of $\mathscr{S}$. However, in the spirit of our above proof, in place of $\mathcal{U}_r^c$, we can think of $\mathcal{Z}$ as the actual set of bits being resampled.

Remark 8.5. It may be concluded from a slight variation on the proof of Proposition 8.2 that in the setting of the triangular grid if the Hausdorff limit

$$F := \lim_{R\to\infty} \frac{\mathcal{U}_R^c}{R}$$

exists and has Hausdorff dimension strictly less than $\frac{5}{4}$, then $\mathcal{U}_R$ is asymptotically decisive. Indeed, assuming that $s{:=}\dim(F){<}\frac{5}{4}$, for every $\varepsilon{>}0$ one may find a countable collection of points z_j and radii ϱ_j, such that $\sum_j \varrho_j^{s+\varepsilon}{<}\varepsilon$ and the union of the disks with these centers and radii contains a neighborhood of F. The probability that $\mathcal{S}_{f_R}$ comes within distance $O(1)\varrho_j R$ of Rz_j is bounded from above by $O(1)\varrho_j^{5/4-\varepsilon}$, if z_j is not too close to the boundary of $\mathcal{Q}$ and R is sufficiently large. A sum bound and the above argument for dealing with a neighborhood of the boundary of $\mathcal{Q}$ complete the proof.

Remark 8.6. We obtain here a somewhat sharp result for "sensitivity to selective noise", though it would be even more satisfying to have a necessary and sufficient condition for a family $(\mathcal{U}_R)_{R>1}$ to be asymptotically clueless. We believe that $(\mathcal{U}_R)_{R>1}$ is asymptotically clueless if and only if $\mathbb{P}[\varnothing{\neq}\mathscr{P}_R{\subseteq}\mathcal{U}_R]{\to}0$, where $\mathscr{P}_R$ is the set of pivotals. Similarly, $(\mathcal{U}_R)_{R>1}$ should be asymptotically decisive if and only if $\mathbb{P}[\mathscr{P}_R{\subseteq}\mathcal{U}_R]{\to}1$. In other words, even though $\mathscr{P}_R$ and $\mathscr{S}_R$ are asymptotically quite different (compare, e.g., Remark 4.6 with Proposition 4.1), they should have the same polar sets.

Remark 8.7. Tsirelson [T] distinguishes two types of noise sensitivities: micro and block sensitivities, where the latter is stronger than the micro sensitivity we have been considering so far. He gives the following illustrative examples. Consider the two functions on $\{-1,1\}^n$:

$$f_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i \quad \text{and} \quad f_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n-n^{1/2}} \prod_{k=i}^{i+n^{1/2}} x_k.$$

Both correspond to renormalized random walks which converge to Brownian motion, but the first is stable while the second is noise-sensitive. Block sensitivity is defined as follows: instead of resampling the bits one by one, each with probability ε, we resample simultaneously blocks of bits. For $\delta{>}0$, divide the n bits into about δ^{-1} blocks of about δn bits: $B_i{:=}\mathbb{N}{\cap}[i\delta n,(i{+}1)\delta n)$. Each block is now resampled (i.e., all bits within the block) with probability ε. A sequence of functions is block sensitive if for any fixed $\varepsilon{>}0$, the lim sup as n goes to infinity of the correlation in this block procedure is bounded by a function of δ which goes to 0 when δ goes to 0. It is easy to see that the sequence of functions f_2 $(n{=}1,2,\dots)$ is not block-sensitive. This is related to the fact that the sensitivity of f_2 is "localized", in the sense that its spectral sample $\mathcal{S}_{f_2}$, when rescaled

by $1/n$, converges in law to a finite (random) set of points. As we will see in §10, this is not at all the case with the spectral sample of percolation. It is easy to check that percolation crossing events are indeed block sensitive.

9. Applications to dynamical percolation

In this section we prove Theorems 1.4 and 1.5.

As in §7.3, we consider the 0-1 indicator function $f=f_R$ of the percolation crossing event from $\partial([-1,1]^2)$ to $\partial([-R,R]^2)$. Then $\mathbb{E}[f]=\mathbb{E}[f^2]\asymp\alpha_1(R)$. We let ω_t be the dynamical percolation configuration at time t, started at the stationary distribution at $t=0$. Recall that we have

$$\mathbb{E}[f(\omega_0)f(\omega_t)]=\sum_{k=0}^{\infty}e^{-kt}\sum_{|S|=k}\hat{f}(S)^2,\quad t>0. \tag{9.1}$$

As in §8.1, for $s\geqslant 1$ define $\varrho(s)$ as the least $r\in\mathbb{N}_+$ such that $r^2\alpha_4(r)\geqslant s$, and break the sum over k in (9.1) into parts according to the $j\in\mathbb{N}$ satisfying $j/t\leqslant k<(j+1)/t$. Then, the same way we got (8.6), just now using Theorem 7.3 and the estimate

$$\alpha_1\left(\varrho\left(\frac{1}{t}\right)\right)\leqslant j^{O(1)}\alpha_1\left(\varrho\left(\frac{j}{t}\right)\right),$$

which follows from (8.4) and (2.3), we get

$$\mathbb{E}[f(\omega_0)f(\omega_t)]\leqslant O(1)\frac{\alpha_1(R)^2}{\alpha_1(\varrho(1/t))}=O(1)\frac{\mathbb{E}[f(\omega_0)]^2}{\alpha_1(\varrho(1/t))}. \tag{9.2}$$

Let $\mathscr{E}$ denote the set of exceptional times $t\in[0,\infty)$ for the event that the origin is in an infinite open cluster. To give a lower bound on the Hausdorff dimension of $\mathscr{E}$, a well-known technique is Frostman's criterion, see e.g. [MP, Theorem 4.27] or [M, Theorem 8.9]. Combined with a compactness argument, it gives the following; see [SSt, Theorem 6.1]. For any $\gamma>0$, let

$$M_\gamma(R):=\int_0^1\int_0^1\frac{\mathbb{E}[f_R(\omega_s)f_R(\omega_t)]}{\mathbb{E}[f_R(\omega_0)]^2|t-s|^\gamma}\,dt\,ds. \tag{9.3}$$

If $\sup_R M_\gamma(R)<\infty$, then $\mathscr{E}\cap[0,1]$ is non-empty with positive probability, and on this event a.s. its dimension is at least γ. It is easy to see that there is a constant d such that $\dim_H(\mathscr{E})=d$ a.s. Therefore, $\sup_R M_\gamma(R)<\infty$ also implies that $\dim_H(\mathscr{E})\geqslant\gamma$ a.s.

Proof of Theorem 1.4. To start with, we have $\varrho(s)=s^{4/3+o(1)}$, by (2.5). Secondly,

$$\alpha_1(r)=r^{-5/48+o(1)}$$

by [LSW2]. Thus, translation invariance and (9.2) give

$$\frac{\mathbb{E}[f(\omega_s)f(\omega_t)]}{\mathbb{E}[f(\omega)]^2}\leqslant O(1)|t-s|^{-(4/3)(5/48)+o(1)},$$

as $|t-s|\to 0$. Therefore, as long as $\gamma<1-\frac{4}{3}\frac{5}{48}$, we have $\sup_R M_\gamma(R)<\infty$. The above discussion therefore gives $\dim_H(\mathscr{E})\geqslant\frac{31}{36}$ a.s. The matching upper bound is given by Theorem 1.9 of [SSt]. This implies statement (1) of the theorem.

The proof of statement (2) is similar. For the 0-1 indicator function f^+ of the crossing event between radius 1 and R in a half-plane one gets the following analogue of Theorem 7.3: if $k\in\mathbb{N}_+$ satisfies $k\leqslant\alpha_4(r)r^2$, then

$$\mathbb{Q}[|\mathscr{S}_{f^+}|<k]\leqslant O(1)\frac{\alpha_1^+(R)^2}{\alpha_1^+(r)}.$$

The proof is similar to the proof of Theorem 7.3, and is left to the reader. The bound corresponding to (9.2) is then

$$\frac{\mathbb{E}[f^+(\omega_0)f^+(\omega_t)]}{\mathbb{E}[f^+(\omega_0)^2]}=O(1)\alpha_1^+\left(\varrho\left(\frac{1}{t}\right)\right)^{-1}. \tag{9.4}$$

By [SW, Theorem 3], we have $\alpha_1^+(r)=r^{-\xi_1^++o(1)}$, with $\xi_1^+=\frac{1}{3}$. The proof of the lower bound of $1-\frac{4}{3}\xi_1^+$ on the Hausdorff dimension then proceeds as above. For the upper bound, we refer to [SSt, Theorem 1.13]. This proves part (2).

For the proof of the third part, we now let f be the indicator function of the event that there is a white crossing in the upper half-plane (here, the set of hexagons whose center has non-negative imaginary part) and a black crossing in the lower half-plane (the set of hexagons whose center has negative imaginary part), both crossings from some fixed radius $r_0=2$ to radius R. The two half-planes are chosen here in such a way that the percolation configurations in the two halves are independent, and $r_0=2$ is chosen so that the event has positive probability for all $R>r_0$. Then, by independence, we get from (9.4) that

$$\frac{\mathbb{E}[f(\omega_0)f(\omega_t)]}{\mathbb{E}[f(\omega_0)^2]}=O(1)\alpha_1^+\left(\varrho\left(\frac{1}{t}\right)\right)^{-2}.$$

Thus, in this case we get the lower bound of $\frac{1}{9}$ for the corresponding Hausdorff dimension, which completes the proof. □

Proof of Theorem 1.5. Let $f=f_R$ be the indicator function for the existence of an open crossing from 0 to $\partial([-R,R]^2)$. We will apply the relation between α_1, α_4 and α_5 that comes from the $k=2$ case of Proposition A.1 in our appendix. Since $\alpha_5(r)\asymp r^{-2}$ (see [KSZ, Lemma 5] or [SSt, Corollary A.8]), the estimate (A.1) says that there are some constants $c_1,\varepsilon>0$ such that

$$\alpha_1(r)\alpha_4(r)>c_1r^{\varepsilon-2} \tag{9.5}$$

holds for all $r>1$. Thus,

$$\begin{aligned}\frac{\mathbb{E}[f(\omega_0)f(\omega_t)]}{\mathbb{E}[f(\omega_0)]^2}&\leqslant\frac{O(1)}{\alpha_1(\varrho(1/t))}&&\text{by (9.2)}\\&\leqslant O(1)\varrho\biggl(\frac{1}{t}\biggr)^{2-\varepsilon}\alpha_4\biggl(\varrho\biggl(\frac{1}{t}\biggr)\biggr)&&\text{by (9.5)}\\&\leqslant\frac{O(1)}{t}\varrho\biggl(\frac{1}{t}\biggr)^{-\varepsilon}&&\text{by (8.3)}\\&\leqslant O(1)t^{\varepsilon/2-1},\end{aligned}$$

where the last inequality follows from the definition of ϱ. Therefore, if we take $\gamma\in\bigl(0,\frac{1}{2}\varepsilon\bigr)$, then $\sup_R M_\gamma(R)<\infty$, and the set of exceptional times for having an infinite cluster almost surely has a positive Hausdorff dimension. □

Finally, note that if there were exceptional times with two distinct infinite white clusters with positive probability, then there would also be times with the 4-arm event from the origin to infinity. It was shown in [SSt] that this does not happen on the triangular lattice, and that there are no exceptional times on $\mathbb{Z}^2$ with three infinite white clusters. However, one can also easily prove the stronger result for $\mathbb{Z}^2$. Recall that (2.6) implies that $\alpha_4(r)^2r^2<O(1)r^{-\varepsilon}$ for some $\varepsilon>0$. This implies that the expected number of pivotals for the 4-arm event between radius 4 and R tends to zero as $R\to\infty$. (One should also take into account the sites near the outer boundary and near the inner boundary. Indeed, the total expected number of pivotals is $O(1)R^2\alpha_4(R)^2$.) Hence [SSt, Theorem 8.1] says that there are a.s. no exceptional times for the 4-arm event even on $\mathbb{Z}^2$.

10. Scaling limit of the spectral sample

Given $\eta>0$ let μ_η denote the law of Bernoulli$\bigl(\frac{1}{2}\bigr)$ site percolation on the triangular grid T_η of mesh η. Let ω denote a sample from μ_η. Given a quad $\mathcal{Q}\subset\mathbb{C}$, we can consider the event that $\mathcal{Q}$ is crossed by ω. To make this precise in the case where $\mathcal{Q}$ is not adapted to the grid, we may consider the white and black coloring of the hexagonal grid dual to T_η, as in §2.1. We let $f_\mathcal{Q}$ denote the ±1-indicator function of the crossing event. Let $\hat{\mu}_\eta^\mathcal{Q}$

denote the law of the spectral sample of $f_{\mathcal{Q}}$, that is, if $\mathcal{X}$ is a collection of subsets of the vertices of T_η, then

$$\hat{\mu}_\eta^{\mathcal{Q}}(\mathcal{X}) = \sum_{S \in \mathcal{X}} \hat{f}_{\mathcal{Q}}(S)^2.$$

Let d_0 denote the spherical metric on $\widehat{\mathbb{C}}=\mathbb{C}\cup\{\infty\}$ (with diameter π). If $S_1, S_2 \subseteq \widehat{\mathbb{C}}$ are closed and non-empty, let $d_H(S_1, S_2)$ be the Hausdorff distance between S_1 and S_2 with respect to the underlying metric d_0. If $S \neq \varnothing$, define $d_H(\varnothing, S)=d_H(S, \varnothing):=\pi$, and set $d_H(\varnothing, \varnothing)=0$. Then d_H is a metric on the set $\mathfrak{S}$ of closed subsets of $\widehat{\mathbb{C}}$. Since $(\mathfrak{S}\setminus\{\varnothing\}, d_H)$ is compact, the same holds for $(\mathfrak{S}, d_H)$. We may consider the probability measure $\hat{\mu}_\eta^{\mathcal{Q}}$ as a Borel measure on $(\mathfrak{S}, d_H)$.

THEOREM 10.1. *Let $\mathcal{Q}$ be a piecewise smooth quad in $\mathbb{C}$. Then the weak limit*

$$\hat{\mu}^{\mathcal{Q}} := \lim_{\eta \searrow 0} \hat{\mu}_\eta^{\mathcal{Q}}$$

(with respect to the metric d_H) exists. Moreover, it is conformally invariant, in the sense that if ϕ is conformal in a neighborhood of $\mathcal{Q}$ and $\mathcal{Q}' := \phi(\mathcal{Q})$, then $\hat{\mu}^{\mathcal{Q}} = \hat{\mu}^{\mathcal{Q}'} \circ \phi$.

As mentioned in the introduction, the existence of the limit follows from Tsirelson's theory (see [T, Theorem 3c5]) and [SS]. Nevertheless, we believe that our exposition below might be helpful.

Remark 10.2. The proof of the existence of the limit also works for subsequential scaling limits of critical bond percolation on $\mathbb{Z}^2$: if $\eta_j \searrow 0$ is a sequence along which bond percolation on $\eta_j \mathbb{Z}^2$ has a limit (in the sense of [SS], say), then the corresponding spectral sample measures also have a limit. (The existence of such sequences $\{\eta_j\}_{j=1}^{\infty}$ follows from compactness.)

In the proof of Theorem 10.1, we will use the following result.

PROPOSITION 10.3. ([SS]) *Let $\mathcal{Q}$ be a piecewise smooth quad in $\mathbb{C}$. Suppose that $\alpha \subset \mathbb{C}$ is a finite union of finite length paths, and that $\alpha \cap \partial\mathcal{Q}$ is finite. Then for every $\varepsilon > 0$ there is a finite collection of piecewise smooth quads $\mathcal{Q}_1, \mathcal{Q}_2, ..., \mathcal{Q}_n \subset \mathbb{C}\setminus\alpha$ and a function $g\colon \{-1,1\}^n \to \{-1,1\}$ such that*

$$\lim_{\eta \searrow 0} \mu_\eta[f_{\mathcal{Q}}(\omega) \neq g(f_{\mathcal{Q}_1}(\omega), f_{\mathcal{Q}_2}(\omega), ..., f_{\mathcal{Q}_n}(\omega))] < \varepsilon.$$

Another result from [SS] we will need is that for any finite sequence $\mathcal{Q}_1, \mathcal{Q}_2, ..., \mathcal{Q}_n$ of piecewise smooth quads in $\mathbb{C}$, the law of the vector $(f_{\mathcal{Q}_1}(\omega), f_{\mathcal{Q}_2}(\omega), ..., f_{\mathcal{Q}_n}(\omega))$ under μ_η has a limit as $\eta \searrow 0$, and that this limiting joint law is conformally invariant.

Proof of Theorem 10.1. Let $U \subset \mathbb{C}$ be an open set such that ∂U is a disjoint finite union of smooth simple closed paths and $\partial U \cap \partial \mathcal{Q}$ is finite. In order to establish the existence of the limit $\hat{\mu}^{\mathcal{Q}}$, it is clearly enough to show that for every such U the limit

$$W(\mathcal{Q}, U) := \lim_{\eta \searrow 0} \hat{\mu}_\eta^{\mathcal{Q}}(\mathscr{S} \subseteq U)$$

exists, and for the conformal invariance statement, it suffices to show that

$$W(\mathcal{Q}', \phi(U)) = W(\mathcal{Q}, U).$$

Fix some $\varepsilon > 0$ arbitrarily small. In Proposition 10.3, take $\alpha := \partial U$ and let $\mathcal{Q}_1, ..., \mathcal{Q}_n$ and g be as guaranteed there. Let $J := \{j \in \{1, 2, ..., n\} : \mathcal{Q}_j \subset U\}$ and $J' := \{1, ..., n\} \setminus J$. Then $\mathcal{Q}_j \cap U = \varnothing$ when $j \in J'$. Set $x := (f_{\mathcal{Q}_1}(\omega), ..., f_{\mathcal{Q}_n}(\omega)) \in \{-1, 1\}^n$, and let x_J and $x_{J'}$ denote the restrictions of x to J and J', respectively. Then for all η sufficiently small $x_{J'}$ is independent of ω_U and x_J is determined by ω_U.

Let $G = G(\omega) := g(x)$, let ν_η be the law of x under μ_η, and let $\nu := \lim_{\eta \searrow 0} \nu_\eta$. By (2.9), we have, for all η sufficiently small,

$$W_\eta(G, U) := \sum_{S \subset U} \widehat{G}(S)^2 = \mathbb{E}[\mathbb{E}[G(\omega) \,|\, \omega_U]^2].$$

We may write $g(x)$ as a sum

$$g(x) = \sum_{y \in \{-1,1\}^J} 1_{x_J = y} g_y(x_{J'})$$

with some functions $g_y : \{-1, 1\}^{J'} \to \{-1, 1\}$. Then

$$W_\eta(G, U) = \sum_y \nu_\eta(x_J = y) \nu_\eta[g_y]^2 \to \sum_y \nu(x_J = y) \nu[g_y]^2 \quad \text{as } \eta \searrow 0.$$

(Here, $\nu[g_y]$ denotes the expectation of g_y with respect to ν, and similarly for ν_η.) Hence, $W(G, U) := \lim_{\eta \searrow 0} W_\eta(G, U)$ exists.

By (2.7), we have

$$|\hat{\mu}_\eta^{\mathcal{Q}}(\mathscr{S} \subseteq U) - W_\eta(G, U)| \leqslant 4 \mu_\eta(G \neq f_{\mathcal{Q}})^{1/2}.$$

For η sufficiently small, the right-hand side is smaller than $4\sqrt{\varepsilon}$, by our choice of g. Since $W(G, U) = \lim_{\eta \searrow 0} W_\eta(G, U)$, we conclude that

$$|\hat{\mu}_\eta^{\mathcal{Q}}(\mathscr{S} \subseteq U) - W(G, U)| < 5\sqrt{\varepsilon}$$

for all η sufficiently small. (But we cannot say that $\lim_{\eta \searrow 0} \hat{\mu}_\eta^{\mathcal{Q}}(\mathscr{S} \subseteq U) = W(G, U)$, since G depends on ε.) This implies that

$$\limsup_{\eta \searrow 0} \hat{\mu}_\eta^{\mathcal{Q}}(\mathscr{S} \subseteq U) - \liminf_{\eta \searrow 0} \hat{\mu}_\eta^{\mathcal{Q}}(\mathscr{S} \subseteq U) \leqslant 10\sqrt{\varepsilon}.$$

Since ε is an arbitrary positive number, this establishes the existence of the limit $W(\mathcal{Q}, U)$. The proof of conformal invariance is similar, and left to the reader. □

We now describe some a.s. properties of the limiting law.

THEOREM 10.4. *If $\mathcal{S}$ is a sample from $\hat{\mu}^{\mathcal{Q}}$, then a.s. $\mathcal{S}$ is contained in the interior of $\mathcal{Q}$ and for every open $U\subset\mathbb{C}$, if $U\cap\mathcal{S}\neq\varnothing$, then $U\cap\mathcal{S}$ has Hausdorff dimension $\frac{3}{4}$. In particular, $\mathcal{S}$ is a.s. homeomorphic to a Cantor set, unless it is empty.*

Proof. It follows from (7.8) from §7.4 that $\mathcal{S}$ is $\hat{\mu}^{\mathcal{Q}}$-a.s. contained in the interior of $\mathcal{Q}$. Now fix some open U whose closure is contained in the interior of $\mathcal{Q}$. Fix $\eta>0$ and let $\mathcal{S}_\eta$ denote a sample from $\hat{\mu}^{\mathcal{Q}}_\eta$. Let λ_η denote the counting measure on $\mathcal{S}_\eta\cap U$ divided by $\eta^{-2}\alpha_4(1,1/\eta)$. We may consider λ_η as a random point in the metric space of Borel measures on $\mathcal{Q}$ with the Prokhorov metric. By (3.6) and the estimate (2.4), we have $\limsup_{\eta\searrow 0}\mathbb{E}[\lambda_\eta(U)]<\infty$. Therefore, the law of λ_η is tight as $\eta\searrow 0$. Likewise, the law of the pair $(\mathcal{S}_\eta,\lambda_\eta)$ is tight. Hence, there is a sequence $\eta_j\to 0$ such that the law of the pair $(\mathcal{S}_{\eta_j},\lambda_{\eta_j})$ converges weakly as $j\to\infty$. Let $(\mathcal{S},\lambda)$ denote a sample from the weak limit. Then λ is a.s. a measure whose support is contained in $\mathcal{S}$.

Now let $B\subset U$ be a closed disk. Let $B'\subset B$ be a concentric open disk with smaller radius, and let δ denote the distance from $\partial B'$ to ∂B. Theorem 7.1 with B' and B in place of U' and U implies that $\lambda(B)>0$ a.s. on the event $\varnothing\neq\mathcal{S}\cap B\subset B'$. (Note that $\varnothing\neq\mathcal{S}\cap B\subset B'$ is an open condition on $\mathcal{S}$, since it is equivalent to having $\varnothing\neq\mathcal{S}\cap B'$ and $\mathcal{S}\cap(B\setminus B')=\varnothing$.) But $B\setminus B'$ can be covered by $O(\delta^{-1})$ disks of radius δ. By (3.4) and (2.5), for each of these radius-δ disks, the probability that $\mathcal{S}$ intersects it is $O(\delta^{5/4+o(1)})$. Therefore, $\mathbb{P}[\mathcal{S}\cap B\not\subset B']=O(\delta^{1/4+o(1)})$. In particular, we have

$$\mathbb{P}[\mathcal{S}\cap B\neq\varnothing,\lambda(B)=0]=o(1)$$

as $\delta\searrow 0$; that is, $\mathbb{P}[\mathcal{S}\cap B\neq\varnothing,\lambda(B)=0]=0$. By considering a countable collection of disks covering U it follows that on $U\cap\mathcal{S}\neq\varnothing$ we have $\lambda(U)>0$ a.s. The correlation estimate (3.3) and the asymptotics $\alpha_4(r)=r^{-5/4+o(1)}$ from (2.5) imply that for $\eta>0$ and every $s>-\frac{3}{4}$ we have

$$\mathbb{E}\left[\int_U\int_U(|x-y|\vee\eta)^s\,d\lambda_\eta(x)\,d\lambda_\eta(y)\right]=O(1).$$

This implies that a.s.

$$\int_U\int_U|x-y|^s\,d\lambda(x)\,d\lambda(y)<\infty.$$

Therefore, Frostman's criterion implies that the Hausdorff dimension of $\mathcal{S}$ is a.s. at least $\frac{3}{4}$ on the event $\mathcal{S}\cap U\neq\varnothing$. By Lemma 3.2, the expected number of disks of radius r needed to cover $\mathcal{S}\cap U$ is bounded from above by $r^{3/4+o(1)}$. Hence, the Hausdorff dimension of $U\cap\mathcal{S}$ is a.s. at most $\frac{3}{4}$ on the event $U\cap\mathcal{S}\neq\varnothing$. This proves the claim for any fixed U. The assertion for every U then follows by considering a countable basis for the topology (i.e., disks having rational radius and centers with rational coordinates). □

Remark 10.5. It would be interesting to prove the weak convergence of the law of $(\mathscr{S}_\eta, \lambda_\eta)$ as $\eta \searrow 0$.

Note that the proof above shows that for any subsequential scaling limit $(\mathscr{S}, \lambda)$ of $(\mathscr{S}_\eta, \lambda_\eta)$, the support of the measure λ a.s. is the whole of $\mathscr{S}$.

Proof of Theorem 1.6. Since $[0,R]^2$ is a square, we have $\mathbb{E}[f_R] \to 0$ a.s. Therefore $\mathbb{P}[\mathscr{S}_{f_R} = \varnothing] \to 0$. Consequently, the claims follow from Theorems 10.1 and 10.4. □

11. Some open problems

Here is a list of some questions and open problems:

(1) For any Boolean function $f: \{-1,1\}^n \to \{-1,1\}$, define its *spectral entropy* to be

$$\mathrm{Ent}(f) = \sum_{S \subseteq \{1,\dots,n\}} \hat{f}(S)^2 \log \frac{1}{\hat{f}(S)^2}.$$

Friedgut and Kalai conjectured in [FK] that there is some absolute constant $C > 0$ such that for any Boolean function f,

$$\mathrm{Ent}(f) \leqslant C \sum_{S \subseteq \{1,\dots,n\}} \hat{f}(S)^2 |S| = C\mathbb{E}[|\mathscr{S}_f|];$$

in other words, that the spectral entropy is controlled by the total influence. As was pointed out to us by Gil Kalai, it is natural to test the conjecture in the setting of percolation: if f_R is the ± 1-indicator function of the left-right crossing in the square $[0,R]^2$, is it true that $\mathrm{Ent}(f_R) = O(R^2 \alpha_4(R))$?

(2) Our paper deals with noise sensitivity of percolation and its applications to dynamical percolation. One could ask similar questions about the Ising model, for which a natural dynamics is the Glauber dynamics. For instance, Broman and Steif ask in [BrS, Question 1.8] if there exist exceptional times for the Ising model on $\mathbb{Z}^2$ at $\beta = \beta_c$ for which there is an infinite up-spin cluster. Since SLE_3 (which is supposedly the scaling limit of critical Ising interfaces, see Smirnov's recent breakthrough [Sm2]) does not have double points, there should be very few pivotals, and thus such exceptional times should not exist, but the missing argument is a quasi-multiplicativity property for the probabilities of the alternating 4-arm events in the Ising model. Similar questions can be asked for the FK random cluster model, Potts models, etc.

(3) Prove the weak convergence of the law of $(\mathscr{S}_\eta, \lambda_\eta)$; see Remark 10.5. In a forthcoming paper [GPS1], we will prove the weak convergence of the law of $(\mathscr{P}_\eta, \tilde{\lambda}_\eta)$, where $\tilde{\lambda}_\eta$ is the counting measure on the set of pivotals renormalized by $\eta^{-2}\alpha_4(1, 1/\eta)$.

(4) Prove that $\mathscr{P}_R$ and $\mathscr{S}_R$ asymptotically have the same "polar sets". See Remark 8.6 for a more precise description.

(5) Prove that the laws of $\mathscr{P}_R$ and $\mathscr{S}_R$ are asymptotically mutually singular, or that their scaling limits are singular. Remark 4.6 suggests that this should be the case, since both these sets should be statistically self similar, in some sense.

(6) Do we have $\mathbb{E}[|\mathscr{S}_{R\mathcal{Q}}|]=\mathbb{E}[|\mathscr{P}_{R\mathcal{Q}}|]\asymp R^2\alpha_4(R)$ for any quad $\mathcal{Q}\subset\mathbb{C}$, as R goes to infinity? (See Theorem 7.4).

(7) In the same fashion, prove that the sharp tightness as in Theorem 1.1 still holds for general quads. With our techniques, this would require a uniform control over the domain on the constants involved in Proposition 5.11, as well as a statement analogous to Proposition 4.1 for the case of general quads.

(8) Prove that the main statement in §5 (Proposition 5.1) still holds for non-monotone functions such as the ℓ-arm annulus crossing events. (See §5.3 for an explanation why we needed the monotonicity assumption for the first moment.) If such a generalization was proved, then it would imply in particular that for the triangular lattice the set of exceptional times with both infinite black and white clusters has dimension $\frac{2}{3}$ a.s., strengthening the last statement in Theorem 1.4.

For $\ell>1$, there is a further small complication when ℓ is odd: a bit can be pivotal for the ℓ-arm event even without having the exact 4-arm event around its tile. (That is why we restricted Proposition 4.7 to the $\ell\in\{1\}\cup 2\mathbb{N}_+$ case.) Resolving this technicality and the non-monotonicity problem would imply the existence of exceptional times where there are polychromatic three arms from 0 to infinity (the dimension of this set of exceptional times would then be $\frac{1}{9}$). We cannot prove the existence of such times with the results of the present paper.

(9) Let us conclude with a computational problem: find any "efficient" algorithmic way to sample $\mathscr{S}$ in the case of percolation, say (in order, for instance, to make pictures of it), or prove that such an algorithm does not exist.

As we have recently learnt from Gil Kalai, the fact that the crossing function itself is computable in polynomial time (in the number of input bits) implies that there is a polynomial time *quantum algorithm* to sample $\mathscr{S}$ [BV, Theorem 8.4.2]. In fact, the Fourier sampling problem provided the first formal evidence that quantum Turing machines are more powerful than bounded-error probabilistic Turing machines.

Appendix A. An inequality for multi-arm probabilities

We prove here an estimate regarding the multi-arm crossing probabilities for annuli in critical bond percolation on $\mathbb{Z}^2$, which is due to Vincent Beffara (private communication)

and included here with his permission.

PROPOSITION A.1. *Fix $k\in\mathbb{N}_+$ and consider bond percolation on $\mathbb{Z}^2$ with parameter $p=\frac{1}{2}$. There are constants C, $\varepsilon>0$, which may depend on k, such that for all $1<r<R$,*

$$\alpha_{2k+1}(r,R)\leqslant C\alpha_1(r,R)\alpha_{2k}(r,R)\Big(\frac{r}{R}\Big)^{\varepsilon}. \tag{A.1}$$

The method of proof can be generalized to give a few similar results. However, new ideas seem to be necessary for the corresponding statement with $k=\frac{1}{2}$. The case $k=\frac{1}{2}$ is of particular significance: it was proved in [SSt] that it implies the existence of exceptional times for Bernoulli$\big(\frac{1}{2}\big)$ bond percolation on $\mathbb{Z}^2$. In the present paper we prove their existence using (A.1) instead.

Proof. For simplicity of notation, we will restrict the proof to the case $k=2$, which is the case we need, but the proof very easily carries over to the general case. Let $A(r,R)$ denote the annulus which is the closure of $B(0,R)\setminus B(0,r)$. If we have the 4-arm event in $A(r,R)$, i.e., four crossings of alternating colors between the two boundary components of $A(r,R)$, with white (primal) and black (dual) colors on the tiles given in §2.1, then there are at least 4 interfaces γ_1, γ_2, γ_3 and γ_4 separating these clusters. These interfaces are simple paths on the grid $\mathbb{Z}^2+\big(\frac{1}{4},\frac{1}{4}\big)$ in $A(r,R)$, and each of them has one point on each boundary component $A(r,R)$.

Let $\gamma=(\gamma_1,\gamma_2,\gamma_3)$ be a triple of three simple paths that can arise as three consecutive interfaces in cyclic order. Let $\mathcal{A}_\gamma$ denote the event that these are actual interfaces between crossing clusters. Let $S:=S_\gamma$ denote the connected component of $A(r,R)\setminus(\gamma_1\cup\gamma_3)$ that does not contain γ_2, and let $\mathcal{B}_\gamma$ denote the event that $\mathcal{A}_\gamma$ occurs and there are at least two disjoint primal crossings in S. Our first goal is to prove that

$$\mathbb{P}[\mathcal{B}_\gamma\,|\,\mathcal{A}_\gamma]\leqslant O(1)\alpha_1(r,R)\Big(\frac{r}{R}\Big)^{\varepsilon}, \tag{A.2}$$

with some constant $\varepsilon>0$.

Note that on $\mathcal{A}_\gamma$, we have in S at least one primal crossing and at least one dual crossing, which are adjacent to γ_1 and γ_3. For the sake of definiteness, we will assume that the primal crossing is adjacent to γ_1 and the dual crossing is adjacent to γ_3. (This can be determined from γ.) Let $S'=S'_\gamma$ denote the set of edges in S that are not adjacent to γ_1. Then given $\mathcal{A}_\gamma$, we have $\mathcal{B}_\gamma$ if and only if there is a primal crossing also in S'. Therefore, (A.2) follows once we show that for every such γ the probability that there is a crossing in S' is bounded from above by the right-hand side of (A.2).

We may consider a percolation configuration ω in the whole plane and also restrict it to S'. Let ω_ϱ denote the restriction of ω to $B(0,\varrho)$, for any $\varrho\in[r,R]$. Write $\varrho\leftrightarrow\varrho'$ for the

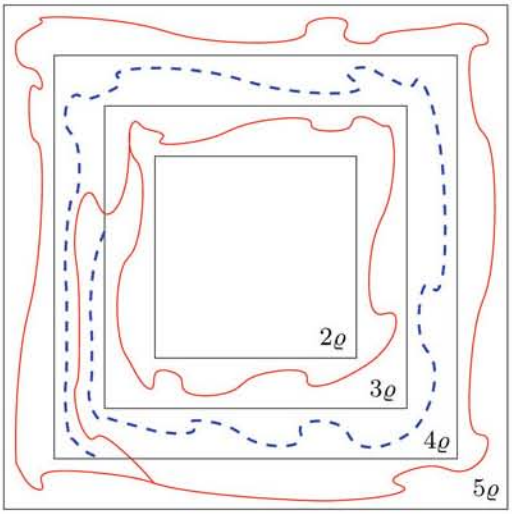


Figure A.1. The event $\mathcal{D}_j$ with $j=1$.

event that there is a crossing of ω between the two boundary components of the annulus $A(\varrho,\varrho')$, and write $\varrho\overset{D}{\longleftrightarrow}\varrho'$ for the existence of such a crossing within some specified set D. We will prove that for some constant $a\in(0,1)$ and every ϱ and ϱ' satisfying $r\vee 100\leqslant\varrho\leqslant\frac{1}{8}\varrho'\leqslant\frac{1}{8}R$ we have

$$\mathbb{P}[r\overset{S'}{\longleftrightarrow}\varrho'\,|\,\omega_\varrho, r\longleftrightarrow R]\leqslant a. \tag{A.3}$$

Using induction, this implies (A.2) with $\varepsilon=\log_8(1/a)$.

The interface γ_1 crosses the annulus $A(3\varrho,4\varrho)$ one or more times. Let w denote the winding number around 0 of one of these crossings, that is, the signed change of the argument along the crossing divided by 2π. Suppose that β is a simple path in $A(3\varrho,4\varrho)$ with one endpoint on each boundary component of the annulus and let w_β denote the winding number of β. If $\beta\cap\gamma_1=\varnothing$, then we may adjoin to $\beta\cup\gamma_1$ two arcs on the boundary components of the annulus to form a simple closed curve which has winding number in $\{0,\pm1\}$. Therefore, we see that $|w-w_\beta|>3$ implies that $\beta\cap\gamma_1\neq\varnothing$.

Let $\mathcal{D}_j$ denote the event that there is a dual crossing in ω of $A(3\varrho,4\varrho)$ with winding number in the range $\left[j-\frac{1}{2},j+\frac{1}{2}\right]$, and there are primal circuits in $A(2\varrho,3\varrho)$ and in $A(4\varrho,5\varrho)$, each of them separating the two boundary components of its annulus, and these primal circuits are connected to each other in ω; see Figure A.1. By the RSW theorem, there is a constant $\delta>0$ such that $\mathbb{P}[\mathcal{D}_{\pm10}]\geqslant\delta$. We claim that

$$\mathbb{P}[\mathcal{D}_{\pm10}\,|\,r\longleftrightarrow R,\omega_\varrho]\geqslant\delta. \tag{A.4}$$

If we condition on $\mathcal{D}_{10}$, and we further condition on the outermost primal circuit α_0 in $A(2\varrho,3\varrho)$ and on the innermost primal circuit α_1 in $A(4\varrho,5\varrho)$, then the configuration inside α_0 and the configuration outside α_1 remains unbiased, and if additionally α_0 is connected to the inner boundary component of $A(r,R)$ and α_1 is connected to the outer boundary component of $A(r,R)$, then we also have $r\leftrightarrow R$. This implies that $\mathbb{P}[r\leftrightarrow R|\omega_\varrho,\mathcal{D}_{10}]\geqslant\mathbb{P}[r\leftrightarrow R|\omega_\varrho]$. The same holds for $\mathcal{D}_{-10}$, and since $\mathcal{D}_{\pm10}$ is independent of ω_ϱ, the inequality (A.4) easily follows.

Now note that if $\mathcal{D}_j$ holds, then every primal crossing of $A(3\varrho,4\varrho)$ in ω has winding number in the range $[j-4,j+4]$. Hence, if $|j-w|>7$, then

$$\mathcal{D}_j\cap\{3\varrho\overset{S'}{\longleftrightarrow}4\varrho\}=\varnothing.$$

Consequently, at least one of the two events $\mathcal{D}_{\pm10}$ is disjoint from $\{\varrho\overset{S'}{\longleftrightarrow}4\varrho\}$. This gives (A.3) with $a:=1-\delta$. As we have argued before, (A.2) follows.

The 5-arm crossing event is certainly contained in $\bigcup_\gamma\mathcal{B}_\gamma$, where the union ranges over all γ as above. Hence, (A.2) gives

$$\alpha_5(r,R)\leqslant O(1)\Big(\frac{r}{R}\Big)^{\varepsilon}\alpha_1(r,R)\mathbb{E}\bigg[\sum_\gamma 1_{\mathcal{A}_\gamma}\bigg]. \tag{A.5}$$

If X is the number of interfaces crossing the annulus $A(r,R)$ (which is necessarily even), then $\sum_\gamma 1_{\mathcal{A}_\gamma}$ is not more than $X^3 1_{X\geqslant4}$. Since for all $j\in\mathbb{N}$ we have

$$\mathbb{P}[X\geqslant j]\leqslant O(1)\Big(\frac{r}{R}\Big)^{j\varepsilon_0}$$

with some constant $\varepsilon_0>0$ (by the RSW estimate and the BK inequality) and

$$\mathbb{P}[X\geqslant4]\geqslant\frac{1}{O(1)}\Big(\frac{r}{R}\Big)^{\varepsilon_1}$$

for some $\varepsilon_1\in(0,\infty)$, we have $\mathbb{E}[X^3 1_{X\geqslant4}]\leqslant O(1)\mathbb{P}[X\geqslant4]$ when $R>2r$, say. Therefore,

$$\mathbb{E}\bigg[\sum_\gamma 1_{\mathcal{A}_\gamma}\bigg]\leqslant\mathbb{E}[X^3 1_{X\geqslant4}]\leqslant O(1)\mathbb{P}[X\geqslant4]=O(1)\alpha_4(r,R).$$

When combined with (A.5), this proves the proposition in the case $k=2$. The general case is similarly obtained. □

References

[ADA] Aizenman, M., Duplantier, B. & Aharony, A., Path-crossing exponents and the external perimeter in 2D percolation. *Phys. Rev. Let.*, 83 (1999), 1359–1362.

[BHPS] Benjamini, I., Häggström, O., Peres, Y. & Steif, J. E., Which properties of a random sequence are dynamically sensitive? *Ann. Probab.*, 31 (2003), 1–34.

[BKS] Benjamini, I., Kalai, G. & Schramm, O., Noise sensitivity of Boolean functions and applications to percolation. *Inst. Hautes Études Sci. Publ. Math.*, 90 (1999), 5–43 (2001).

[BS] Benjamini, I. & Schramm, O., Exceptional planes of percolation. *Probab. Theory Related Fields*, 111 (1998), 551–564.

[BMW] van den Berg, J., Meester, R. & White, D. G., Dynamic Boolean models. *Stochastic Process. Appl.*, 69 (1997), 247–257.

[BV] Bernstein, E. & Vazirani, U., Quantum complexity theory. *SIAM J. Comput.*, 26 (1997), 1411–1473.

[BrS] Broman, E. I. & Steif, J. E., Dynamical stability of percolation for some interacting particle systems and ε-movability. *Ann. Probab.*, 34 (2006), 539–576.

[FK] Friedgut, E. & Kalai, G., Every monotone graph property has a sharp threshold. *Proc. Amer. Math. Soc.*, 124 (1996), 2993–3002.

[GPS1] Garban, C., Pete, G. & Schramm, O., Pivotal, cluster and interface measures for critical planar percolation. Preprint, 2010. `arXiv:1008.1378v1 [math.PR]`.

[GPS2] — The scaling limits of dynamical and near-critical percolation. In preparation.

[G] Grimmett, G., *Percolation.* Grundlehren der Mathematischen Wissenschaften, 321. Springer, Berlin–Heidelberg, 1999.

[HPS] Hammond, A., Pete, G. & Schramm, O., Local time for dynamical percolation, and the incipient infinite cluster. In preparation.

[H] Hoffman, C., Recurrence of simple random walk on $\mathbb{Z}^2$ is dynamically sensitive. *ALEA Lat. Am. J. Probab. Math. Stat.*, 1 (2006), 35–45.

[HP] Häggström, O. & Pemantle, R., On near-critical and dynamical percolation in the tree case. *Random Structures Algorithms*, 15 (1999), 311–318.

[HPSt] Häggström, O., Peres, Y. & Steif, J. E., Dynamical percolation. *Ann. Inst. Henri Poincaré Probab. Statist.*, 33 (1997), 497–528.

[JS] Jonasson, J. & Steif, J. E., Dynamical models for circle covering: Brownian motion and Poisson updating. *Ann. Probab.*, 36 (2008), 739–764.

[KKL] Kahn, J., Kalai, G. & Linial, N., The influence of variables on boolean functions, in *29th Annual Symposium on Foundations of Computer Science*, pp. 68–80. IEEE Computer Society, Los Alamitos, CA, 1988.

[KS] Kalai, G. & Safra, S., Threshold phenomena and influence: perspectives from Mathematics, Computer Science, and Economics, in *Computational Complexity and Statistical Physics*, St. Fe Inst. Stud. Sci. Complex., pp. 25–60. Oxford Univ. Press, New York, 2006.

[K1] Kesten, H., The incipient infinite cluster in two-dimensional percolation. *Probab. Theory Related Fields*, 73 (1986), 369–394.

[K2] — Scaling relations for 2D-percolation. *Comm. Math. Phys.*, 109 (1987), 109–156.

[KSZ] Kesten, H., Sidoravicius, V. & Zhang, Y., Almost all words are seen in critical site percolation on the triangular lattice. *Electron. J. Probab.*, 3 (1998), 75 pp.

[Kh] Khoshnevisan, D., Dynamical percolation on general trees. *Probab. Theory Related Fields*, 140 (2008), 169–193.

[KLM] KHOSHNEVISAN, D., LEVIN, D. A. & MÉNDEZ-HERNÁNDEZ, P. J., Exceptional times and invariance for dynamical random walks. *Probab. Theory Related Fields*, 134 (2006), 383–416.

[LSW1] LAWLER, G. F., SCHRAMM, O. & WERNER, W., Values of Brownian intersection exponents. II. Plane exponents. *Acta Math.*, 187 (2001), 275–308.

[LSW2] — One-arm exponent for critical 2D percolation. *Electron. J. Probab.*, 7 (2002), 13 pp.

[LSS] LIGGETT, T. M., SCHONMANN, R. H. & STACEY, A. M., Domination by product measures. *Ann. Probab.*, 25 (1997), 71–95.

[LMN] LINIAL, N., MANSOUR, Y. & NISAN, N., Constant depth circuits, Fourier transform, and learnability. *J. Assoc. Comput. Mach.*, 40 (1993), 607–620.

[M] MATTILA, P., *Geometry of Sets and Measures in Euclidean Spaces*. Cambridge Studies in Advanced Mathematics, 44. Cambridge University Press, Cambridge, 1995.

[MP] MÖRTERS, P. & PERES, Y., *Brownian Motion*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2010.

[N] NOLIN, P., Near-critical percolation in two dimensions. *Electron. J. Probab.*, 13 (2008), 1562–1623.

[PSS] PERES, Y., SCHRAMM, O. & STEIF, J. E., Dynamical sensitivity of the infinite cluster in critical percolation. *Ann. Inst. Henri Poincaré Probab. Stat.*, 45 (2009), 491–514.

[PS] PERES, Y. & STEIF, J. E., The number of infinite clusters in dynamical percolation. *Probab. Theory Related Fields*, 111 (1998), 141–165.

[R] REIMER, D., Proof of the van den Berg–Kesten conjecture. *Combin. Probab. Comput.*, 9 (2000), 27–32.

[Sch] SCHRAMM, O., Conformally invariant scaling limits: an overview and a collection of problems, in *International Congress of Mathematicians* (Madrid, 2006). Vol. I, pp. 513–543. Eur. Math. Soc., Zürich, 2007.

[SS] SCHRAMM, O. & SMIRNOV, S., On the scaling limits of planar percolation. To appear in *Ann. Probab.*

[SSt] SCHRAMM, O. & STEIF, J. E., Quantitative noise sensitivity and exceptional times for percolation. *Ann. of Math.*, 171 (2010), 619–672.

[Sm1] SMIRNOV, S., Critical percolation in the plane: conformal invariance, Cardy's formula, scaling limits. *C. R. Acad. Sci. Paris Sér. I Math.*, 333 (2001), 239–244.

[Sm2] — Towards conformal invariance of 2D lattice models, in *International Congress of Mathematicians* (Madrid, 2006). Vol. II, pp. 1421–1451. Eur. Math. Soc., Zürich, 2006.

[SW] SMIRNOV, S. & WERNER, W., Critical exponents for two-dimensional percolation. *Math. Res. Lett.*, 8 (2001), 729–744.

[T] TSIRELSON, B., Scaling limit, noise, stability, in *Lectures on Probability Theory and Statistics*, Lecture Notes in Math., 1840, pp. 1–106. Springer, Berlin–Heidelberg, 2004.

[W] WERNER, W., Lectures on two-dimensional critical percolation, in *Statistical Mechanics*, IAS/Park City Math. Ser., 16, pp. 297–360. Amer. Math. Soc., Providence, RI, 2009.

Christophe Garban
CNRS
Département de mathématiques (UMPA)
École normale supérieure de Lyon
FR-69364 Lyon Cedex 07
France
christophe.garban@umpa.ens-lyon.fr

Gábor Pete
Department of Mathematics
University of Toronto
40 St George St
Toronto, ON M5S 2E4
Canada
gabor@math.toronto.edu

Oded Schramm
was at the Theory Group
of Microsoft Research
One Microsoft Way
Redmond, WA 98052-6399
U.S.A.
http://research.microsoft.com/
en-us/um/people/schramm/

Received May 6, 2008
Received in revised form November 12, 2009

Part III
Random Walks and Graph Limits

Electronic Journal of Probability

Vol. 6 (2001) Paper no. 23, pages 1–13.

Journal URL
http://www.math.washington.edu/~ejpecp/
Paper URL
http://www.math.washington.edu/~ejpecp/EjpVol6/paper23.abs.html

RECURRENCE OF DISTRIBUTIONAL LIMITS OF FINITE PLANAR GRAPHS

Itai Benjamini
The Weizmann Institute and Microsoft Research, Mathematics Department
Weizmann Institute of Science, Rehovot 76100, Israel
itai@wisdom.weizmann.ac.il

Oded Schramm
Microsoft Research, One Microsoft way, Redmond, WA 98052
schramm@microsoft.com

Abstract Suppose that G_j is a sequence of finite connected planar graphs, and in each G_j a special vertex, called the root, is chosen randomly-uniformly. We introduce the notion of a distributional limit G of such graphs. Assume that the vertex degrees of the vertices in G_j are bounded, and the bound does not depend on j. Then after passing to a subsequence, the limit exists, and is a random rooted graph G. We prove that with probability one G is recurrent. The proof involves the Circle Packing Theorem. The motivation for this work comes from the theory of random spherical triangulations.

Keywords Random triangulations, random walks, mass trasport, circle packing, volume growth

Mathematics subject classification 2000 82B41, 60J45 ,05C10

Submitted to EJP on March 14, 2001. Final version accepted on September 19, 2001.

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_15, © Springer Science+Business Media, LLC 2011

1 Introduction

1.1 Random triangulations

In recent years, physicists were interested in the study of random surfaces [ADJ97]. Random triangulations turned out to be a useful model for exact calculations, non rigorous arguments, and Monte-Carlo simulations regarding the geometry of random surfaces and the behaviour of physical systems on these surfaces. From a mathematical viewpoint, natural measures that were considered are the uniform measures on isomorphism classes of triangulations of the sphere with a fixed number of vertices. In [AAJ+98, ANR+98] diffusion on some random surfaces and random walks on random triangulations including the uniform measure have been considered. It was suggested there that the probability for the random walk to be at time t at its starting vertex should decay like t^{-1}, provided that t is not too large relative to the size of the triangulation, and that the mean square displacement at time t is $t^{1/2}$. Motivated by these observations, we decided to study the recurrence versus transience dichotomy for limits of random rooted spherical triangulations. It turned out that in the end the results apply to the more general setting of planar bounded-degree graphs. It will be proven that under the assumption of a uniform bound on the vertex degrees, limits of finite planar graphs are recurrent (provided that the root is chosen uniformly).

1.2 Limits of graphs

In addition to studying asymptotic properties of large random objects, it is mathematically appealing and natural to introduce a limiting infinite object and study its properties. In order to define the limit of a sequence of (possibly random) triangulations or graphs, it is necessary to keep track of a basepoint; or a **root**.

A **rooted graph** is just a pair (G, o), where G is a (connected) graph and o is a vertex in G. A rooted graph (G, o) is isomorphic to (G', o') if there is an isomorphism of G onto G' which takes o to o'. In this paper, we only consider **locally finite** graphs; that is, each vertex has finitely many neighbors.

The space $\mathcal{X}$ of isomorphism classes of rooted connected (locally finite) graphs has a natural topology, which is induced by the following metric. Let $(G, o), (G', o') \in \mathcal{X}$. For $r = 1, 2, \ldots$, let $B_G(o, r)$ be the closed ball of radius r about o in G, and similarly for G'. Let k be the supremum of all r such that $\big(B_G(o, r), o\big)$ and $\big(B_{G'}(o', r), o'\big)$ are isomorphic as rooted graphs, and set $d\big((G, o), (G', o')\big) := 2^{-k}$, where $2^{-\infty} := 0$. Then d is a metric on $\mathcal{X}$. Moreover, it is easy to verify that for every $M < \infty$ the subspace $\mathcal{X}_M \subset \mathcal{X}$ of graphs with degrees bounded by M is compact in this topology.

Suppose that H is a finite connected graph. We cannot think of H as an element of $\mathcal{X}$, unless a root o is chosen. The most natural way to choose a root is to make the choice random, and uniform among the vertices of H. In this way, a finite unrooted graph H corresponds to a probability measure μ_H on $\mathcal{X}$. More explicitly, for every Borel subset

$\mathcal{A} \subset \mathcal{X}$, $\mu_H(\mathcal{A})$ is equal to the probability that $(H, o) \in \mathcal{A}$ when o is chosen uniformly and randomly among the vertices of H.

Suppose that (G, o) is a random rooted finite graph. (This means that (G, o) is a sample from a Borel probability measure μ on $\mathcal{X}$, which is called the **law** of (G, o), and μ is supported on the set of finite graphs.) Then (G, o) is **unbiased** if its law is in the closed convex hull of the measures μ_H. In other words, for every finite graph H, conditioned on the event that G is isomorphic to H, the distribution of (G, o) is μ_H (provided that G is isomorphic to H with positive probability). Informally, a random rooted finite graph (G, o) is unbiased, if given G the root o is uniformly distributed among the vertices $V(G)$.

Let (G, o) and $(G_1, o_1), (G_2, o_2), \ldots$ be random connected rooted graphs. We say that (G, o) is the **distributional limit** of (G_j, o_j) as $j \to \infty$ if for every $r > 0$ and for every finite rooted graph (H, o'), the probability that (H, o') is isomorphic to $\big(B_{G_j}(o_j, r), o_j\big)$ converges to the probability that (H, o') is isomorphic to $\big(B_G(o, r), o\big)$. This is equivalent to saying that the law of (G, o), which is a probability measure on $\mathcal{X}$, is the weak limit of the law of (G_j, o_j) as $j \to \infty$.

It is easy to see, by compactness or diagonalization, that if $(G_j, o_j) \in \mathcal{X}_M$ a.s., $M < \infty$, then there is always a subsequence of the sequence (G_j, o_j) having a distributional limit.

Theorem 1.1. *Let $M < \infty$, and let (G, o) be a distributional limit of rooted random unbiased finite planar graphs G_j with degrees bounded by M. Then with probability one G is recurrent.*

Remarks. To illustrate the theorem, the reader may wish to consider the case where each G_j is a finite binary tree of depth j. (In that case the distance of the root from the leaves of the tree is approximately a geometric random variable. Hence, G is not the 3-regular tree.)

The assumption of planarity in the theorem is necessary, as can be seen by considering the intersection of $\mathbb{Z}^3$ with larger and larger balls. In this case, since the surface area to volume ratio tends to zero, the root is not likely to be close to the boundary. The distributional limit will then be $(\mathbb{Z}^3, 0)$ a.s. Also, the 3-regular tree can be obtained as the a.s. limit of unbiased finite graphs (3 regular graphs with girth going to infinity).

A natural extension of the collection of all planar graphs is the collection of all graphs with an excluded minor. It is reasonable to guess that if H is any graph and one replaces the assumption of planarity by the assumption that all G_j do not have H as a minor, then the theorem still holds, because many theorems on planar graphs generalize to excluded minor graphs (see, e.g., [Tho99]).

Because of the assumption of a uniform bound on the degrees, the interesting case when G_n is uniformly distributed among all isomorphism classes of triangulations of size n is unfortunately not included. We **conjecture** that the theorem still holds for limits of these measures. (See [Mal99, GW00] for a study of the degree distribution for those measures.) If G_n is uniformly distributed among spherical triangulations of size n with degree at most M, then the theorem above does apply to any (subsequential) limit of G_n.

A key to almost all the rigorous results regarding random surfaces and triangulations is the enumeration techniques, which originated with the fundamental work of Tutte [Tut62]. Schaeffer [Sch99] found a simpler proof for Tutte's enumeration, and thereby produced a good sampling algorithm. In contrast, our approach in this paper makes no use of enumeration. The proof of 1.1 is based on the theory of circle packings with specified combinatorics.

Following is another statement of the theorem from a slightly different perspective. Let G be a finite graph and let $X_0, X_1, \ldots$ be simple random walk on G started from a random-uniform vertex $X_0 \in V(G)$. Let $\phi(n, G)$ be the probability that $X_j \neq X_0$ for all $j = 1, 2, \ldots, n$, and set

$$\phi_M(n) := \sup\bigl\{\phi(n, G) : G \in \mathcal{G}_M\bigr\},$$

where $\mathcal{G}_M$ denotes the collection of finite planar graphs with maximum degree at most M.

Corollary 1.2. *For all $M < \infty$,*

$$\lim_{n\to\infty} \phi_M(n) = 0.$$

Proof. Let $h := \inf_{n\in\mathbb{N}} \phi_M(n)$. Then there is a sequence of graphs $G_j \in \mathcal{G}_M$ with $\phi(n, G_j) \geq \phi(j, G_j) \geq h/2$ for all $n \leq j$. Let (G, o) be a distributional limit of a subsequence of (G_j, o_j), where o_j is unbiased in $V(G_j)$. Then for every $n \in \mathbb{N}$ the probability that simple random walk on G started from o will not revisit o in the first n steps is at least $h/2$. By the theorem, G is recurrent, and hence $h = 0$, as required. □

Problem 1.3. *Determine the rate of decay of $\phi_M(n)$ as a function of n.*

The example of an $n \times n$ square in the grid $\mathbb{Z}^2$ shows that $\inf_n \phi_M(n) \log n > 0$. It might be reasonable to guess that $\phi_M(n)$ decays like $c(M)/\log n$.

In the next section, a proof of Theorem 1.1 is given, and the last section is devoted to miscellaneous remarks, including examples of planar triangulations with uniform growth r^α, $\alpha \notin \{1, 2\}$.

2 Recurrence

Theorem 1.1 will be proved using the theory of combinatorially specified circle packings. At the foundation of this theory is the Circle-Packing Theorem, which states that for every finite planar graph G there is a disk-packing P in the plane whose tangency graph is G; which means that the disks in P are indexed by the vertices $V(G)$ of G and two disks are tangent iff the corresponding vertices share an edge. When G is the 1-skeleton of a triangulation of the sphere, the packing P is unique up to Möbius transformations.

The Circle Packing Theorem was first proved by Koebe [Koe36] and later various generalizations have been obtained. See, e.g., [dV91] for a proof of this result. The Circle Packing Theorem has been instrumental in answering some problems about the geometry and potential theory of planar graphs [MP94, MT90, BS96, JS00]. (The relations between circle-packings and analytic function theory are important, fascinating, and much studied, but are not as relevant to the present paper.) Some surveys of the theory of combinatorially specified circle packings are also available [Ber92, Sac94, Ste99].

The major step in the proof of Theorem 1.1 is the case of triangulations, that is,

Proposition 2.1 (The triangulation case). *Let $M < \infty$, and let (T, o) be a distributional limit of rooted random unbiased (finite) triangulations of the sphere T_j with degrees bounded by M. Then with probability one T is recurrent.*

Proof. We assume, as we may, that T is a.s. infinite. By the Circle-Packing Theorem, for each j, there is a disk-packing P^j in the plane with tangency graph T_j. Since T^j is random, also P^j is random. (Actually, the randomness in T^j plays no role in the proof, and we could assume that T^j is deterministic. The essential randomness is that of o_j.) We choose P^j to be independent of o_j given T^j. Our immediate goal is to take an appropriate limit of the disk-packings P^j to obtain a disk-packing P with tangency graph T.

There is a unique triangle t_j in T_j whose vertices correspond to three disks of P^j which intersect the boundary of the unbounded component of $\mathbb{R}^2 \setminus P^j$. For every vertex $v \in V(T_j) \setminus t_j$, the disks corresponding to neighbors of v in T_j surround the disk P^j_v like petals of a flower, as in Figure 2.1. By shrinking the packing, if necessary, assume that P^j is contained in the unit disk $B(0, 1)$.

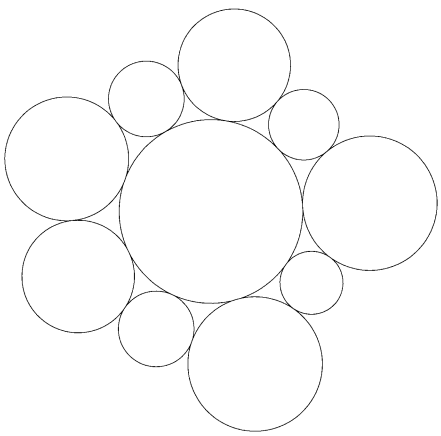


Figure 2.1: The Ring Lemma setup.

Let $\hat{P}^j$ be the image of the packing P^j under the map $z \mapsto az + b$, where $a \in (0, \infty)$ and $b \in \mathbb{R}^2$ are chosen so that $\hat{P}^j_{o_j}$ is the unit disk $B(0, 1)$.

Here is a simple but important fact about disk packings, known as the Ring Lemma [RS87]. If a disk P_0 is surrounded by n other disks $P_1, \ldots, P_n$, as in Figure 2.1, then the ratio r_0/r_1 between the radius of P_0 and the radius of P_1 is bounded from above by a constant which depends only on n. Since the vertex degrees in the triangulations T_j are all bounded by M, it follows that for every d there is some upper bound $c = c(d, M)$ for the ratio r/r' between any two radii of disks corresponding to vertices at combinatorial distance at most d from o_j, provided that o_j is at combinatorial distance at least $d+1$ from t_j. Because $|V(T_j)| \to \infty$ as $j \to \infty$, with probability tending to 1, $P^j_{o_j}$ is not one of the boundary disks corresponding to vertices of t_j. Moreover, by the uniform bound on the degrees, the combinatorial distance in T^j from o_j to t_j tends to infinity in probability as $j \to \infty$, because for every $r \geq 1$ the number of vertices of T^j at distance at most r from t_j is bounded, and hence o_j is unlikely to be there. We may therefore conclude that for every d there is a constant $c = c(d)$ so that with probability tending to 1 all the disks in $\hat{P}^j$ corresponding to vertices at combinatorial distance at most d from o_j have radii in $[1/c, c]$. By compactness, we may take a subsequence of j tending to ∞ so that the law of $\hat{P}^j$ tends (in the appropriate topology) to a random disk packing P whose tangency graph is T. Assume, with no loss of generality, that there is no need to pass to a subsequence.

An **accumulation point** of P is a point $z \in \mathbb{R}^2$ such that every open set containing z intersects infinitely many disks in P. Below, we prove the following result.

Proposition 2.2. *With probability* 1*, there is at most one accumulation point in* $\mathbb{R}^2$ *of the packing* P*.*

Proof of Proposition 2.1, continued. Assuming 2.2, the proof is completed as follows. In [HS95, Thms. 2.6, 3.1.(1), 8.1] and independently in [McC98] it was shown that if G is a bounded degree tangency graph of a disk-packing in the plane which has no accumulation points in the plane, then G is recurrent. This completes the proof if P has no accumulation points in $\mathbb{R}^2$. If P has one accumulation point $p \in \mathbb{R}^2$, then consider the subgraph G_1 of T spanned by the vertices corresponding to disks contained in the disk $B(p, 1)$ of radius 1 about p. By inverting in the circle of radius 1 about p, it follows that G_1 is recurrent. (Note that the above quoted results do not require the graph to be the 1-skeleton of a triangulation.) Similarly, the subgraph G_2 spanned by the vertices of T corresponding to disks that intersect the complement of $B(p, 1)$ is also recurrent. Since $V(T) = V(G_1) \cup V(G_2)$ and the boundary separating G_1 and G_2 is finite, it follows that T is recurrent. □

For the proof of Proposition 2.2, a lemma will be needed, but some notations must be introduced first. Suppose that $C \subset \mathbb{R}^2$ is a finite set of points. (In the application below, C will be the set of centers of disks in P^j.) When $w \in C$, we define its **isolation radius** as $\rho_w := \inf\{|v - w| : v \in C \setminus \{w\}\}$. Given $\delta \in (0, 1)$, $s > 0$ and $w \in C$, we say that w is (δ, s)-**supported** if in the disk of radius $\delta^{-1}\rho_w$, there are more than s points of C outside of every disk of radius $\delta\rho_w$; that is, if

$$\inf_{p \in \mathbb{R}^2} \left| C \cap B(w, \delta^{-1}\rho_w) \setminus B(p, \delta\rho_w) \right| \geq s\,.$$

Lemma 2.3. *For every $\delta \in (0,1)$ there is a constant $c = c(\delta)$ such that for every finite $C \subset \mathbb{R}^2$ and every $s \geq 2$ the set of (δ, s)-supported points in C has cardinality at most $c|C|/s$.*

Proof. Let $k \in \{3, 4, 5, \dots\}$. Consider a bi-infinite sequence $\mathfrak{S} = (\mathfrak{S}_n : n \in \mathbb{Z})$ of square-tilings $\mathfrak{S}_n$ of the plane, where for all $n \in \mathbb{Z}$ all the squares in the tiling $\mathfrak{S}_{n+1}$ have the same size and each square in the tiling $\mathfrak{S}_{n+1}$ is tiled by k^2 squares in the tiling $\mathfrak{S}_n$. Let $\hat{\mathfrak{S}}$ denote the collection of all the squares in all the tilings $\mathfrak{S}_n$. Assume that no point of C lies on the boundary of a square in $\hat{\mathfrak{S}}$. For every $n \in \mathbb{Z}$, say that a square $S \in \mathfrak{S}_n$ is s-supported if for every square $S' \in \mathfrak{S}_{n-1}$, we have $\big|C \cap S \setminus S'\big| \geq s$. To estimate the number of s-supported squares in $\hat{\mathfrak{S}}$, we now define a "flow" f on $\hat{\mathfrak{S}}$. Set

$$f(S', S) := \begin{cases} \min\big\{s/2, |S' \cap C|\big\} & S \in \mathfrak{S}_{n+1},\ S' \in \mathfrak{S}_n,\ S' \subset S, \\ 0 & S \in \mathfrak{S}_{n+1},\ S' \in \mathfrak{S}_n,\ S' \not\subset S, \\ -f(S, S') & S \in \mathfrak{S}_n,\ S' \in \mathfrak{S}_{n+1}, \\ 0 & S \in \mathfrak{S}_n,\ S' \in \mathfrak{S}_m,\ |m-n| \neq 1. \end{cases}$$

Let $a \in \mathbb{Z}$ be small enough so that each square of $\mathfrak{S}_a$ contains at most one point of C, and let $b \in \mathbb{Z}$, $b > a$. Then

$$\sum_{S' \in \mathfrak{S}_a} \sum_{S \in \mathfrak{S}_{a+1}} f(S', S) = |C|\,, \qquad \sum_{S' \in \mathfrak{S}_b} \sum_{S \in \mathfrak{S}_{b+1}} f(S', S) \geq 0\,.$$

Therefore

$$\sum_{n=a+1}^{b} \sum_{S \in \mathfrak{S}_n} \sum_{S' \in \hat{\mathfrak{S}}} f(S', S) \leq |C|\,, \tag{2.1}$$

because for every pair $S' \in \mathfrak{S}_n$, $S \in \mathfrak{S}_{n+1}$ with $n \in \{a+1, \dots, b-1\}$, the corresponding two terms $f(S, S')$ and $f(S', S)$ both appear on the left hand side and they cancel each other. By the definition of f, for every $S \in \hat{\mathfrak{S}}$ we have $\sum_{S' \in \hat{\mathfrak{S}}} f(S', S) \geq 0$. On the other hand, if $S \in \hat{\mathfrak{S}}$ is s-supported, then $\sum_{S' \in \hat{\mathfrak{S}}} f(S', S) \geq s/2$. Therefore, (2.1) implies that the number of s-supported squares in $\hat{\mathfrak{S}}$ is at most $2|C|/s$.

To estimate the number of (δ, s)-supported points of C, we will compare it with the expected number of s-supported squares in $\hat{\mathfrak{S}}$, where $\mathfrak{S}$ is chosen randomly, as follows. Take $k := \lceil 20\delta^{-2} \rceil$ as the parameter for the sequence $\mathfrak{S}$. Let $\mathfrak{S}$ have the distribution such that the distribution of $\hat{\mathfrak{S}}$ is invariant under translation and rescaling, and such that the diameter of any square in $\mathfrak{S}_0$ is in the range $[1, k)$.

To be explicit, we now construct this distribution. Let $(\alpha_n, n \in \mathbb{Z})$ be a sequence of independent random variables with each α_n uniform in $\{0, 1, \dots, k-1\}^2$. Let β be uniform in $[0, \log k)$ and independent from the sequence (α_n). Then we may take

$$\mathfrak{S}_n = \Big\{ e^\beta k^n \big([j, j+1] \times [j', j'+1]\big) + e^\beta \sum_{m=-\infty}^{n-1} k^m \alpha_m : j, j' \in \mathbb{Z} \Big\}.$$

Let N be the number of (δ, s)-supported points in C. Say that a point $w \in C$ is a **city** in a square $S \in \hat{\mathfrak{S}}$ if the edge-length of S is in the range $[4\delta^{-1}\rho_w, 5\delta^{-1}\rho_w]$ and the distance from w to the center of S is at most $\delta^{-1}\rho_w$. It is easy to see that there is a constant $c_0 = c_0(k) > 0$, which does not depend on C or w, such that w is a city for some square of $\hat{\mathfrak{S}}$ with probability at least c_0. By the above choice of k, if w is a city in S and w is (δ, s)-supported, then S is s-supported. Consequently, the expected number of pairs (w, S) such that w is a city in S and S is s-supported is at least $c_0 N$. However, by area considerations it is clear that there is a constant $c_1 = c_1(\delta)$ such that any square S cannot have more than c_1 cities in it. Hence, the expected number of s-supported squares is at least $c_0 N/c_1$. However, we have seen above that the number of s-supported squares is at most $2|C|/s$. Hence $N \le 2{c_0}^{-1}c_1|C|/s$, which proves the lemma. □

Proof of Proposition 2.2. Suppose that there is a positive probability that P has two distinct accumulation points in $\mathbb{R}^2$. Then there is a $\delta \in (0,1)$ and an $\epsilon > 0$ such that with probability at least ϵ there are two accumulation points p_1, p_2 in $B(0, \delta^{-1})$ such that $|p_1 - p_2| \ge 3\delta$. But this implies that for arbitrarily large s and for infinitely many j there is probability at least ϵ that the center of $P^j_{o_j}$ is (δ, s)-supported in the set C^j of centers of the disks in P^j, contradicting Lemma 2.3. □

Proof of Theorem 1.1. We claim that there is a constant c such that for all $j = 1, 2, \ldots$ there is a triangulation T_j of the sphere with maximum vertex degree at most cM, which contains a subgraph isomorphic to G_j, and such that $|V(T_j)| \le c|V(G_j)|$. Indeed, embed G_j in the plane, and let f be a face of this embedding. Let $v_0, \ldots, v_{k-1}$ be the vertices on the boundary of f, in clockwise order. First, suppose that $v_j \ne v_i$ for $j \ne i$, and that v_j does not neighbor with v_i unless $|i - j| = 1$ or $\{i, j\} = \{0, k-1\}$. In that case, we triangulate f by a zigzag pattern; that is, put into T_j the edges $[v_j, v_{k-j}]$, $j = 1, 2, \ldots, \lfloor k/2 \rfloor - 1$ and the edges $[v_j, v_{k-1-j}]$, $j = 1, 2, \ldots, \lfloor (k-1)/2 \rfloor - 1$. Otherwise, first put inside f a cycle of length k, $u_0, \ldots, u_{k-1}$, then put the edges $[u_j, v_j]$, $j = 0, \ldots, k-1$ and $[u_j, v_{j+1}]$, $j = 0, \ldots, k-2$ and $[u_{k-1}, v_0]$, and then triangulate the face bounded by the new cycle. It is easy to verify that when this construction is applied to every face f of the embedding of G, the resulting triangulation satisfies the required conditions for some appropriate constant c.

Let o_j be a vertex chosen uniformly in $V(T_j)$. Since, $|V(T_j)| \le c|V(G_j)|$, we have $\mathbf{P}\big[o_j \in V(G_j)\big] \ge 1/c$. Proposition 2.1 implies that a subsequential limit of (T_j, o_j) is recurrent a.s. By Rayleigh monotonicity, G is recurrent a.s. □

3 Concluding Remarks

3.1 Limits of uniform spherical triangulations

Let $M \in [6, \infty]$; and let T_j^M be chosen randomly-uniformly among isomorphism classes of spherical triangulations with j vertices and maximum degree at most M, and given T_j^M

let o_j be chosen uniformly among the vertices of T_j^M.

Conjecture 3.1. *The distributional limit of* (T_j^M, o_j) *exists.*

This holds when $M = 6$, and then the limit is the hexagonal grid (since by Euler's formula there can be in this case at most 12 vertices with degree smaller than 6). After a first draft of the current paper has been distributed, a proof of the conjecture for the case $M = \infty$ (that is, no restriction on the degrees) has been obtained by Omer Angel and Oded Schramm. A paper with this result is forthcoming.

Assuming the conjecture for now, let (T^M, o) denote the limit random rooted triangulation. Let $p_n(T^M)$ denote the probability that the simple random walk starting from o will be at o at time n, given T^M. Following the discussion in the introduction, it should be believed that $p_n(T^M)$ decays like n^{-1} for almost all T^M. Theorem 1.1 shows that when $M < \infty$ the decay cannot be faster than n^α with $\alpha < -1$.

3.2 Intrinsic Mass Transport Principle

Proposition 2.2 implies that every distributional limit of unbiased random bounded degree triangulations of the sphere has at most two ends. (If $m \in \mathbb{N}$, the statement that a graph G has m ends is equivalent to the statement that m is the maximum number of infinite components of $G \setminus W$ as W ranges over finite subsets of $V(G)$. The definition of the space of ends is a bit more complicated, and can be found in many textbooks on point-set topology.) It is easy to verify that the proof applies also to the case where there is no uniform bound on the degrees, assuming that the limit graph is locally finite a.s. In fact, one can show that under the assumptions of Theorem 1.1, G has at most two ends. Moreover, if G is the distributional limit of unbiased finite rooted graphs, then a.s. if G is recurrent it has at most two ends. This can be proved using an intrinsic version of the Mass Transport Principle (MTP). The extrinsic version of MTP [Häg97, BLPS99] in its simplest form says that $\sum_{x\in\Gamma} f(x,y) = \sum_{x\in\Gamma} f(y,x)$ where Γ is a discrete countable group and $f : \Gamma \times \Gamma \to [0,\infty)$ is invariant under the diagonal action of Γ on $\Gamma \times \Gamma$, $\gamma : (g,h) \to (\gamma g, \gamma h)$.

We now briefly explain the intrinsic MTP. Suppose that $f(G,x,y)$ is a function which takes a graph G and two vertices x, y in G and returns a non-negative number. We assume that f is isomorphism invariant, in the sense that $f(\psi(G), \psi(x), \psi(y)) = f(G,x,y)$, when ψ is a graph isomorphism. Also assume that f is measurable on the appropriate space of connected graphs with two distinguished vertices. Then f will be called a **transport function**. (For example, if $g(G,v,u)$ is the degree of v if v neighbors with u in G and zero otherwise, then g is a transport function.) A probability measure μ on $\mathcal{X}$ is said to satisfy the intrinsic MTP (IMTP) if for every transport function f we have

$$\mathbf{E}\Big[\sum_v f(G,o,v)\Big] = \mathbf{E}\Big[\sum_v f(G,v,o)\Big], \tag{3.1}$$

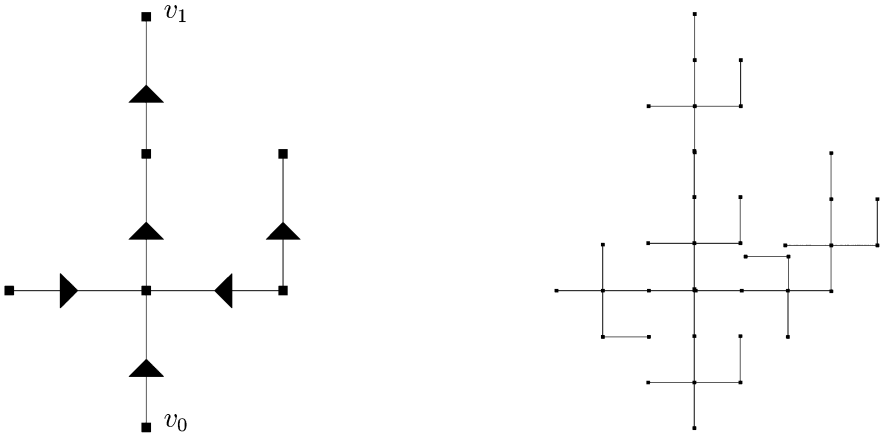


Figure 3.1: A subdivision rule for a tree.

where the sum on both sides is over all the vertices of G. It is easy to verify that every unbiased probability measure on finite graphs in $\mathcal{X}$ satisfies the IMTP. We claim that this also extends to weak limits of such measures. Indeed, let μ be a weak limit of Borel probability measures $\mu_1, \mu_2, \ldots$ on $\mathcal{X}$ satisfying the IMTP. Given a transport function f and some $k > 0$, let $f_k(G, v, u) := f(G, v, u)$ when $f \leq k$ and the distance from v to u in G is at most k, and $f_k(G, v, u) := 0$, otherwise. It is clear that both sides of (3.1) for μ and f_k are the limits of the corresponding terms for μ_j, hence they are equal. The fact that μ satisfies the IMTP now follows by taking the supremum with respect to k.

Many of the applications of MTP, such as appearing in [BLPS99], for example, can therefore be applied to (limits of finite) unbiased random graphs, we hope to further pursue this in a future work.

3.3 Examples of triangulations with uniform growth

It was observed by physicists [ADJ97] that for some random triangulations the average volume of balls of radius r is r^4. This looks surprising and calls for an intuitive explanation. We can't provide one, but to shed some light on the geometry of random surfaces, we can construct for every $\alpha > 1$ a triangulation of the plane for which every ball of radius r has r^α vertices, up to a multiplicative constant.

The simplest such construction is based on a tree. Consider a finite tree t_1 with two distinct marked vertices v_0, v_1, both having degree 1. Direct the edges of t_1 arbitrarily. Let t_2 be the tree obtained from t_1 by replacing each directed edge $[u_0, u_1]$ by a new copy of t_1, where v_0 replaces u_0 and v_1 replaces u_1. See Figure 3.1 for an example. Inductively, let t_n be obtained from t_{n-1} by replacing each edge of t_{n-1} with a copy of t_1. Note that the maximum degree in t_n is the maximum degree in t_1. Suppose that t_1 has k edges and the distance from v_0 to v_1 in t_1 is Δ. It is then clear that the diameter of t_n is Δ^n,

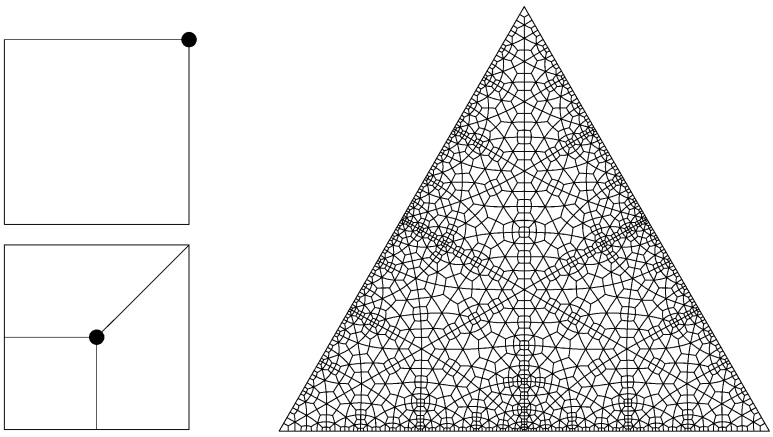


Figure 3.2: A subdivision rule and a resulting self-similar tiling.

up to an additive constant, and the number of edges of t_n is exactly k^n. It follows that for $m \leq n$ every ball of radius Δ^m in t_n has k^m edges, up to a multiplicative constant, because t_n is obtained by appropriately replacing each edge of t_{n-m} by a copy of t_m. Then every ball of radius $r \leq \mathrm{diam}(t_n)$ in t_n has about $r^{\log k/\log \Delta}$ vertices. We may pick a root in each t_n and take a subsequential limit, to obtain a tree t_∞ where every ball of radius r has about r^α vertices, where $\alpha = \log k/\log \Delta$.

One can easily modify the construction to obtain a similar tree with growth r^α where $\alpha > 1$ is not the ratio of logs of integers, by letting the replacement rule from t_{n-1} to t_n appropriately depend on n.

It is easy to make planar triangulations with similar properties. For example, suppose that the maximum degree in t_∞ is M, and let T be a triangulation of the sphere with at least M disjoint triangles. Then we may replace each vertex v of t_∞ by a copy T_v of T and for each edge $[v, u]$ in t_∞ glue a triangle in T_v to a triangle in T_u, in such a way that every vertex of T_v is glued to at most one other vertex. The resulting graph is the 1-skeleton of a planar triangulation, as required.

Another example which is somewhat similar but not tree-like appears in Figure 3.2. It is obtained by starting with a quadrilateral with a marked corner, subdividing it as in the figure to obtain three quadrilaterals with the interior vertex as the marked corner of each, and continuing inductively. The result is a map of the plane with quadrilateral faces and maximum degree 6.

These examples answer Problem 1.1 from [Bab97]

Acknowledgment. We thank Bertrand Duplantier for introducing us to the fascinating subject of random spherical triangulations.

References

[AAJ+98] Jan Ambjørn, Konstantinos N. Anagnostopoulos, Lars Jensen, Takashi Ichihara, and Yoshiyuki Watabiki. Quantum geometry and diffusion. *J. High Energy Phys.*, (11):Paper 22, 16 pp. (electronic), 1998.

[ADJ97] Jan Ambjørn, Bergfinnur Durhuus, and Thordur Jonsson. *Quantum geometry.* Cambridge University Press, Cambridge, 1997. A statistical field theory approach.

[ANR+98] Jan Ambjørn, Jakob L. Nielsen, Juri Rolf, Dimitrij Boulatov, and Yoshiyuki Watabiki. The spectral dimension of 2D quantum gravity. *J. High Energy Phys.*, (2):Paper 10, 8 pp. (electronic), 1998.

[Bab97] László Babai. The growth rate of vertex-transitive planar graphs. In *Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms (New Orleans, LA, 1997)*, pages 564–573, New York, 1997. ACM.

[Ber92] Marcel Berger. Les placements de cercles. *Pour la Science (French Scientific American)*, 176, 1992.

[BLPS99] I. Benjamini, R. Lyons, Y. Peres, and O. Schramm. Group-invariant percolation on graphs. *Geom. Funct. Anal.*, 9(1):29–66, 1999.

[BS96] I. Benjamini and O. Schramm. Harmonic functions on planar and almost planar graphs and manifolds, via circle packings. *Invent. Math.*, 126:565–587, 1996.

[dV91] Yves Colin de Verdière. Une principe variationnel pour les empilements de cercles. *Inventiones Mathematicae*, 104:655–669, 1991.

[GW00] Zhicheng Gao and Nicholas C. Wormald. The distribution of the maximum vertex degree in random planar maps. *J. Combin. Theory Ser. A*, 89(2):201–230, 2000.

[Häg97] Olle Häggström. Infinite clusters in dependent automorphism invariant percolation on trees. *Ann. Probab.*, 25(3):1423–1436, 1997.

[HS95] Zheng-Xu He and O. Schramm. Hyperbolic and parabolic packings. *Discrete Comput. Geom.*, 14(2):123–149, 1995.

[JS00] Johan Jonasson and Oded Schramm. On the cover time of planar graphs. *Electron. Comm. Probab.*, 5:85–90, 2000. paper no. 10.

[Koe36] P. Koebe. Kontaktprobleme der Konformen Abbildung. *Ber. Sächs. Akad. Wiss. Leipzig, Math.-Phys. Kl.*, 88:141–164, 1936.

[Mal99] V. A. Malyshev. Probability related to quantum gravitation: planar gravitation. *Uspekhi Mat. Nauk*, 54(4(328)):3–46, 1999.

[McC98] Gareth McCaughan. A recurrence/transience result for circle packings. *Proc. Amer. Math. Soc.*, 126(12):3647–3656, 1998.

[MP94] S. Malitz and A. Papakostas. On the angular resolution of planar graphs. *SIAM J. Discrete Math.*, 7:172–183, 1994.

[MT90] Gary L. Miller and William Thurston. Separators in two and three dimensions. In *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing*, pages 300–309. ACM, Baltimore, May 1990.

[RS87] Burt Rodin and Dennis Sullivan. The convergence of circle packings to the Riemann mapping. *J. Differential Geom.*, 26(2):349–360, 1987.

[Sac94] Horst Sachs. Coin graphs, polyhedra, and conformal mapping. *Discrete Math.*, 134:133–138, 1994.

[Sch99] Gilles Schaeffer. Random sampling of large planar maps and convex polyhedra. In *Annual ACM Symposium on Theory of Computing (Atlanta, GA, 1999)*, pages 760–769 (electronic). ACM, New York, 1999.

[Ste99] Kenneth Stephenson. Approximation of conformal structures via circle packing. In N. Papamichael, St. Ruscheweyh, and E. B. Saff, editors, *Computational Methods and Function Theory 1997, Proceedings of the Third CMFT Conference*, volume 11, pages 551–582. World Scientific, 1999.

[Tho99] Robin Thomas. Recent excluded minor theorems for graphs. In *Surveys in combinatorics, 1999 (Canterbury)*, pages 201–222. Cambridge Univ. Press, Cambridge, 1999.

[Tut62] W. T. Tutte. A census of planar triangulations. *Canad. J. Math.*, 14:21–38, 1962.

Commun. Math. Phys. 241, 191–213 (2003)
Digital Object Identifier (DOI) 10.1007/s00220-003-0932-3

Communications in
Mathematical Physics

Uniform Infinite Planar Triangulations

Omer Angel[1], **Oded Schramm**[2]

[1] Department of Mathematics, Weizmann Institute of science, Rehovot, 76100, Israel.
E-mail: omer@math.weizmann.ac.il
[2] Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA.
E-mail: schramm@microsoft.com

Received: 13 August 2002 / Accepted: 28 February 2003
Published online: 19 September 2003 – © Springer-Verlag 2003

Abstract: The existence of the weak limit as $n \to \infty$ of the uniform measure on rooted triangulations of the sphere with n vertices is proved. Some properties of the limit are studied. In particular, the limit is a probability measure on random triangulations of the plane.

1. Introduction

1.1. Motivation. What is a generic planar geometry?

There are many different planar geometries. The most commonly used one is the Euclidean plane, but is it generic? Is it more natural than, say, the hyperbolic plane?

For simplicity, consider discrete planar geometries (realized as planar graphs). Now there are still many choices. The lattice $\mathbb{Z}^2$ is the graph most commonly associated with planar geometry, but there is no a priori reason to prefer it over the triangular lattice, or any other lattice. One possible approach is based on convenience, preferring at each time the most convenient framework to work with. Even by that criterion no single geometry is always the best. Thus, some recent results are naturally adapted to the triangular lattice [31].

When we use a lattice, we force much more structure into our geometry than the topological condition of planarity necessitates. Random planar graphs, such as Delauny triangulations, have less enforced structure, but they still arise from the underlying Euclidean geometry. Is there a clear reason to prefer the Euclidean over the hyperbolic plane?

The approach used here is to consider a probability measure – in some sense a uniform measure – on planar geometries. Then we can ask what properties does a typical sample of that measure have. The way this is done is by considering discrete geometries, realized in the form of infinite planar triangulations, and finding an interesting distribution on them. Over finite planar triangulations the uniform measure is a natural choice. We prove the existence of a probability measure on infinite planar triangulations which is the limit

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_16, © Springer Science+Business Media, LLC 2011

of the uniform distributions on finite planar triangulations as their size tends to infinity. A sample of this measure is called the uniform infinite planar triangulation (UIPT). This model was suggested in [9], where Benjamini and Schramm show a.s. parabolicity of a wide class of distributions on infinite planar graphs under the condition of a uniform bound on the vertex degrees. Alas, the results there require the vertex degrees to be bounded, and hence do not apply to the UIPT.

The uniform *finite* planar triangulation and related objects have been studied by both combinatorists and physicists. Mathematical study is traced back to the 1960's with Tutte's attempts at the four color problem. In a series of papers Tutte was able to count the number of planar maps of a given size of various classes, including triangulations [32–35]. One of the conjectures he raised is that almost all planar maps are asymmetric, i.e., have no non-trivial automorphisms.

Tutte later proved his conjecture for a specific class of planar maps [36]. Random planar maps (and triangulations among them) have been studied extensively since then by others, proving Tutte's conjecture in a more general setting [29].

Previous research here focused on finite triangulations, but many of the results are about the asymptotic properties of planar maps and can be translated directly into claims about the infinite triangulations we study. Thus, there are results about the distribution of degrees in a uniformly chosen triangulation [18], the size of 3-connected components [8], and probabilistic 0-1 laws [7].

Schaeffer found a bijection between certain types of planar maps and labeled trees [30]. Chassaing and Schaeffer [13] recently used that bijection to show a connection between the asymptotic distribution of the radius of a random map and the integrated super-Brownian excursion. They deduce from this connection that the diameter of such a map of size n scales as $n^{1/4}$. While they work with planar quadrangulations and we with triangulations, it appears that such local differences are insignificant when large scale observations such as diameter, growth, separation, etc. are concerned. This phenomena is referred to as universality.

The physicists study such triangulations under the title of *2-dimensional quantum gravity*. The essential idea is to develop a quantum theory of gravity by extending to higher dimensions the concept of Feynman integrals on paths. Triangulations are used as a discretized version of a 2 dimensional manifold, and a function is averaged over all of them [3, 10]. More often physicists are interested not in the discretized planar triangulation but in a continuous scaling limit of it which is believed to exist.

Physicists introduced here the methods of random matrix models [15]. Through these methods and other heuristics many conjectures were made on the structure of such triangulations. In particular, it is believed that the Hausdorff dimension of the scaling limit of 2-dimensional quantum gravity is 4 [3]. For a good general exposition of quantum gravity see [2], as well as [1, 14].

Of particular interest is the KPZ relation [22] which relates critical exponents for a number of models on the plane and in 2 dimensional quantum gravity. This relation has been used to predict various exponents such as non intersection exponents for Brownian motion in the plane [16, 17]. Later a rigorous derivation of the same values was found using the SLE process [24–26].

Section 2 summarizes some results on counting triangulations which are the basis for much of what follows. Section 3 describes some properties of the UIPT that follow directly from the formulas for counting triangulations. In particular, it is shown that a.s. the UIPT has one end, i.e., the limiting process does not add any topological complications to the triangulation.

In Sect. 4 the existence of the limit distribution is proved. In Sect. 5 we give another characterization of the UIPT by a locality property. This roughly means that different regions in the triangulation are independent of one another and that each region is uniformly distributed among all triangulations of a given size (and hence the name uniform triangulation for the infinite graph). Section 6 describes a multi-type Galton-Watson tree naturally associated with a UIPT.

In Sect. 7 we show a relation between two types of infinite planar triangulations that demonstrates the universality principle. Through this relation we also get an infinite form of the main result of [8] (see also [6]).

In a forthcoming paper [4], an alternative method of constructing and sampling the UIPT is given. Using this method, it is shown there that up to polylogarithmic factors the UIPT has growth rate r^4, agreeing both with the heuristics for the Hausdorff dimension [3] and with the asymptotics for the radius of finite maps [13]. That paper also proves that the component of the boundary of the ball of radius r separating it from infinity has size roughly r^2. The method also enables an analysis of site percolation on the UIPT.

We proceed now to give formal definitions of the types of triangulations we study. An exact formulation of our main results will follow.

1.2. Definitions. The notion of a triangulation is very similar to the topological notion of a simplicial complex, although since we deal with the combinatorial aspects rather then the topological ones we will use a graph theoretic approach.

The notion of a triangulation has a bit of ambiguity around it. There are several variations on the definition, and they have much in common although there are some minor differences between them. The common thread to all variations is that a triangulation is a graph embedded in the sphere S^2 so that all faces are triangles. We will work with two types of triangulations.

Definition 1.1. *Consider a finite graph G embedded in the sphere S^2. A* **face** *is a connected component of $S^2 \setminus G$. The face is a* **triangle** *if its boundary meets precisely three edges of the graph. Similarly, a face is an* ***m*-gon** *if it meets m edges. A* **triangulation** *T is such a graph G together with a subset of the triangular faces of G.*

Let the **support** *$S(T) \subset S^2$ of T be the union of G and the triangles in T. Two triangulations T, T' are considered* **equivalent** *if there is a homeomorphism of $S(T)$ and $S(T')$ that corresponds T and T'. T is a triangulation of the sphere if $S(T) = S^2$. It is a triangulation of an m-gon if $S^2 \setminus S(T)$ is a single m-gon.*

For convenience, we usually abbreviate "equivalence class of triangulations" to "triangulation". This should not cause much confusion. The definition extends naturally to other manifolds, though we will not be concerned with that generality here.

Following the terminology found in [2] for types of triangulations, we define three classes of triangulations, types I, II and III. These differ according to which graphs are permitted in 1.1. In type I, there may be more than one edge connecting a pair of vertices, and loops (i.e., edges with both endpoints attached to the same vertex) are allowed as well. Type I triangulations will not be considered here, though some of the results (and proofs) apply to them as well.

Definition 1.2. *A* **type II triangulation** *is a triangulation where the underlying graph has no loops, but may have multiple edges.*

Definition 1.3. *A* **type III triangulation** *is a triangulation where the underlying graph is a simple graph (having no multiple edges or loops).*

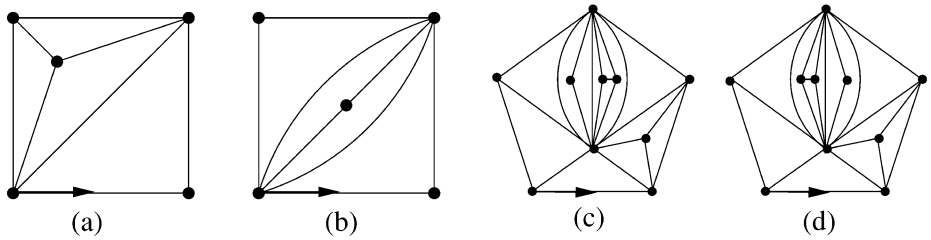


Fig. 1.1a–d. A type III and three type II triangulations. Triangulations **(a)** and **(b)** are triangulations of a square, while **(c)** and **(d)** are triangulations of a pentagon

Type II (resp. type III) triangulations are also referred to as 2-connected (resp. 3-connected) triangulations, since they are the triangulations with 2 or 3 connected underlying graphs.

If T is a triangulation of a domain in the plane which may have several holes (i.e., several boundary components), we will refer to the holes of the domain as *external* or *outer* faces of T. An external face may have 3 vertices on its boundary and then it is a triangle in itself. In that case that face is still distinguished from the triangles of T. In the case of type II, an external face can also have only 2 vertices on its boundary.

It is worthwhile noting that the circle packing theorem [23] gives a canonical embedding in the sphere (up to Moebius transformations) of a type III triangulation of the sphere.

The vertices of T lying on the boundary of its support $S(T)$ are called boundary vertices, and those in the interior of $S(T)$ are internal vertices. When we consider triangulations of a domain in the sphere with a number of boundary components we will usually fix the number of boundary vertices in each component as part of the domain. Thus, for example, a disc with m boundary vertices will be distinguished from a disc with $m' \neq m$ boundary vertices. Such a disc is referred to as an m-gon.

The size of a triangulation T, denoted $|T|$, is defined as the number of internal vertices. Since all faces are triangles, by Euler's characteristic formula, if E (resp. F) is the number of edges (resp. faces) of T, then $3|T| - E$ (resp. $2|T| - F$) is determined by the number and size of the boundary components of $|T|$. In particular, for a sphere all vertices are internal, and so $3|T| - E = 6$ and $2|T| - F = 4$.

Note that for a type III triangulation of the sphere (and even slightly more generally) the underlying graph determines the triangulation, i.e., whether any three edges form a triangle or not. When multiple edges are allowed there may be several distinct embeddings of the graph in the sphere giving distinct triangulations. E.g., in Fig. 1.1 c and d are distinct triangulations that have the same underlying graph.

A fundamental problem encountered when studying planar maps (triangulations included) is that of symmetries, namely that some maps have non-trivial automorphism groups. It seems plausible that most triangulations are asymmetric. While this has been proved [36, 29], we dispose of this problem in another manner. A simple way of eliminating any symmetries there is by adding a root to the triangulation.

Definition 1.4. *A* **root** *in a triangulation T consists of a triangle t of T called the* **root face***, with an ordering of its vertices (x, y, z). The vertex x is the* **root vertex** *and the directed edge (x, y) is the* **root edge**.

Note that in type II triangulations there may be more than one triangle with the same three vertices, so marking only the three vertices does not generally suffice. In

a triangulation of the sphere, if the root edge is given, then there are exactly two possibilities for the root. We will usually mark only the root edge as in Fig. 1.1, by an arrow.

When T has a boundary we will usually assume that the root edge lies on the boundary. Since a disc with m boundary vertices is referred to as an m-gon, triangulations (a) and (b) of Fig. 1.1 are of a square, while (c) and (d) are of a pentagon.

There are many possible variations on the definition of triangulation. Restricting to 2 or 3-connected underlying graphs (or even 4 or 5-connected) gives slightly different definitions. It is possible to restrict the degrees of vertices, or to allow faces that are not triangles. General convex polytopes can thus be described as a variation on the notion of triangulation. Most of the results proved below should have analogues for such generalizations, though the proofs do not always carry through. For convenience and brevity, we will deal with type II and type III triangulations here.

The definition of rooted infinite triangulations is very similar to that of finite triangulations. In that case G is infinite, of course, but we always require it to be locally-finite; that is, each vertex is incident to only finitely many edges. The only slightly technical point is that we require the embedding of G to be faithful to the combinatorial structure, in the following sense: if $\{p_n : n = 1, 2, \dots\}$ is a sequence in S^2 belonging to distinct edges of G, then accumulation points of p_n must be outside of $S(T)$.

A triangulation may be endowed with a metric in a number of ways. We will rather use a metric on the vertices of a triangulation – the graph metric induced by the underlying graph. It is also interesting to consider a triangulation as a metric space by having each face be isometric to an equilateral triangle with the shortest path metric on the whole triangulation. Then a triangulation of the sphere is a metric space homeomorphic to the sphere.

For either type, the space of finite and infinite (equivalence classes of) connected planar rooted triangulations is endowed with a natural topology described in [9]. (Note however, that there the root was only a vertex, leading to a closely related but slightly different distribution). A sequence of rooted triangulations converges to a triangulation T if eventually they are equivalent with T on arbitrarily large combinatorial balls around the root. This is a metric topology: e.g., set $d(T, T') = k^{-1}$, where k is the maximal radius such that the combinatorial balls of radius k around the roots are equivalent. In this topology, all finite triangulations are isolated points, and infinite triangulations are their accumulation points. We leave it to the reader to verify that limits of triangulations of the sphere are triangulations of domains in the plane.

Unlike in the setup of [9], the triangulation space is not compact. Consider the sequence T_n of triangulations where T_n contains two vertices of degree n with the same n neighbors forming a cycle (i.e., a double pyramid). Since T_n are distinct and all have diameter 2, $\{T_N\}$ has no convergent subsequence.

We will be interested in the uniform distributions on triangulations:

Definition 1.5. *τ_n^2 (resp. τ_n^3) is the uniform distribution on rooted type II (resp. type III) triangulations of the sphere of size n (i.e., having n vertices).*

The topology on the triangulation space induces a weak topology on the linear space of measures supported on planar triangulations. We study the distribution on infinite planar triangulations which is the weak limit of τ_n as $n \to \infty$. The weak limit may equivalently be defined as a limit with respect to neighborhoods of the root, i.e.,

Definition 1.6. *A measure supported on rooted triangulations τ is a weak limit of a sequence of measures τ_n if for any radius r and any finite triangulation T:*

$$\lim_{n\to\infty} \tau_n\left(B_r(0) = T\right) = \tau\left(B_r(0) = T\right),$$

where $B_r(0)$ is the ball of radius r around the root vertex (0) in the graph metric.

This definition is equivalent to the test-function definition that for every continuous function f on the space of triangulations:

$$\int f d\tau_n \to \int f d\tau.$$

If T_1 and T_2 are rooted triangulations, we say that T_1 is contained in T_2 (and write $T_1 \subset T_2$) if the two roots are the same and T_1 is contained in T_2 as unrooted triangulations. Sometimes we may also write $T_1 \subset T_2$ to mean that there is a triangulation isomorphic to T_1 contained in T_2.

Finally, a word on notation. By $X_n \sim Y_n$ we mean that $X_n/Y_n \to 1$. By $X_n \approx Y_n$ we mean that $\log X_n/\log Y_n \to 1$. We use $c, c_1, c_2, \ldots$ to signify constants, whose actual value may change from one formula to another.

1.3. Main Results. We will first prove that

Theorem 1.7. *There exists a probability measure τ^2 (resp. τ^3) supported on infinite planar triangulations of type II (resp. type III) such that*

$$\tau^i = \lim \tau_n^i.$$

Note. If τ_n are distributions on rooted graphs with uniformly bounded degrees, then since there are finitely many possibilities for $B_r(0)$, by compactness τ_n has a subsequential limit. In our case the triangulations do not have uniformly bounded degrees. We prove a limit exists – not just a subsequential limit. Additionally we prove that the limit is a probability measure; this is not a priori clear because of the lack of compactness.

Having defined the limit measure τ (we will often drop the type notation when results hold for either type) we turn to study the a.s. properties of a sample of τ. Denote such a sample by UIPT. A basic geometric property, one endedness, will show that the limit structure maintains the plane's topology. Recall the definition:

Definition 1.8. *A graph G is said to have* **one end** *(is one-ended) if for any finite subgraph H, $G \setminus H$ contains exactly one infinite connected component.*

Theorem 1.9. *The UIPT is a.s. one ended, and is therefore a triangulation of the plane.*

We also ask about the electrical type of the underlying graph. In [9] it is shown that for any sequence of distributions on planar graphs with degrees uniformly bounded by M, if a root is marked uniformly in each graph then every subsequential limit is a.s. recurrent. This holds, for example, for planar triangulations with uniformly bounded degrees. However, for those distributions it is not clear how to prove that the limit exists (simulations support this [11]). The following conjectures appear in [9]:

Conjecture 1.10. *For every $M \geq 6$, the distributions τ_N conditioned to have degrees uniformly bounded by M are weakly convergent.*

Conjecture 1.11. *The UIPT is a.s. recurrent.*

VEL parabolicity (for vertex extremal length) is a property of infinite graphs, closely related to circle packings for planar graphs. In graphs with bounded degrees it is equivalent to recurrence [21]. The proof in [9] of a.s. VEL parabolicity for uniform infinite triangulations with bounded degrees is still valid for the UIPT, with tightness (Lemma 4.4) filling the role of bounded degrees.

2. Counting

2.1. Classical results. Much of the analysis of triangulations is based on counting them. This is true both for finite triangulations and for infinite triangulations where the asymptotics of the finite triangulations come into play. The following counting results go back to Tutte [32] who counted various types of planar maps and triangulations. The results we use here are not due to Tutte but are derived using the same technique he uses. More details can be found in [12]. A good account of the technique including all results given here can be found in [20].

Theorem 2.1. *1. For $n, m \geq 0$, not both 0, the number of type II triangulations of a disc with $m+2$ boundary vertices and n internal vertices that are rooted on a boundary edge is*

$$\varphi^2_{n,m} = \frac{2^{n+1}(2m+1)!(2m+3n)!}{m!^2 n!(2m+2n+2)!}.$$

2. For $m \geq 1, n \geq 0$, the number of rooted type III triangulations of a disc with $m+2$ boundary vertices and n internal vertices that are rooted on a boundary edge is

$$\varphi^3_{n,m} = \frac{2(2m+1)!(4n+2m-1)!}{(m-1)!(m+1)!n!(3n+2m+1)!}.$$

The case $n = m = 0$ for type II triangulations warrants special attention. A triangulation of a 2-gon must have at least one internal vertex so there are no triangulations with $n = m = 0$, yet the above formula gives $\varphi^2_{0,0} = 1$. It will be convenient to use this value rather than 0 for the following reason. Typically, a triangulation of an m-gon is used not in itself but is used to close an external face of size m of some other triangulation by "gluing" the boundaries together. When the external face is a 2-gon, there is a further possibility of closing the hole by gluing the two edges to each other with no additional vertices. Setting $\varphi^2_{0,0} = 1$ takes this possibility into account.

Since we will consider the asymptotics of large triangulations we will need the following estimates of these numbers. Using the Stirling formula, as $n \to \infty$ we have the following:

$$\varphi^2_{n,m} \sim C^2_m \alpha_2^n n^{-5/2},$$

where $\alpha_2 = 27/2$ and

$$C^2_m = \frac{\sqrt{3}(2m+1)!}{4\sqrt{\pi}m!^2}(9/4)^m \sim C 9^m m^{1/2}.$$

For type III triangulations we have similar estimates:

$$\varphi^3_{n,m} \sim C^3_m \alpha_3^n n^{-5/2},$$

where $\alpha_3 = 256/27$ and as $m \to \infty$:

$$C_m^3 = \frac{2(2m+1)!}{6\sqrt{6\pi}(m-1)!(m+1)!}(16/9)^m \sim C(64/9)^m m^{1/2}.$$

Much of the time we will not distinguish between type II and type III triangulations. The type index will be dropped either when the stated results hold for both types or when it is clear which type is discussed.

We are interested in triangulations of the sphere that have no predefined boundary. The number of those is given by:

Proposition 2.2. *For either type, the number of rooted triangulations of the sphere with n vertices is $\varphi_{n-3,1}$.*

Proof. Adding a triangle that closes the outer face of a triangulation of a triangle makes a triangulation of the sphere. Alternatively, removing the triangle incident on the root edge that is not the root triangle gives a triangulation of a triangle rooted on the boundary. Thus, there is a bijection between triangulations of the sphere with n vertices and triangulations of a triangle with $n-3$ internal vertices. □

We will also be interested in triangulations of discs where the number of internal vertices is not prescribed. The following measure is of particular interest:

Definition 2.3. *The* **free distribution** *on rooted triangulations of an $(m+2)$-gon, denoted μ_m, is the probability measure that assigns weight*

$$\alpha^{-n}/Z_m(\alpha^{-1})$$

to each rooted triangulation of the $(m+2)$-gon having n internal vertices, where

$$Z_m(t) = \sum_n \varphi_{n,m} t^n.$$

As before, μ_m^2 (resp. μ_m^3) will denote free type II (resp. type III) triangulations, and similarly for the partition functions Z_m^2 and Z_m^3. Thus, the probability of a triangulation T, is proportional to $\alpha^{-|T|}$, and Z_m acts as a normalizing factor.

Note that by the asymptotics of φ as $n \to \infty$ we see that the sum defining Z converges for any $t \le \alpha^{-1}$ and for no larger t. The value of the partition functions will be useful. For this we have:

Proposition 2.4. *1. For type II triangulations, if $t = \theta(1-2\theta)^2$:*

$$Z_m^2(t) = \frac{(2m)!((1-6\theta)m + 2 - 6\theta)}{m!(m+2)!}(1-2\theta)^{-(2m+2)}.$$

2. For type III triangulations, if $t = \theta(1-\theta)^3$:

$$Z_m^3(t) = \frac{(2m)!((1-4\theta)m + 6\theta)}{m!(m+2)!}(1-\theta)^{-(2m+1)}.$$

At the critical point $t = \alpha^{-1}$ we will omit t. There Z takes the values:

$$Z_m^2 = Z_m^2(2/27) = \frac{(2m)!}{m!(m+2)!}\left(\frac{9}{4}\right)^{m+1},$$

and

$$Z_m^3 = Z_m^3(27/256) = \frac{2(2m)!}{m!(m+2)!}\left(\frac{16}{9}\right)^{m}.$$

The proof can be found as intermediate steps in the derivation of $\varphi_{m,n}$ in [20]. The above form may be deduced after a suitable reparametrization of the form given there.

2.2. Universality. While the exponential term in the asymptotics of φ is different for type II and III, the next term of $n^{-5/2}$ is the same. Similarities also occur in the asymptotics of C_m and of Z_m for the two types. Those similarities are not coincidental. It turns out that the asymptotic form is quite common when counting 2 dimensional structures. That form of the asymptotics is not dependent on the manifold, and is valid for any 2-dimensional manifold with or without boundaries. The same forms also appear when instead of triangulations other types of maps are considered, and was found to hold for a large variety of map types ([11, 10] and also the result of [13], related to our growth results). We therefore believe that many of the results here hold in a much more general context.

This universality is related to the basic property of the 2-sphere that a cycle partitions it into two parts, i.e., the Jordan Curve Theorem. This leads to a similarity between recurrence relations for different types of structures and through them to similar asymptotics for the solutions. For another instance of universality and some explanation see [5, 6].

It turns out that the exponential part of the asymptotics will cancel out often and when finer properties of infinite triangulations are considered the power term will come into play and determine the observed behavior.

2.3. Some estimates. We will need the following estimates throughout the paper.

Lemma 2.5. *Let*

$$S(k, N, a) = \sum_{\substack{n_1+\ldots+n_k=N \\ n_1,n_2>a}} \left(\prod n_i\right)^{-5/2}$$

then for any k there is a $c = c(k)$ such that for any N:

$$S(k, N, a) \le cN^{-5/2}a^{-3/2}.$$

Proof. Clearly

$$S(k, N, a) \le k! \sum_{\substack{n_1\ge\ldots\ge n_k \\ n_1+\ldots+n_k=N \\ n_1,n_2>a}} \left(\prod n_i\right)^{-5/2},$$

since each term in the sum over ordered k-tuples corresponds to at most $k!$ terms in the original sum, and less if there are any repetitions. Since each possible choice of

$n_2, \ldots, n_k$ determines a unique value for n_1 and always $n_1 \geq N/k$ we can replace n_1 by the smaller N/k and extend the range of summation.

$$\begin{aligned} S(k,N,a) &\leq k!(N/k)^{-5/2} \sum_{\substack{n_2 \geq \ldots \geq n_k \\ n_2 > a}} \left(\prod_{i \neq 1} n_i \right)^{-5/2} \\ &\leq k!(N/k)^{-5/2} \left(\sum_{n_2 \geq a} n_2^{-5/2} \right) \prod_{i>2} \left(\sum_{n_i} n_i^{-5/2} \right) \\ &\leq cN^{-5/2} a^{-3/2}. \end{aligned}$$

□

3. Basic Properties

3.1. Invariance with respect to the random walk. If we are given a finite triangulation, but not the location of the root, what can we say about the location of the root? The following proposition says: not much. For a triangulation T and a possible root r in T let T_r denote the triangulation T with r marked as root (if T is rooted then the old root is no longer marked).

Proposition 3.1. *Let T be a sphere triangulation chosen by τ_n, and r be a root in T chosen uniformly among all possible roots. Then T_r is uniformly distributed among all rooted triangulations (of size n).*

Proof. At first glance this seems trivial: since all rooted triangulations are equally likely no triangle in T should be more likely to be the root than any other. However, there is a subtlety here since there may be several triangles r such that the triangulations T_r are isomorphic. This occurs whenever T has a non trivial automorphism.

The key fact here is that any automorphism of T that preserves a root is necessarily the identity automorphism. If R is the set of possible roots and G is the automorphism group of T as an unrooted triangulation, then G acts naturally on R and a non identity element of G has no fixed points in R. Thus, the size of the orbit of a triangle $r \in R$ is just the size of G, regardless of r.

Since each of the orbits in the action of G on R corresponds to a distinct rooted triangulation, and each orbit has the same size, each possible triangulation is equally likely to result after a new root is selected. □

Note that since each directed edge can be completed in two ways to a root each directed edge is equally likely to be the root edge. From this we see that the UIPT must be invariant with respect to a random walk:

Theorem 3.2. *Let T be a triangulation chosen by τ_N for some N or by a subsequential limit τ. If x is the root vertex of T, y is a uniformly chosen neighbor of x, and (y, z, w) is a triangle in T uniformly chosen among all triangles including y, then $T_{(y,z,w)}$ has the same law as T.*

Proof. For finite N, if a vertex x of degree d is the root vertex, then there are d possibilities for the root edge (and $2d$ options for the root). It follows that the probability that x is the root is proportional to its degree. This is the stable distribution for the random

walk on the graph of T, so as a consequence of Proposition 3.1 we see that $T_{(y,z,w)}$ has the same law as T.

Since this is true for every τ_N, the same holds for any subsequential limit. □

3.2. One endedness. We start with a lemma describing the behavior of a triangulation on a disjoint union of discs.

Lemma 3.3. *Given k disjoint polygons (with given boundary sizes) and a triangulation T of the polygons, let n_i be the number of internal vertices in the ith polygon. Then*

$$\left|\left\{T \mid \sum n_i = N \wedge \exists i, j,\ i \neq j,\ n_i, n_j > a\right\}\right| < C\alpha^N N^{-5/2} a^{-3/2},$$

where C depends only on the number and sizes of the boundaries of the polygons.

Proof. We prove that the number of triangulations where $n_1, n_2 > a$ is small, as required. By symmetry, the number for any other pair (i, j) has the same bound. Since the number of such pairs, $\binom{k}{2}$, does not depend on a or on N, this suffices.

We use the upper bound $\varphi_{n,m} \leq \beta_m (n+1)^{-5/2} \alpha^n$ (+1 is only necessary to account for $n = 0$, and is not essential). Assume the ith domain has boundary size $m_i + 2$. The number of triangulations we wish to bound is:

$$\begin{aligned}
\sum_{\substack{n_1+\ldots+n_k=N \\ n_1,n_2>a}} \prod_i \varphi_{n_i,m_i} &\leq \sum_{\substack{n_1+\ldots+n_k=N \\ n_1,n_2>a}} \prod_i \beta_{m_i} (n_i+1)^{-5/2} \alpha^{n_i} \\
&= \alpha^N \prod \beta_{m_i} \sum_{\substack{n_1+\ldots+n_k=N \\ n_1,n_2>a}} \prod (n_i+1)^{-5/2} \\
&\leq c_1 \alpha^N \sum_{\substack{n_1+\ldots+n_k=N+k \\ n_1,n_2>a}} \prod (n_i)^{-5/2} \\
&\leq c_2 \alpha^N N^{-5/2} a^{-3/2},
\end{aligned}$$

where at the end we used Lemma 2.5. □

Generally a limit T of a sequence of finite sphere triangulations need not have support $S(T)$ which is homeomorphic to the sphere or even the plane. While the limit is still planar, when embedded in the sphere $S(T)$ may have any number of accumulation points. One accumulation point gives a punctured sphere, i.e., the plane. More than one means that $S(T)$ has a more complicated topological structure; it is no longer simply-connected.

Corollary 3.4. *Every subsequential limit of τ_N a.s. has one end.*

Proof. Suppose that a subsequential limit τ has more than one end with positive probability. Then for some k and some $\varepsilon > 0$ the probability that a loop of length k including the root partitions a sample of τ into two infinite parts is at least ε. This implies that for any a for infinitely many N the τ_N-probability of having a loop of length k including the root that has at least a vertices on either side is at least $\varepsilon/2$. Call such a loop a separating loop.

Count pairs (T, L) with T a triangulation of size N and L a separating loop included in T. From Lemma 3.3 we know that the total number of such pairs is $O(\alpha^N N^{-5/2} a^{-3/2})$. However, the number of sphere triangulations with N vertices is $\varphi_{N-3,1} \sim C\alpha^N N^{-5/2}$, and by dividing we deduce that the expected number of separating loops is $O(a^{-3/2})$. In particular as $a \to \infty$ the probability that a separating loop exists tends to 0. □

4. Existence of the Limit

4.1. Tightness. The difficulty in establishing the existence of the limit distribution as a probability measure is showing that the size of the ball of radius r around the root is tight with respect to the total number of vertices N. Recall that a family of random variables X_n is tight with respect to n if

$$\lim_{t\to\infty} \mathbb{P}(|X_n| > t) = 0$$

uniformly with respect to n.

To this end, we first prove the following estimates for the degree of the root in either type of triangulation. While the lemmas are very similar in nature, the methods of proof given here are different. This demonstrates the underlying unity of the different models, while local differences make some techniques applicable in one and others in another.

The following lemmas appear in similar form in [18].

Lemma 4.1. *Denote the degree of the root vertex by d_0. For any $\varepsilon > 0$ there is a $c = c(\varepsilon)$ such that*

$$\tau_N^3(d_0 = k) < c\left(\frac{3}{4} + \varepsilon\right)^k,$$

uniformly for all N and

$$\tau_N^3(d_0 = k) \xrightarrow[N\to\infty]{unif} \frac{2(2k-3)!}{(k-3)!(k-1)!}\left(\frac{3}{16}\right)^{k-1}.$$

Proof. A type III triangulation of the sphere where the root vertex has degree k is the union of two triangulations, T_0, T_1 whose intersection is a k-gon: T_0 contains the root vertex and k triangles connecting it to the sides of the k-gon, and T_1 contains all other triangles. The root triangle has one edge in the intersection of T_0 and T_1. Choose this edge to be the root edge of T_1. Now $T \leftrightarrow T_1$ is a bijection between rooted triangulations of the sphere with $d_0 = k$ and rooted triangulations of a k-gon with the root edge on the boundary.

If $|T| = N$, then $|T_1| = N - k - 1$, and we know the number of such triangulations. Dividing by the number of sphere triangulations, we get:

$$\begin{aligned}\tau_N(d_0 = k) &= \frac{\varphi_{N-1-k,k-2}}{\varphi_{N-3,1}} \\ &= \frac{2(2k-3)!}{3(k-3)!(k-1)!}\,\frac{(4N-2k-9)!(3N-6)!(N-3)!}{(4N-11)!(3N-k-6)!(N-k-1)!} \\ &\to \frac{2(2k-3)!}{(k-3)!(k-1)!}\left(\frac{3}{16}\right)^{k-1}.\end{aligned}$$

To prove the uniform exponential bound consider the ratio

$$\frac{\tau_N(d_0 = k+1)}{\tau_N(d_0 = k)} = \frac{(2k-1)(2k-2)(3N-k-6)(N-k-1)}{k(k-2)(4N-2k-9)(4N-2k-10)} < \frac{3}{4} + \varepsilon$$

for any sufficiently large N, k. □

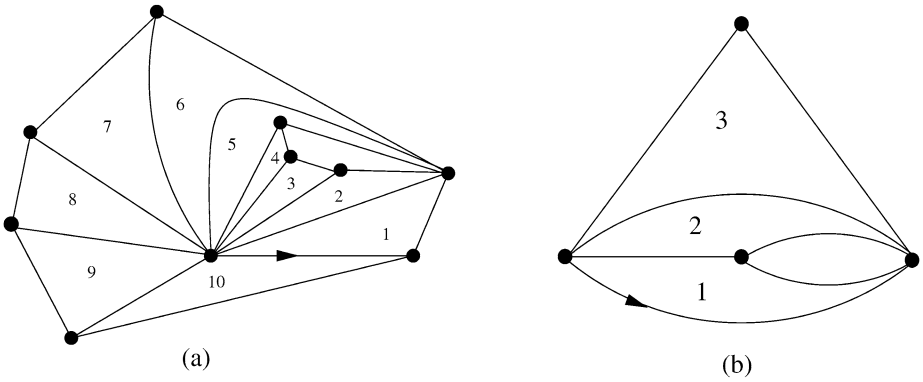


Fig. 4.1a, b. Proof of the tightness of the root's degree

For type II, since multiple edges are present, there are two notions of degree. The vertex degree of v is the number of neighbors it has, while the edge degree of v is the number of edges incident on it. For our purposes bounding the vertex degree is sufficient, but in what follows we bound the larger edge degree.

Lemma 4.2. *Denote the edge degree of the root vertex by d_0, then there is a c such that*

$$\tau_N^2(d_0 = k) < c\left(\frac{5}{3\sqrt{3}}\right)^k$$

uniformly for all N.

Proof. Let $t_1, \ldots, t_{d_0}$ be the triangles incident with the root vertex, ordered counter-clockwise starting with the root triangle t_1. For $s \leq d_0$ let T_s be the sub-triangulation including triangles $t_1, \ldots, t_s$. Adding t_i one at a time, we consider the distribution of T_{s+1} conditioned on T_s, and show that for any T_s there is a probability bounded away from 0 that $d_0 = s + 1$.

T_s may have several external faces. One of those, say F, includes the root vertex, and t_{s+1} is in F. In order for t_{s+1} to be the last triangle adjacent to the root vertex it must include the two edges of F on either side of the root vertex. Thus, in Fig. 4.1a, the triangles incident with the root vertex are numbered. The triangle t_{10} is the final triangle, and it includes both the edge from t_9 and the edge from the root triangle t_1. Note that when triangle t_5 is added, an unknown part of the triangulation is enclosed, but this will not effect the bounds we get.

To bound from below the probability that t_{s+1} is the last triangle conditioned on T_s, assume that the boundary of F has size $m+2$. A first condition on the events that the part of the triangulation inside F has n vertices. The number of possible ways to triangulate F under these constraints is $\varphi_{n,m}$. If t_{s+1} is the last triangle around the root, then adding it leaves a face of boundary size $m + 1$ with n internal vertices. Thus the probability of the next triangle being the last one is:

$$\begin{aligned}\frac{\varphi_{n,m-1}}{\varphi_{n,m}} &= \frac{m^2(2n + 2m + 1)(2n + 2m + 2)}{2m(2m + 1)(3n + 2m - 1)(3n + 2m)} \\ &> \frac{2m(n + m)^2}{(2m + 1)(3n + 2m)^2}.\end{aligned}$$

If $m > 0$, then this is at least 2/27, so the probability that $d_0 > s+1$ is at most 25/27. Since this bound is uniform it also holds when conditioning only on T_s and not on the number of internal vertices in F.

Thus, as new triangles are revealed, each triangle has a probability of at least 25/27 of being the last one, unless $m = 0$. If $m = 0$, as for T_2 in Fig. 4.1b, then after a triangle is added we must have $m = 1$ and so out of every two consecutive s, at least one has $m > 0$. It follows that the probability of having more than k edges leaving the root vertex is at most $(25/27)^{(k-1)/2}$, as claimed. □

Note. For type III triangulations, Lemma 4.1 gives the exact probability of any given degree in the UIPT. To a large extent, this is possible because the radius 1 neighborhood of the root has a simple structure. When multiple edges are allowed, even the ball of radius 1 around the root can have a complicated structure, making an exact calculation harder to get. On the other hand, for type II triangulations, we can calculate the exact probability that a certain triangle is present in the triangulation conditioned on some sub-triangulation (e.g., the probability that t_{s+1} is the last triangle around the root conditioned on T_s, as in the proof). This is much harder to do for type III triangulations, because we need to keep track of which pairs of vertices already have edges between them, whereas in type II triangulations adding another edge is always legal.

At this point we will rigorously define the ball B_r of radius r around the root (or any other vertex, for that matter). This ball is a sub-triangulation, but there is some subtlety in its definition. The vertices of B_r are all those vertices at distance at most r from the root vertex, but not all edges and triangles between these vertices are necessarily part of B_r.

Definition 4.3. *B_0 is just the root vertex itself. B_{r+1} is composed of all triangles incident on any vertex of B_r together with their vertices and edges.*

Note that there may be edges between vertices on the boundary of B_r that are not part of B_r itself. Next, we turn our attention to the size of the ball B_r. The following lemma holds for both types.

Lemma 4.4. *For any fixed r the random variables $M_r = \max\{d_v \mid v \in B_r\}$ (i.e., the maximal degree in B_r) defined on the space of triangulations with measure τ_N are tight with respect to N.*

Proof. For $r = 0$, B_r is just the root, and Lemmas 4.1 and 4.2 show that the degree of the root is tight with respect to N for either type.

We proceed by induction on r. Suppose that M_r is tight with respect to N. To show that M_{r+1} is also tight we use Theorem 3.2. Let T denote a sample of τ_n, and let $X_0, X_1, \ldots$ be a simple random walk on T started at the root vertex X_0. Denote by $\mathbb{P}$ the resulting probability measure on triangulations with paths beginning at the root. It follows from Theorem 3.2 that for any i the degree of X_i has the same distribution as the degree of the root. Fixing $M' > M > 0$ we estimate the probability that $M_r \le M$ and yet $M_{r+1} > M'$. Conditioned on this event, there is at least one vertex $u \in B_{r+1} \setminus B_r$ with $d_u > M'$. Since there is a path of length $r+1$ from the root vertex to u, and all vertices on the path have degrees at most M,

$$\mathbb{P}(d_{X_{r+1}} > M' \mid M_r \le M < M' < M_{r+1}) \ge M^{-(r+1)},$$

and so

$$\mathbb{P}(d_{X_{r+1}} > M') \ge M^{-(r+1)} \mathbb{P}(M_r \le M < M' < M_{r+1}).$$

By Theorem 3.2 the LHS does not depend on r and is simply $\tau_n(d_0 > M')$. The RHS does not depend on the random walk either, so for any M,

$$M^{r+1}\tau_n(d_0 > M') \geq \tau_n(M_{r+1} > M') - \tau_n(M_r > M).$$

By induction, for all $\varepsilon > 0$ we may choose $M = M(\varepsilon)$ such that $\tau_n(M_r > M) \leq \varepsilon/2$ for all n. Then we take $M' = M'(\varepsilon) > M$ sufficiently large so that $\tau_n(d_0 > M') < M^{-(r+1)}\,\varepsilon/2$ for all n. This gives $\tau_n(M_{r+1} > M') < \varepsilon$ for all n, and completes the proof. □

Corollary 4.5. *For any fixed r the random variables $|B_r|$ are tight with respect to N.*

Proof. Since M_r is tight with respect to N, and $|B_r| < (1 + M_r)^r$, $|B_r|$ is tight as well. □

Corollary 4.6. *Every subsequential limit of τ_N is a probability measure.*

In [28] it is shown that for every finite triangulation T there is a constant c such that asymptotically, in almost every sphere triangulation of size n the number of times T appears is roughly cn. This $c(T)$ is roughly the probability that a neighborhood of the root in the UIPT is isomorphic to T. In fact, the result of [28] is stronger, since it gives not just an annealed probability of seeing T but that the quenched probability is constant. We bring here a simpler calculation just for the annealed probability, since the results of the calculation are useful in what follows.

It will be easier to work with rooted triangulations having the property that if they are a sub-triangulation of the UIPT, then they appear in it exactly once. This is not always the case: a root triangle together with a cycle of some length may appear in the triangulation in several different ways.

Definition 4.7. *A rooted triangulation A is* **rigid** *if it is connected and no triangulation includes two distinct copies of A with coinciding roots.*

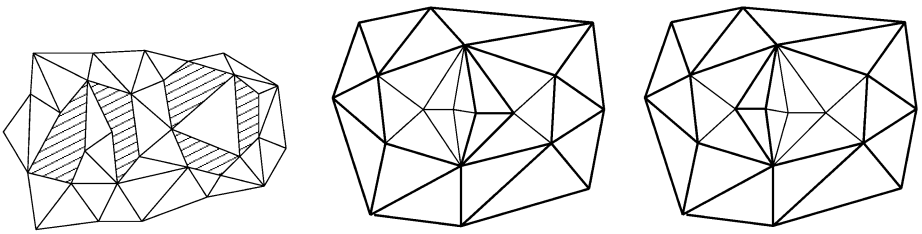
The balls B_r of a triangulation are rigid, as is evident from the following sufficient criterion for rigidity (the proof is left to the reader).

Lemma 4.8. *If in the dual graph of triangulation A the vertices corresponding to the triangles of A form a connected set, and every vertex of A is incident on a triangle, then A is rigid.*

This criterion is not necessary for rigidity, as is demonstrated by Fig. 4.2 a, where there is an isolated triangle. In fact a sufficient and necessary criterion is that the support $S(T)$ be 3-connected. In order to complete a planar triangulation to a sphere triangulation we need to fill each of its external faces with some triangulation. The advantage of rigid triangulations is that filling the external faces in different ways must lead to distinct sphere triangulations, whereas for non-rigid triangulations different ways of filling the faces may give rise to the same complete triangulation. Figure 4.2 b, c give an example of a non-rigid triangulation and how two completions give rise to the same triangulation.

A second consequence of the construction of B_r, whose proof (an application of the Jordan curve theorem) is left to the reader is:

Lemma 4.9. *In the ball B_r there are no edges between two vertices of any external face except those making the face itself.*



(a) (b) (c)

Fig. 4.2a–c. A rigid triangulation (with shaded outer faces) and two isomorphic completions of a non-rigid triangulation

Proposition 4.10. *Let A be a rigid rooted triangulation having no edges between two vertices of an external face except those making the face itself. Assume A has n vertices, some of which are on k external boundary components of sizes $m_1 + 2, \ldots, m_k + 2$. Then every subsequential limit τ of τ_N has:*

$$\tau(A \subset T) = \frac{\alpha^{3-n}}{C_1}\left(\prod Z_{m_i}\right)\sum \frac{C_{m_i}}{Z_{m_i}}.$$

Moreover, the probability that the ith face is the infinite one corresponds to the ith term in the sum, i.e.:

$$\frac{\alpha^{3-n}}{C_1} C_{m_i} \prod_{j \neq i} Z_{m_j}.$$

Note. For type II triangulations the restriction on edges between vertices of an external face is not necessary. For type III triangulations it is needed, since when such an edge exists it imposes restrictions on the component inside the face. In general, the probability that $A \subset T$ can be found using the proposition together with the inclusion-exclusion principle. The requested probability is a linear combination of a fixed number of terms and each of them has a limit as above.

Proof. Let τ be a subsequential limit of τ_N. Denote by $Q = Q(A, n_2, \ldots, n_k)$ the event that $A \subset T$ (with the A's root corresponding to T's root) and that the part of T in the ith external face of A contains n_i internal vertices. This is defined in the finite as well is the infinite setting (though we keep $n_2, \ldots, n_k < \infty$). In what follows n_1 denotes the number of vertices in the 1st external face, i.e., $n_1 = N - n - \sum_{i>1} n_i$. The probability of Q is:

$$\tau_N(Q) = \frac{\prod_{i=1}^{k} \varphi_{n_i,m_i}}{\varphi_{N-3,1}},$$

and $\tau_N(A \subset T)$ is the sum over all possible vectors (n_i) of this probability.

We first consider the limit:

$$\lim_{N\to\infty} \tau_N(Q(A, n_2, \ldots, n_k)) = \left(\prod_{i>1} \varphi_{n_i,m_i}\right) \lim_{N\to\infty} \frac{\varphi_{n_1,m_1}}{\varphi_{N-3,1}}$$

$$= \left(\prod_{i>1} \varphi_{n_i,m_i}\right) \lim_{N\to\infty} \frac{C_{m_1} n_1^{-5/2} \alpha^{n_1}}{C_1 (N-3)^{-5/2} \alpha^{N-3}}$$
$$= \left(\prod_{i>1} \varphi_{n_i,m_i}\right) \frac{C_{m_1} \alpha^{-n-\sum_{i>1} n_i}}{C_1 \alpha^{-3}}.$$

Since the limit exists, it equals $\tau(Q)$. This may be written as:

$$\tau(Q) = \frac{\alpha^{3-n}}{C_1} C_{m_1} \prod_{i>1} \varphi_{n_i,m_i} \alpha^{-n_i}. \tag{4.1}$$

Of course, a similar expression holds when the role of the 1st face is filled by some other face, i.e., the sizes of all but the ith face are fixed.

Let $R_i = R_i(A)$ denote the event that $A \subset T$ and all the external faces of A except possibly the ith one contain finitely many vertices. Obviously,

$$R_1 = \bigcup_{n_2,\ldots,n_k<\infty} Q(A, n_2, \ldots, n_k).$$

Using (4.1) we get for any subsequential limit τ of τ_N:

$$\tau(R_1) = \frac{\alpha^{3-n}}{C_1} C_{m_1} \prod_{i>1} Z_{m_i}, \tag{4.2}$$

and a similar formula for R_j, $j > 1$. It is clear that $\tau(R_i \cap R_j) = 0$ for $i \neq j$ in $\{1, \ldots, k\}$. Moreover, Corollary 3.4 (one end) implies $\tau\big(\{A \subset T\} \setminus \cup_i R_i\big) = 0$. Hence,

$$\tau(A \subset T) = \sum_{i=1}^{k} \tau(R_i) = \frac{\alpha^{3-n}}{C_1} \left(\prod Z_{m_i}\right) \sum \frac{C_{m_i}}{Z_{m_i}}.$$

□

Proof of Theorem 1.7. Since for any r the size of B_r is tight in N it suffices to show that for any possible ball A the probability $\tau_N(B_r = A)$ converges to some limit as $N \to \infty$. However, we know that the ball B_r satisfies the conditions of Proposition 4.10, and for any triangulation satisfying the conditions of Proposition 4.10 the limit exists. □

Proof of Theorem 1.9. This follows from Theorem 1.7 and Corollary 3.4. □

5. Locality

Next, we look at another basic property of the UIPT, namely locality. The meaning of locality is that isolated regions of the UIPT are almost independent. In the following, $R_i = R_i(A)$ will denote the event defined in the proof of Proposition 4.10.

Theorem 5.1. *Let A be a finite rigid triangulation (for type III, with no edges between vertices on external faces). Assume A has k external faces of sizes $m_1 + 2, \ldots, m_k + 2$. Condition on the event $R_i(A)$, and let T_j denote the component of the UIPT in the jth face. Then:*

1. *The triangulations T_j are independent.*
2. *T_i has the same law as the UIPT of an $(m_i + 2)$-gon (that is, the $N \to \infty$ limit of the uniform measure on rooted triangulations of an $(m_i + 2)$-gon with N internal vertices).*
3. *For $j \neq i$, T_j has the same law as the free triangulation of an $(m_j + 2)$-gon.*

Proof. Without loss of generality, assume $i = 1$. From Eqs. 4.2 and 4.1 we see that

$$\tau(R_1) = \frac{\alpha^{3-n}}{C_1} C_{m_1} \prod_{j>1} Z_{m_j}$$

and so

$$\tau\big(|T_j| = n_j \text{ for } j > 1 \mid R_1\big) = \prod_{j>1} \frac{\varphi_{n_j,m_j} \alpha^{-n_j}}{Z_{m_j}}.$$

Thus, we see that conditioned on $R_1(A)$ the sizes of the T_j's are independent, and $|T_j|$ is distributed like the free triangulation of an $(m_j + 2)$-gon. Consider τ_N. Conditioned on $Q(A, n_2, \ldots, n_k)$, since all possible triangulations of the sphere with the prescribed component sizes are equiprobable, the same holds for each component T_i. Thus, for any N the joint distribution of $(T_1, \ldots, T_j)$ conditioned on their sizes is a product distribution. As $N \to \infty$, these joint distributions converge to the product distribution, where T_j is uniform on triangulations with $|T_j| = n_j$.

Finally, the marginal of T_1 has size tending to infinity, and so converges to the UIPT of an $(m_1 + 2)$-gon. □

6. Ball Structure

Recall that Theorem 5.1 tells us that conditioned on a sub-triangulation T, with some external faces, the probability that a face of size $m + 2$ is the infinite one is proportional to $\frac{C_m}{Z_m}$. In the case of type II or III triangulations we have:

$$\frac{C_m^2}{Z_m^2} = \frac{(m+1)(m+2)(2m+1)}{3\sqrt{3\pi}},$$

$$\frac{C_m^3}{Z_m^3} = \frac{m(m+2)(2m+1)}{6\sqrt{6\pi}},$$

so in either case the probability of a face of size m being the infinite face is roughly proportional to m^3.

We wish to study the relation between the ball of radius r and the ball of radius $r + 1$. The ball of radius r is a finite triangulation with any number of external faces with any combination of boundary sizes. Moving to $r + 1$ we add in each outer face some triangles around its circumference. These added triangles can fill up the face, or they can split that face up into a number of sub-faces of different sizes. Figure 6.1a shows a ball with several finite faces, and the layer of the triangulation between radius r and $r + 1$ in the finite faces. The shaded areas are some of the faces of the ball of radius $r + 1$. The infinite face may contain additional sub-faces.

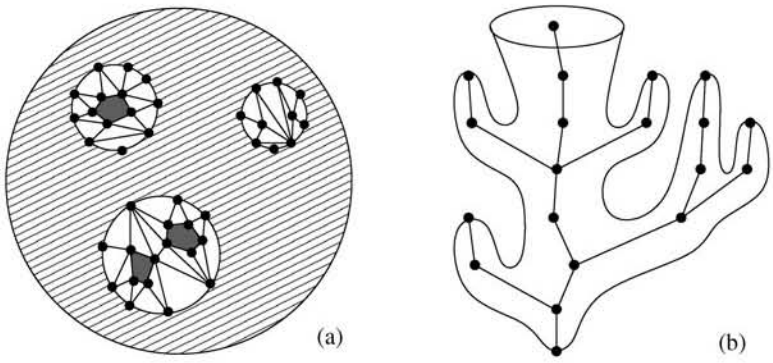


Fig. 6.1. **(a)** A possible ball in a planar triangulation. **(b)** A tree corresponding to a surface. Height corresponds to distance from the root

This gives rise to a tree-like structure for the triangulation, as in Fig. 6.1b. Each outer face of the ball of radius r corresponds to a vertex in the rth level of the tree. The face corresponding to a child is contained in the face of the parent vertex. An infinite triangulation will yield an infinite tree. Similarly, if a triangulation is one ended, then so is the corresponding tree, i.e., the tree is composed of a single infinite branch from the root with finite sub-trees growing from it. Note that while any triangulation determines a tree, the converse is false. The tree does not determine the triangulation.

A vertex in the tree corresponds to an external face of some triangulation, so there are different types of vertices depending on the face sizes. Labeling each vertex with the boundary size of the corresponding face, we see that the UIPT gives rise to a multi-type tree process.

From Theorem 5.1 we see that if we condition the first r levels of the tree and on which vertex in the rth level is in the infinite branch, then Theorem 5.1 tells us that the remaining sub-trees are independent. Thus, we see that at each level, one vertex, with a known distribution, has an infinite sub-tree above it, and the others have independent numbers of offspring of independent types. The tree process is thus just a multi-type Galton Watson process conditioned to survive. Without the conditioning we get the tree corresponding to a free triangulation of the sphere, which we know to be a.s. finite. However, since the free process has a power tail on its size, it is critical. Thus, the above description is just the construction of a critical Galton Watson process conditioned on survival (see [27]).

7. Type Relations

The two types of UIPT are part of a wider class of random planar objects satisfying common properties. This was first hinted at by the universality of the asymptotic formulas for counting various planar objects. Between type II and type III triangulations there is a more fundamental relation, enabling us to find a direct transformation between type II and type III triangulations. A similar transformation also holds between type I and type II triangulations as well as other pairs of classes of planar objects.

Roughly, the idea for passing from a type II triangulation to a type III triangulation is to take each double edge and to remove all the triangles and vertices inside it. The two edges are then glued together to get again a triangulation of the plane or sphere as the

case may be. Conversely, to get from a type III triangulation to a type II one, we will take each edge and replace it with a double (or multiple) edge with some distribution on the triangulation inside the resulting 2-gons. Recall that we allowed the triangulation of the 2-gon with no internal vertices, and gluing it in a 2-gonal outer face meant gluing the two edges together. Thus, with some probability ($1/Z_0 = 8/9$ actually) this empty triangulation is used and the edge remains a single edge.

Both directions pose some difficulties. A 2-gon partitions a triangulation to two components. How do we decide which is the inside and which the outside? In an infinite triangulation of the plane we wish to contract the finite side, but for a finite triangulation of the sphere it is not so clear. Also, there is the possibility that the root of the triangulation is deleted in this way, and then a new root is needed.

In the opposite direction, there is the question of the distribution for the triangulation of the 2-gon added. The natural candidate in the infinite case is the free triangulation of the 2-gon. Again, in the finite setting things are more delicate. Since we define the infinite triangulation measures as limits of the finite ones, we need to find some transformation of the finite measures first.

For a type II triangulation T, when we contract 2-gons as described above, until no double edges remain, the result is a maximal (with respect to inclusion) 3-connected sub-triangulation, since the only way 2 vertices could separate the graph is by forming a 2-gon. Therefore, the transition can be summarized as taking a single maximal 3-connected subgraph of the triangulation. The natural choice for this is to take the 3-connected component containing the root triangle. This also saves us the trouble of choosing a new root in the case that the old root was in one of the contracted 2-gons. Note that the root vertex or even the root edge is not enough, since 2 vertices may be in the intersection of two distinct 3-connected components. However the 3 vertices of any triangle determine a unique 3-connected component.

Definition 7.1. *Let T be a rooted type II triangulation. Define $\tilde{T}$ to be the type III triangulation composed of the 3-connected component of the root in T, with all double edges identified into single edges. For a measure ν on rooted type II triangulations let $\tilde{\nu}$ be the resulting measure on type III triangulations, i.e., for any event R:*

$$\tilde{\nu}(R) = \nu\big(\{T \mid \tilde{T} \in R\}\big).$$

This operation is known as taking the core of a structure [6]. In general, for two classes of rooted planar objects, one more restricted than the other, the core of a member of the wider class is its largest partial structure containing the root included in the smaller class (when it is unique).

Lemma 7.2. *For any finite N, for some coefficients $a_{n,i}$:*

$$\widetilde{\tau_N^2} = \sum_{i \le N} a_{N,i} \tau_i^3.$$

In the limit, for some constants $a_i, a_\infty > 0$:

$$\widetilde{\tau^2} = a_\infty \tau^3 + \sum a_i \tau_i^3.$$

In the infinite case this means that the 3-connected component of the root is either a finite sphere triangulation with some distribution on the size where all triangulations of the same size are equiprobable, or it is an infinite triangulation. Conditioned on the latter case it is just the infinite type III UIPT. The asymptotics of the coefficients $a_{n,i}$ are described in [6].

Proof. Consider first the finite case. All we need to show is that any two type III triangulations have the same probability of appearing as the 3-connected component of the root in τ_N^2, i.e., that for any triangulation U the number of triangulations T with $|T| = N$ and $\tilde{T} = U$ depends only on $|U|$.

This is clear, since any two triangulations of the same size have the same number of edges. Specifically, U has $3|U| - 6$ edges. Formally, if $Z_0^2(x) = \sum_n \varphi_{0,n}^2 x^n$ is the generating function for triangulations of a 2-gon, then the number of ways U can come about is the coefficient of $x^{N-|U|}$ in $(Z_0(x))^{3|U|-6}$, which is, of course, determined by $|U|$.

The infinite case follows from the finite case by taking a weak limit. The map $T \to \tilde{T}$ is continuous with respect to the topology on the spaces of type II and III triangulations. Since τ_N^3 is supported on triangulations with N vertices, they have disjoint supports for distinct N. Therefore, necessarily:

$$\begin{aligned}\widetilde{\tau^2} &= \lim_{N\to\infty} \widetilde{\tau_N^2} \\ &= \lim_{N\to\infty} \sum_i a_{N,i}\tau_i^3 \\ &= \sum a_i \tau_i^3 + a_\infty \tau^3,\end{aligned}$$

where $a_i = \lim_N a_{N,i}$ must exist and

$$a_\infty = \lim_{s\to\infty} \lim_{N\to\infty} \sum_{i>s} a_{N,i}$$

is the part of the measure that tends to infinity. In the infinite case, we can also give an explicit formula for a_i. This is done in much the same way that we calculated the probability of a given ball when proving the limit of τ_N exists. Indeed, to find the probability $\widetilde{\tau^2}(T)$ for some type III triangulation T with $|T| = n$ we just need to find the probability $\tau^2(T)$ when each edge is replaced by an external face of size 2. By Proposition 4.10 this is:

$$(3n-6)\alpha^{3-n}(Z_0^2)^{3n-7}C_0^2/C_1^2.$$

(A sphere triangulation with n vertices has $3n - 6$ edges.) Substituting $Z_0^2 = 9/8$ and the values of C this translates to:

$$\frac{2^{19}}{3^7}(n-2)\Big(\frac{27}{256}\Big)^n.$$

Since there are $\varphi_{n-3,1}^3$ possible triangulations of size n, the probability $a_n = \tau^2(|\tilde{T}| = n)$ is:

$$a_n = \frac{2^{20}(4n-11)!}{3^7(n-3)!(3n-7)!}\Big(\frac{27}{256}\Big)^n.$$

In order to find a_∞ we need to sum a_n. Since $(256/27)^n a_n$ is a linear combination of $\varphi_{n-3,1}$ and $n\varphi_{n-3,1}$, the generating function $A(t) = \sum a_n (256t/27)^n$ is a linear combination of $t^{-3}Z_1(t)$ and its derivative. Using that we find: $\sum a_n = 1/2$ and the remainder:

$$a_\infty = 1/2.$$

□

Since we know that the UIPT is a.s. one ended, it has at most one infinite 3-connected component. The above calculations tell us more. We see that with probability 1/2 the root triangle is part of the infinite 3-connected component. In fact, if the root is in a finite 3-connected component, then this component has a number of 2-gonal external faces, and the infinite one contains a triangulation with the same LAW as the original UIPT. Iterating this we see that there is always a unique infinite 3-connected component, and with probability 1/2 the root is part of it. This is an infinite version of an asymptotic result on finite triangulations found in [19]. In fact, we know now the distribution of the size of the 3-connected component of the root, as a_n is the asymptotic probability that the component has size n.

How do we get back from the type III UIPT to the type II UIPT? We need to find the distribution of a UIPT conditioned on including an infinite triangulation. Theorem 5.1 deals with the UIPT conditioned on containing a finite sub-triangulation, and by conditioning on a growing subsequence of triangulations, we see that to get back from the infinite 3-connected component to the whole type II triangulation we need to replace each edge of $\tilde{T}$ with a free triangulation of a 2-gon.

Note that the expected number of triangles in a free triangulation of a 2-gon is twice the expected number of internal vertices and so is 2/3 (again, this is the derivative of $Z_0^3(t)$ at α^{-1}). Since a triangulation contains 3/2 times as many edges as triangles, we see that in some sense in the resulting type II triangulation 1/2 the triangles were in the original type III triangulation and 1/2 were added.

As a consequence of this relation, some results on the type II UIPT are valid for type III as well. Those include the results on growth and on percolation derived in [4], among others.

Acknowledgement. We thank Itai Benjamini and Bálint Virág for inspiring conversations. Part of this research was done during visits of the first author to Microsoft Research. The first author thanks his hosts for these visits.

References

1. Ambjørn, J.: Quantization of Geometry. *Lectures presented at the 1994 Les Houches Summer School "Fluctuating Geometries in Statistical Mechanics and Field Theory"*. `arXiv:hep-th/9411179`
2. Ambjørn, J., Durhuus, B., Jonsson, T.: *Quantum Gravity, a Statitstical Field Theory Approach*. Cambridge Monographs on Mathematical Physics, Cambridge: Cambridge University Press, 1997
3. Ambjørn, J., Watabiki, Y.: Scaling in quantum gravity. Nucl. Phys. B **445**(1), 129–142 (1995)
4. Angel, O.: *Growth and Percolation on the Uniform Infinite Planar Triangulation*. In preparation
5. Banderier, C., Flajolet, P., Schaeffer, G., Soria, M.: Planar maps and Airy phenomena. In: *Automata, lang. and prog. (Geneva, 2000)*, Lecture Notes in Comput. Sci., **1853**, Berlin: Springer, 2000, pp. 388–402
6. Banderier, C., Flajolet, P., Schaeffer, G., Soria, M.: Random maps, coalescing saddles, singularity analysis, and Airy phenomena. In: *Analysis of algorithms (Krynica Morska, 2000)*. Random Struc. Alg. **19**(3–4), 194–246 (2001)
7. Bender, E.A., Compton, K.J., Richmond, B.L.: 0-1 laws for maps. Random Struc. Alg. **14**(3), 215–237 (1999)
8. Bender, E.A., Richmond, B.L., Wormald, N.C.: Largest 4-connected components of 3-connected planar triangulations. Random Struc. Alg. **7**(4), 273–285 (1995)
9. Benjamini, I., Schramm, O.: Recurrence of Distributional Limits of Finite Planar Graphs. Elec. J. Prob. **6**, 23 (2001)
10. Boulatov, D.V., Kazakov, V.A.: The Ising Model on a Random Planar Lattice: the Structure of the Phase Transition and the Exact Critical Exponents. Phys. Lett. B **186**(3–4), 379–384 (1987)
11. Bowick, M.J., Catterall, S.M., Thorleifsson, G.: Minimal Dynamical Triangulations of Random Surfaces. Phys. Lett. B **391**(3–4), 305–309 (1997)

12. Brown, W.G.: Enumeration of triangulations of the disk. Proc. London Math. Soc. (3) **14**, 746–768 (1964)
13. Chassaing, P., Schaeffer, G.: *Random Planar Lattices and Integrated Super-Brownian Excursion*. arXiv:math.CO/0205226
14. David, F.: Simplicial Quantum Gravity and Random Lattices. In: *Gravitation et quantifications (Les Houches, 1992)*, Amsterdam: North-Holland, Amsterdam, 1995, pp. 679–749, arXiv:hep-th/9303127
15. Di Francesco, P., Ginsparg, P., Zinn-Justin, J.: 2D Gravity and Random Matrices. Phys. Rep. **254**(1–2), (1995)
16. Duplantier, B.: Random walks, polymers, percolation, and quantum gravity in two dimensions. In: STATPHYS 20 (Paris, 1998). Phys. A **263**(1–4), 452–465 (1999)
17. Duplantier, B.: Random walks and quantum gravity in two dimensions. Phys. Rev. Lett. **81**(25), 5489–5492 (1998)
18. Gao, Z., Richmond, B.L.: Root vertex valency distributions of rooted maps and rooted triangulations. Europ. J. Comb. **15**(5), 483–490 (1994)
19. Gao, Z., Wormald, N.C.: The size of the largest component in random planar maps. Siam J. Disc. Math. **12**(2), 217–228 (1999)
20. Goulden, I.P., Jackson, D.M.: *Combinatorial enumeration*. New York: Wiley, 1983
21. He, Z.X., Schramm, O.: Hyperbolic and parabolic packings. Disc. Comp. Geom. **14**(2), 123–149 (1995)
22. Knizhnik, V.G., Polyakov, A.M., Zamolodchikov, A.B.: Fractal structure of 2D-quantum gravity. Mod. Phys. Lett. A **3**(8), 819–826 (1988)
23. Koebe, P.: Kontaktprobleme der konformen abbildung, Berichte Verhande. Sächs. Akad. Wiss. Leipzig, Math. -Phys. Klasse **88**, 141–164 (1936)
24. Lawler, G., Schramm, O., Werner, W.: Values of Brownian intersection exponents. I. Half-plane exponents. Acta Math. **187**(2), 237–273 (2001)
25. Lawler, G., Schramm, O., Werner, W.: Values of Brownian intersection exponents. II. Plane exponents. Acta Math. **187**(2), 275–308 (2001)
26. Lawler, G., Schramm, O., Werner, W.: Values of Brownian intersection exponents III: Two-sided exponents. Ann. Inst. H. Poincare Prob. Stat. **38**(1), 109–123 (2002)
27. Lyons, R., Peres, Y.: *Probability on Trees and Networks*. www.math.gatech.edu/ rdlyons/prbtree/prbtree.html
28. Richmond, B.L., Wormald, N.C.: Random triangulations of the plane. Euro. J. Comb. **9**(1), 61–71 (1988)
29. Richmond, B.L., Wormald, N.C.: Almost all maps are asymmetric. J. Comb. Theory Ser. B **63**(1), 1–7 (1995)
30. Schaeffer, G.: *Conjugaison s'arbres et cartes combinatoires aléatoires*. PhD. Thesis, Université Bordeaux I, 1998, Bordeaux
31. Smirnov, S.: Critical percolation in the plane: conformal invariance, Cardy's formula, scaling limits. C. R. Acad. Sci. Paris Sé4. I Math. **333**(3), 239–244 (2001)
32. Tutte, W.T.: A census of planar triangulations. Canad. J. Math. **14**, 21–38 (1962)
33. Tutte, W.T.: A census of Hamiltonian polygons. Canad. J. Math. **14**, 402–417 (1962)
34. Tutte, W.T.: A census of slicings. Canad. J. Math. **14**, 708–722 (1962)
35. Tutte, W.T.: A census of planar maps. Canad. J. Math. **15**, 249–271 (1963)
36. Tutte, W.T.: On the enumeration of convex polyhedra. J. Comb. The. Ser. B **28**(2), 105–126 (1980)

Communicated by M. Aizenman

ISRAEL JOURNAL OF MATHEMATICS **147** (2005), 221–243

COMPOSITIONS OF RANDOM TRANSPOSITIONS

BY

ODED SCHRAMM

Microsoft Research, One Microsoft Way, Redmond, WA 98052-6399, USA

In loving memory of my parents, Hanna and Mickey Schramm

ABSTRACT

Let $Y = (y_1, y_2, \ldots)$, $y_1 \geq y_2 \geq \cdots$, be the list of sizes of the cycles in the composition of cn transpositions on the set $\{1, 2, \ldots, n\}$. We prove that if $c > 1/2$ is constant and $n \to \infty$, the distribution of $f(c)Y/n$ converges to $PD(1)$, the Poisson–Dirichlet distribution with parameter 1, where the function f is known explicitly. A new proof is presented of the theorem by Diaconis, Mayer-Wolf, Zeitouni and Zerner stating that the $PD(1)$ measure is the unique invariant measure for the uniform coagulation-fragmentation process.

1. Introduction

Consider the composition $\pi_t = T_t \circ T_{t-1} \circ \cdots \circ T_2 \circ T_1$ of random, uniform, independent traspositions T_j of $V := \{1, 2, \ldots, n\}$. How large must t be in order for π_t to "look like" a random-uniform permutation π of V? As we will see, the answer depends on the precise meaning given to the term "look like".

It is easy to check that $\mathbf{P}[\pi(v) = v] = 1/n$ for all $v \in V$. Therefore, the expected number of fixed points of π is 1. However, if v does not appear in any of the transpositions $T_1, T_2, \ldots, T_t$, then $\pi_t(v) = v$. By the familiar solution of the coupon collector's problem, we see that when $t = o(n \log n)$, the probability that π_t has at most one fixed point is small. In this sense, π_t and π are rather different when $t = o(n \log n)$. On the other hand, when $t > cn \log n$, $c > 1/2$, the total variation distance between the law of π_t and that of π tends to zero as $n \to \infty$ [DS81].

We now consider the situation where $t \leq cn$ with $c < 1/2$. Let G^t be the graph on $V = \{1, 2, \ldots, n\}$ where $\{v, u\}$ is an edge in G^t if and only if the

Received May 11, 2004

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_17, © Springer Science+Business Media, LLC 2011

transposition (v, u) appears in $\{T_1, \ldots, T_t\}$. Let V_G^t denote the set of vertices of the largest connected component of G^t (with arbitrary tie breaking if there is more than one). By the Erdős–Rényi Theorem, when $t \leq cn$, $c < 1/2$, we have $|V_G^t| = O(\log n)$ asymptotically almost surely (a.a.s.). It follows that the largest cycle (orbit) of π_t is also of size $O(\log n)$. When $c = 1/2$, the same holds, but with $\log n$ replaced by any function growing faster than $n^{2/3}$. This contrasts with the fact that for every $k \in \{1, 2, \ldots, n\}$ the probability that the cycle of π containing 1 has size $\leq k$ is precisely k/n. Thus, π_t is very different from π when $t/n \leq c \leq 1/2$.

Our main theorem deals with the case where $t/n \geq c > 1/2$. Confirming a conjecture by Aldous, we prove that in this range, the large cycles of π_t, when normalized by their total length, have a distribution that is close to that of the large cycles of π. A more precise statement of this result will be given shortly.

Let σ be some permutation of V. Let $X(\sigma)$ denote the set of cycles (orbits) of elements of V under σ. The **cycle structure** $\mathfrak{X}(\sigma)$ is then the sorted list of the lengths of the cycles, that is, the list $(|C| : C \in X(\sigma))$ sorted in nonincreasing order. Thus, $\mathfrak{X}_i(\sigma)$ denotes the size of the i-th largest cycle of σ. If i is larger than the number of cycles of σ, then we set $\mathfrak{X}_i(\sigma) = 0$, by convention. Since each T_j is chosen uniformly among the transpositions, it follows that for each fixed permutation σ of V the distribution of π_t is the same as that of $\sigma \circ \pi_t \circ \sigma^{-1}$. Thus, the distribution of π_t is determined by the distribution of the conjugacy class of π_t. Now, the conjugacy class of π_t is determined by $\mathfrak{X}(\pi_t)$. Consequently, the distribution of $\mathfrak{X}(\pi_t)$ determines the distribution of π_t.

We are now ready to state our main theorem, which gives a positive answer to a conjecture by David Aldous as stated in [BD].

THEOREM 1.1: *Let $c > 1/2$, and take $t \geq cn$. As $n \to \infty$, the law of $\mathfrak{X}(\pi_t)/|V_G^t|$ converges weakly to the Poisson–Dirichlet distribution $PD(1)$ with parameter 1 (which is defined below).*

A more explicit statement of the theorem is as follows. Given $c > 1/2$ and $\epsilon > 0$, there is an $n(c, \epsilon)$ such that for every $n > n(c, \epsilon)$ and every $t \geq cn$ there is a coupling of the sequence of transpositions T_j and a $PD(1)$ sample Y such that

$$\mathbf{P}[\|Y - \mathfrak{X}(\pi_t)/|V_G^t|\|_\infty < \epsilon] > 1 - \epsilon.$$

Weak convergence has several equivalent formulations (see [Dud89]), and we have opted to use the coupling version here.

The $PD(1)$ distribution is a probability measure on the infinite dimensional

simplex

$$\Omega := \{y \in [0,1]^{\mathbb{N}_+} : \sum_i y_i = 1, y_1 \geq y_2 \geq \cdots\}$$

and may be defined as follows. Let $U_1, U_2, \ldots$ be an i.i.d. sequence of random variables uniformly distributed in $[0,1]$. Set $x_1 := U_1$ and inductively $x_j := U_j(1 - \sum_{i=1}^{j-1} x_i)$. Let (y_i) be the sequence (x_i) sorted in nonincreasing order. The $PD(1)$ distribution is defined as the law of (y_i). See, for example, [Hol01] for other definitions and a discussion of some of the properties of the Poisson–Dirichlet distributions.

The behaviour of the size of the largest cluster of G^t, which is the normalizing quantity $|V_G^t|$ in the theorem, is known precisely. The Erdős–Rényi theorem (see, e.g., [AS00]) tells us that

$$|V_G^t|/n \to z(2t/n) \tag{1.1}$$

in probability as $n \to \infty$, where $z(s)$ is the survival probability of a Galton–Watson branching process with offspring distribution which is the Poisson random variable with mean s. Moreover, $z(s)$ is the positive solution of the equation $1 - z = \exp(-sz)$ when $s > 1$ and $z(s) = 0$ for $s \in [0,1]$.

Berestycki and Durrett [BD] have analysed other aspects of the chain π_t which exhibit a phase transition near $t = n/2$: they investigate the minimal number of transpositions necessary to write π_t as a composition.

In [DMP95], Diaconis, McGrath and Pitman discuss the Riffle shuffle, which is another example where the large cycles appear relaxed well before the permutation is uniformly distributed.

The evolution of $\mathfrak{X}(\pi_t)$ is also known as the discrete uniform coagulation-fragmentation process. Let us briefly describe the transition from $\mathfrak{X}(\pi_t)$ to $\mathfrak{X}(\pi_{t+1})$. Suppose that T_{t+1} is the transposition (a,b). Then a and b are selected uniformly from V, and are "almost independent". (We could also allow $a = b$; then $T = (a,a)$ would be the identity transposition, and a and b would be independent. That would not change anything significant in the following.) Let $X_i, X_j \in X(\pi_t)$ satisfy $a \in X_i$, $b \in X_j$. Then X_i and X_j are size biased selections from $X(\pi_t)$, and are nearly independent given π_t. If $X_i \neq X_j$, then in π_{t+1} the two cycles X_i and X_j are replaced by the single cycle whose vertices are $X_i \cup X_j$. If $X_i = X_j$, then this cycle splits into two cycles of π_{t+1}. If $k = |X_i|$ and $m \in \mathbb{N}_+$ is the least positive integer satisfying $\pi_t^m(a) = b$, then the resulting two cycles of π_{t+1} are $(a, \pi_t(a), \ldots, \pi_t^{m-1}(a))$ and $(b, \pi_t(b), \ldots, \pi_t^{k-m-1}(b))$. Note that given X_i and given $X_i = X_j$, the resulting two new cycles have sizes m and $|X_i| - m$, where m is chosen uniformly in $\{1, 2, \ldots, |X_i| - 1\}$.

There is a similar continuous coagulation-fragmentation process, which is a discrete time Markov chain on the infinite dimensional simplex Ω. The transition kernel M of the chain operates as follows. Given $Y = (Y_1, Y_2, \ldots) \in \Omega$, we choose two indices $i, j \in \mathbb{N}_+$ independently, with $\mathbf{P}[i = k|Y] = \mathbf{P}[j = k|Y] = Y_k$. If $i \neq j$, then let Y' be obtained from Y by replacing the two entries Y_i and Y_j with the single entry $Y_i + Y_j$ and resorting. If $i = j$, then given (Y, i, j), a random variable v is selected uniformly in $[0, Y_i]$ and Y' is obtained by splitting the entry Y_i into the two entries v and $Y_i - v$, and resorting. Then Y' is the new state of the Markov chain.

It is known that the probability measure $PD(1)$ is invariant under M. Apparently, this was first proved in [Wat76]; references for several other proofs of this fact are given in [DMWZZ]. Vershik conjectured that $PD(1)$ is the only invariant measure. Subsequently, this was proved by Diaconis, Mayer-Wolf, Zeitouni and Zerner:

THEOREM 1.2 ([DMWZZ]): *The invariant measure for M is unique.*

See [DMWZZ] for more information and bibliography regarding the history of the problem, including some earlier established special cases.

The proof of [DMWZZ] relies on coupling the discrete and the continuous coagulation-fragmentation processes, and using representation theory on the symmetric group to understand the discrete process. In the present paper, we use a different coupling to handle the continuous process directly, and thereby give a different proof of Theorem 1.2. Moreover, a slight modification of this coupling will be essential in the proof of Theorem 1.1.

The problems addressed in this paper are a mean-field version of a statistical physics model suggested by Tóth [Tót93], which may be described as follows. Consider a locally finite graph $G = (V, E)$, and fix a parameter $\beta > 0$. For each (unoriented) edge $e \in E$, let $Z_e \subset [0, 1]$ be an independent Poisson point process of intensity β on $[0, 1]$. Let $v_0 \in V$. We now describe a walk $v(t)$ starting at $v(0) = v_0$. Let t_1 be the first $t > 0$ such that there is an edge $e_1 = [v_0, v_1]$ incident with v_0 such that $t_1 \in Z_{e_1} + \mathbb{Z}$. If there is no such t_1, then $v(t) = v_0$ for all $t \geq 0$. But if t_1 exists, then let $v(t) = v_0$ for $t \in [0, t_1)$ and $v(t_1) = v_1$. Inductively, assume that t_j and v_j are defined and $v(t_j) = v_j$. Let t_{j+1} be the first $t > t_j$ such that there is an edge $e_{j+1} = [v_j, v_{j+1}]$ incident with v_j such that $t \in Z_{e_{j+1}} + \mathbb{Z}$; set $v(t) = v_j$ for $t \in (t_j, t_{j+1})$ and $v(t_{j+1}) = v_{j+1}$.

In the case where G is the complete graph on V, it is easy to see that the orbit of 1 in π_t is analogous to the range of this walk starting at 1, where $\beta = t/n$. The essential difference between the two is the distinction between continuous

time and discrete time.

There are several known open problems regarding Tóth's model. Is it true that for (connected) bounded degree graphs G, the simple random walk on G is transient iff Tóth's walk v_j visits infinitely many vertices with positive probability for some $\beta > 0$? In particular, is this true for $G = \mathbb{Z}^d$? For finite graphs G, one may ask about the distribution of the size of the image of the walk $\{v_j\}$, for example. See [Ang03] for an analysis of Tóth's model on regular trees and for a list of some open problems, including those mentioned above.

Returning to the symmetric group, one may ask about the typical cycle structure near the transition point $t = n/2$. A very thorough analogous theory exists for the Erdős–Rényi transition. See, for example, [Spe94, As00, JŁR00] and the references cited there.

NOTATIONS. For the convenience of the reader, we list here some of the notations used extensively, with hyperlinks and page numbers of the definitions, and a brief description, where appropriate.

V	$\{1,2,\ldots,n\}$	221
$T_1, T_2, \ldots$	i.i.d. uniform transpositions on V	221
π_t	$T_t \circ T_{t-1} \circ \cdots \circ T_1$	221
$X(\sigma)$	set of cycles of a permutation σ	222
X^s	$X(\pi_s)$	226
$\mathfrak{X}(\sigma)$	cycle structure of σ	222
$X^s(v)$	cycle of π_s containing v	226
$V_X^s(k)$	union of cycles of π_s of size at least k	226
G^t	graph whose edges correspond to transpositions $T_i, i \le t$	222
V_G^t	largest cluster in G^t	222
$V_G^t(k)$	union of clusters of G^t of size at least k	226
$z(s)$	function in the Erdős–Rényi theorem	223
Ω	$\{y \in [0,1]^{\mathbb{N}_+} : \sum_i y_i = 1, y_1 \ge y_2 \ge \cdots\}$	223
$PD(1)$	Poisson–Dirichlet distribution with parameter 1	223
M	coagulation-fragmentation transition kernel	224
$\tilde{M}$	the coupling	230
$I(Y,Z)$	indexes of matched entries	230
Q	sum of matched entries	232
$\tilde{Y}, \tilde{Z}, \hat{Y}, \hat{Z}$	partitions used in defining $\tilde{M}$	230
u, v	random variables used in the definition of $\tilde{M}$	230
$\bar{\epsilon}$	ϵ + fragments smaller than ϵ	233
N^t	unmatched entries larger than ϵ	233

y_1^t, z_1^t largest unmatched entries in Y^t and Z^t 233

2. Big pieces

The main goal of the present section is to show in a quantitative way that most vertices in V_G^t are in reasonably large cycles of π_t.

Suppose that π is a permutation on V and $T = (x, y)$ a transposition. If x and y are in different cycles in π, then in $T \circ \pi$ these two cycles are joined, and the other cycles remain unchanged. Now suppose that $C = (x_0, x_1, \ldots, x_m)$ is a cycle of π which contains x and y. Say, $x = x_j$, $y = x_i$, and $j < i$. Then in $T \circ \pi$ the cycle C is split into the cycles $(x_i, x_{i+1}, \ldots, x_{j-1})$ and $(x_j, x_{j+1}, \ldots, x_m, x_0, x_1, \ldots, x_{i-1})$. The other cycles remain unchanged, of course. This clearly implies the following

LEMMA 2.1: *Let π be a permutation of V and $s \in \mathbb{N}$. Let T be a uniform-random transposition on V. Then the probability that some cycle of π is split in $T \circ \pi$ into two cycles at least one of which has length $\leq s$ is at most $2s/(n-1)$.*

This will be used in the next lemma. Let $X^s = X(\pi_s)$ be the set of cycles of π_s and for $v \in V$ let $X^s(v)$ be the cycle in X^s containing v. Let $V_G^s(k) \subset V$ be the union of those connected components of G^s which have at least k vertices, and let $V_X^s(k) \subset V$ be the union of the cycles in X^s that have at least k vertices.

LEMMA 2.2:

$$\mathbf{E}|V_G^s(k) \setminus V_X^s(k)| \leq 4sk^2/(n-1)$$

holds for every $k, s \in \mathbb{N}$.

Proof: Let I be the set of $t \in \mathbb{N}$ such that there is a cycle $A \in X^{t-1}$ which splits into two nonempty cycles in X^t, $A = A_1 \cup A_2$, $A_1, A_2 \in X^t$ and at least one of these cycles, say A_1, satisfies $|A_1| \leq k$. The above lemma shows that $\mathbf{P}[t \in I] \leq 2k/(n-1)$ for every $t \in \mathbb{N}$, and hence $\mathbf{E}[|I \cap [0, s]|] \leq 2sk/(n-1)$.

Suppose that $C \in X^s$, $|C| < k$ and $C \subset V_G^s(k)$. There must be some vertex $u \in C$ and some time $t \leq s$ such that $|X^t(u)| < |X^{t-1}(u)|$; otherwise, C would be equal to a component of G^s. Among all such possible pairs (u, t), we choose one that maximizes t. Then we have $X^t(u) \subset C$. Consequently, $t \in I \cap [0, s]$ and at least one of the two elements of V transposed by T_t is in C. Therefore, the number of such C is at most $2|I \cap [0, s]|$. The statement of the lemma now follows from the above bound on $\mathbf{E}|I \cap [0, s]|$. ∎

The following lemma will tell us that if X^t has many vertices in reasonably large cycles at time $t = t_0$, then with high probability at a specified later time t_1 most of these vertices will be in cycles of size at least ϵn.

LEMMA 2.3: *Let $\delta \in (0,1]$, $t_0, j \in \mathbb{N}$, and $\epsilon \in (0, 1/8)$. (The lemma will be useful primarily when $(\log n)^2 \le 2^j \le n^\alpha$ with any constant $\alpha < 1/2$.) Assume that $2^j < \epsilon\delta n$ and that $\mathbf{P}[|V_X^{t_0}(2^j)| > \delta n] > 0$. Set $\rho := 2^j/n$ and*

$$t_1 := t_0 + \lceil 2^6 \delta^{-1} \rho^{-1} \log_2(\rho^{-1}) \rceil. \tag{2.1}$$

Then the number of vertices v that are in cycles of size at least 2^j at time t_0 but are not in cycles of size at least $\epsilon\delta n$ at time t_1 satisfies

$$\mathbf{E}[|V_X^{t_0}(2^j) \setminus V_X^{t_1}(\epsilon\delta n)| \,|\, |V_X^{t_0}(2^j)| > \delta n] \le O(1)\delta^{-1}\epsilon |\log(\epsilon\delta)| n, \tag{2.2}$$

where the constant implied in the $O(1)$ notation is universal.

Two important aspects of this lemma are that the right hand side of (2.2) does not depend on j and that t_1 does not depend on ϵ. (However, $t_1 - t_0$ depends primarily on j and the right hand side of (2.2) depends primarily on ϵ.)

Before we begin with the actual proof, here is an informal outline. Let $v \in V_X^{t_0}(2^j)$. Set $K := \lceil \log_2(\epsilon\delta n) \rceil$. We will choose a sequence of times $\tau_j, \tau_{j+1}, \ldots, \tau_K$. For $s = j, j+1, \ldots, K$, when $t \in [\tau_s, \tau_{s+1})$ we will "expect" the size of $X^t(v)$ to be at least 2^s. This can fail in either of two scenarios: it may happen because a transposition cuts the cycle of v, or it may happen because no transposition merges the cycle of v with a sufficiently large cycle. The probabilities for each of these unfortunate situations will be appropriately estimated. The choice of the time interval $\tau_{s+1} - \tau_s$ is somewhat delicate. If it is too long, then perhaps too many cycles will be cut, while if it is too short, then cycles will not have enough time to merge. It turns out that

$$a_s := 2^4 \delta^{-1} 2^{-s} (\lceil \log_2 n \rceil - s)(n-1)$$

is roughly the right choice, as will become clear in the course of the proof.

Proof: Within the proof below, expectations and probabilities will be conditioned on $|V_X^{t_0}(2^j)| > \delta n$. Let $K := \lceil \log_2(\epsilon\delta n) \rceil$, and let a_s be as above. For $s > j$ let $m_s := \lceil a_s \rceil$ and set $m_j := t_1 - t_0 - \sum_{s=j+1}^{K-1} m_s$. Set $\tau_s := t_0 + \sum_{i=j}^{s-1} m_i$. Note that $\tau_K = t_1$ and $a_s \le m_s \le O(a_s)$ for $s = j, j+1, \ldots, K-1$.

Let $s \in \{j, j+1, \ldots, K-1\}$ and $t \in \{\tau_s + 1, \tau_s + 2 \ldots, \tau_{s+1}\}$. Define $F^t \subset V$ to be the set of vertices $v \in V$ such that $|X^t(v)| < |X^{t-1}(v)|$ and $|X^t(v)| < 2^{s+1}$.

Lemma 2.1 shows that $\mathbf{E}|F^t| \le 2^{2s+4}/(n-1)$. We also set $\tilde{F}^t := \bigcup_{\tau=t_0+1}^{t} F^\tau$. Then

$$\mathbf{E}|\tilde{F}^{t_1}| \le \sum_{s=j}^{K-1} m_s 2^{2s+4}/(n-1) = O(\epsilon|\log(\epsilon\delta)|n).$$

We consider the vertices in F^t as vertices "failing" at time t. However, there are other ways in which vertices can fail. If at time $t \in \{\tau_s, \tau_s+1, \dots, \tau_{s+1}-1\}$ we have $|V_X^t(2^s)| < \delta n/2$, then we consider the whole process as failed, and we set $H^t := V$. Otherwise, take $H^t = \emptyset$. Also set $\tilde{H}^t := \bigcup_{t'=t_0}^{t} H^{t'}$.

The third and last way in which a vertex v may fail is if $X^t(v)$ does not grow in time. Let

$$B^s := V_X^{\tau_s}(2^s) \setminus (\tilde{F}^{\tau_{s+1}} \cup \tilde{H}^{\tau_{s+1}-1} \cup V_X^{\tau_{s+1}}(2^{s+1})),$$

and $\tilde{B}^s := \bigcup_{k=j}^{s} B^k$. The vertices in B^s are vertices whose cycles failed to grow sufficiently between time τ_s and time τ_{s+1}. It is clear that

$$V_X^{t_0}(2^j) \subset V_X^{t_1}(\epsilon\delta n) \cup \tilde{H}^{t_1} \cup \tilde{F}^{t_1} \cup \tilde{B}^K. \tag{2.3}$$

If $v \in B^s$, then it must be the case that for every $t \in [\tau_s, \tau_{s+1}-1]$ we have $|V_X^t(2^s)| \ge \delta n/2$ (since B^s is disjoint from $\tilde{H}^{\tau_{s+1}-1}$) and $v \in V_X^t(2^s) \setminus V_X^t(2^{s+1})$ (since B^s is disjoint from $\tilde{F}^{\tau_{s+1}}$). If we condition on $2^s \le |X^t(v)| < 2^{s+1}$ and on $|V_X^t(2^s)| > \delta n/2$, then there is probability at least

$$2^s(\delta n/2 - 2^{s+1})\binom{n}{2}^{-1} \ge 2^{s-3}\delta(n-1)^{-1}$$

that T_{t+1} transposes an element from $X^t(v)$ and an element from some other cycle of X^t whose size is at least 2^s. If that happens, then $v \in V_X^t(2^{s+1})$ and this implies that v cannot be in B^s. Consequently,

$$\begin{aligned}\mathbf{P}[v \in B^s] &\le (1 - 2^{s-3}\delta/(n-1))^{m_s} \\ &\le \exp(-2^{s-3}\delta m_s/(n-1)) \le O(2^s/n).\end{aligned}$$

Hence,

$$\mathbf{E}|\tilde{B}^K| \le O(1)n2^K n^{-1} = O(\epsilon\delta n).$$

It follows from the definition of H^t that in order for H^t to be nonempty, we must have $|\tilde{F}^t \cup \tilde{B}^{s-1}| \ge \delta n/2$. Therefore,

$$\mathbf{E}|\tilde{H}^{t_1}| \le n\mathbf{P}[|\tilde{F}^{t_1} \cup \tilde{B}^K| \ge \delta n/2] \le 2\delta^{-1}\mathbf{E}|\tilde{F}^{t_1} \cup \tilde{B}^K|.$$

When we combine this with (2.3) and the above estimates for $\mathbf{E}|\tilde{F}^{t_1}|$ and $\mathbf{E}|\tilde{B}^K|$, the lemma follows. ∎

LEMMA 2.4: *Fix some $c > 1/2$, and let $t \geq cn$, $t \in \mathbb{N}$. Let $\epsilon, \alpha \in (0, 1/8)$ and let N be the minimal number of cycles in X^t which cover at least $(1-\epsilon)|V_G^t|$ vertices of V_G^t. Then*

$$\mathbf{P}[N > \alpha^{-1} |\log(\alpha\epsilon)|^2] \leq C_1 \alpha$$

for all $n > n_1$, where C_1 is a constant which depends only on c, and n_1 may depend on c and ϵ.

Proof: First, suppose that $t \leq n^{5/4}$. Choose j such that $n^{1/4} \leq 2^j < 2n^{1/4}$. Let $\delta = z/2$, where $z = z(2t/n)$ is the Galton–Watson survival probability discussed in the introduction. Choose t_0 so that (2.1) holds with t in place of t_1. Note that $t - t_0 = O(n^{3/4} \log n)$. (Here and below, the constants in the $O(\cdot)$ notation may depend on c.) We apply the Erdős–Rényi theorem at time t_0 to conclude that a.a.s. $|V_G^{t_0}| - nz = o(n)$ and the second largest component of G^{t_0} has size less than $(\log n)^2$. Lemma 2.2 with $k = 2^j$ and $s = t_0$ implies that $|V_G^{t_0} \setminus V_X^{t_0}(2^j)| \leq n^{7/8}$ a.a.s. Note also that $|V_G^t \setminus V_G^{t_0}| \leq (t - t_0) O(\log n)^2$ a.a.s., because we know that the components of G^{t_0} other than the largest one are typically smaller than $(\log n)^2$. Hence $|V_G^t \setminus V_X^{t_0}(2^j)| < n^{7/8}$ a.a.s. Now, Lemma 2.3 implies that for every fixed $\epsilon' > 0$ and for every sufficiently large n

$$\mathbf{E}[|V_G^t \setminus V_X^t(\epsilon' n)|] < O(1)\epsilon' |\log \epsilon'| n. \tag{2.4}$$

(Note that $|V_G^{t_0}| \leq n$, and hence the conditioning in (2.2) may be ignored once n is large enough so that $\mathbf{P}[|V_X^{t_0}(2^j)| \leq \delta n] < \epsilon' |\log \epsilon'|$.)

Now, to show that (2.4) holds also without the assumption that $t \leq n^{5/4}$, we note that Lemma 2.3 may be applied with $j = 0$, $\delta = 1$ and t_0 chosen so that (2.1) holds with t in place of t_1. (In this case, we do not need to use Lemma 2.2.)

Set $a(k) := |V_G^t \setminus V_X^t(k)|$. Let i_0 be the smallest integer i such that $a(2^{-i}n) < \epsilon n/2$. Then N is bounded by the number of cycles in $V_X^t(2^{-i_0}n) \cap V_G^t$. Let i_1 be the least integer such that $2^{-i_1} < \alpha\epsilon / |\log(\alpha\epsilon)|$. Then (2.4) shows that $\mathbf{P}[i_0 > i_1] = O(\alpha)$. We may write

$$a(k) = \sum \{|A| : A \subset V_G^t, A \in X^t, |A| < k\}.$$

By considering the contribution of each cycle to the sum

$$S_m := \sum_{i=0}^{m} a(2^{-i}n) 2^i / n,$$

we find that $N = O(S_{i_0})$. On the other hand, (2.4) implies that

$$\mathbf{E}[S_{i_1}] \leq O(1) \sum_{i=0}^{i_1} i \leq O(1)(i_1)^2 \leq O(1)|\log(\alpha\epsilon)|^2.$$

Because $\mathbf{P}[i_0 > i_1] = O(\alpha)$, this completes the proof. ∎

3. Coupling

At this point, it seems likely that the proof of Theorem 1.1 can be completed using some of the results from the work of Diaconis, Mayer-Wolf, Zeitouni and Zerner [DMWZZ]. However, we prefer instead to use a different coupling argument to finish off the proof and also prove the main result of [DMWZZ].

We now describe a coupling in the continuous setting. A similar coupling will also apply to couple between the discrete and continuous setting, but the purely continuous setting avoids several annoying minor notational issues.

The coupling is between two Markov chains Y^t and Z^t starting at possibly different initial starting points $Y^0, Z^0 \in \Omega$ with each separately evolving according to the transition kernel M.

In this coupling, the evolution of (Y^t, Z^t) will also be Markov. Its transition kernel will be denoted by $\tilde{M}$.

The basic idea in the construction of $\tilde{M}$ is that if we have entries in Y^t that are equal to entries in Z^t, then we don't want to ruin this. Consequently, if we make a change to such an entry in Y^t, we want to make a corresponding change to the corresponding entry in Z^t. On the other hand, as much as we can, we do want to produce new entries in Y^t and Z^t that match. Our measure of the discrepancy between Y^t and Z^t will roughly be the number of large unmatched entries, and we will strive to reduce the discrepancy.

In order to define $\tilde{M}$, we need some more notations. Let $(Y, Z) \in \Omega^2$. We will need to match entries in Y with entries in Z of the same length, if such exist, and match as many entries as possible. The matching will be encoded via maps $f_{Z,Y}, f_{Y,Z}\colon \mathbb{N}_+ \to \mathbb{N}$, which are defined as follows. Let $i \in \mathbb{N}_+$, let H be the set of $j \in \mathbb{N}_+$ such that $Y_i = Z_j$, and let $k := |\{j \in \mathbb{N} : j \leq i, Y_j = Y_i\}|$. (Partly because we want to easily generalize to the discrete setting, we do not want to rule out the possibility that $Y_i = Y_j$ for some $i \neq j$.) If $|H| < k$, then set $f_{Y,Z}(i) = 0$. Otherwise, let $f_{Y,Z}(i)$ be the k'th smallest element in H. (By exchanging Y and Z, this also defines the map $f_{Z,Y}$.) Let

$$I(Y, Z) := f_{Y,Z}^{-1}(\mathbb{N}_+) = \{i \in \mathbb{N} : f_{Y,Z}(i) \neq 0\}.$$

The entries Y_i with $i \in I(Y,Z)$ will be referred to as matched. Likewise, Z_j, $j \in I(Z,Y)$, are the matched entries of Z. Observe that $f_{Z,Y} \circ f_{Y,Z}(i) = i$ for every $i \in I(Y,Z)$, $f_{Y,Z}(I(Y,Z)) = I(Z,Y)$ and $Z_{f_{Y,Z}(i)} = Y_i$ for every $i \in I(Y,Z)$. Let

$$Q = Q(Y,Z) := \sum\{Y_i : i \in I(Y,Z)\} = \sum\{Z_j : j \in I(Z,Y)\}.$$

We will now describe the transition kernel $\tilde{M}$. Given $(Y,Z) \in \Omega$, we need to perform one step of M for each of Y and Z, thereby generating new configurations Y' and Z'. We associate with Y and with Z partitions $\tilde{Y} = (\tilde{Y}_i : i \in \mathbb{N}_+)$ and $\tilde{Z} = (\tilde{Z}_i : i \in \mathbb{N}_+)$ of $[0,1]$ into closed intervals, as follows. (See also Figure 3.1.) The length of the interval $\tilde{Y}_i$ is Y_i. The intervals $\tilde{Y}_i$ with $i \in I(Y,Z)$ tile the interval $[1-Q,1]$, while the intervals with $i \notin I(Y,Z)$ tile the interval $[0,1-Q]$. Within each of these classes, let the intervals be ordered according to the indices; that is $\max \tilde{Y}_i \le \min \tilde{Y}_{i'}$ if $i < i'$ when $i,i' \in I(Y,Z)$ and when $i,i' \notin I(Y,Z)$. A partition $\tilde{Z} = (\tilde{Z}_j : j \in \mathbb{N}_+)$ is constructed in the same way. Note that necessarily $\tilde{Z}_{f_{Y,Z}(i)} = \tilde{Y}_i$ whenever $i \in I(Y,Z)$.

Let u and v be two independent uniform random variables in $[0,1]$. Let $a, a' \in \mathbb{N}_+$ be the indices satisfying $u \in \tilde{Y}_a$ and $u \in \tilde{Z}_{a'}$. In this way, u induces a size biased sample from Y and from Z. We will use v to induce a different size biased sample, based on different tilings of $[0,1]$. Let $\hat{Y}$ be the tiling $(\hat{Y}_i : i \in \mathbb{N}_+)$ of $[0,1]$ by intervals that is obtained from $\tilde{Y}$ by shifting the interval $\tilde{Y}_a$ to the beginning. (That is, $\hat{Y}_a = [0, Y_a]$, $\hat{Y}_i = Y_a + \tilde{Y}_i$ if $\max \tilde{Y}_i \le \min \tilde{Y}_a$ and $\hat{Y}_i = \tilde{Y}_i$ if $\max \tilde{Y}_a \le \min \tilde{Y}_i$.) Similarly, $\hat{Z}$ is the tiling obtained from $\tilde{Z}$ by shifting $\tilde{Z}_{a'}$ to the beginning. Note that $\hat{Z}_{f_{Y,Z}(i)} = \hat{Y}_i$ whenever $i \in I(Y,Z)$.

Let b and b' be the indices satisfying $v \in \hat{Y}_b$ and $v \in \hat{Z}_{b'}$. If $a \neq b$, let Y' be obtained from Y by replacing the two entries Y_a and Y_b by the single entry $Y_a + Y_b$ and resorting. If $a = b$, let Y' be obtained from Y by replacing Y_a with the two entries v and $Y_a - v$ and resorting. Similarly, if $a' \neq b'$, let Z' be obtained from Z by replacing the two entries $Z_{a'}$ and $Z_{b'}$ by the single entry $Z_{a'} + Z_{b'}$ and resorting. If $a' = b'$, let Z' be obtained from Z by replacing $Z_{a'}$ with the two entries v and $Z_{a'} - v$ and resorting. This completes the construction of the Markov transition kernel $\tilde{M}$.

Let us observe a few essential features of this coupling. If Y_i and Z_j are split, then one of the two new entries in each of Y' and Z' is equal to v. If $i \in I(Y,Z)$ and Y_i is split or merged, then the same happens to $Z_{f_{Y,Z}(i)}$. Similarly, if

$j \in I(Z,Y)$ and Z_j is split or merged, then the same happens to $Y_{f_{Z,Y}(j)}$.

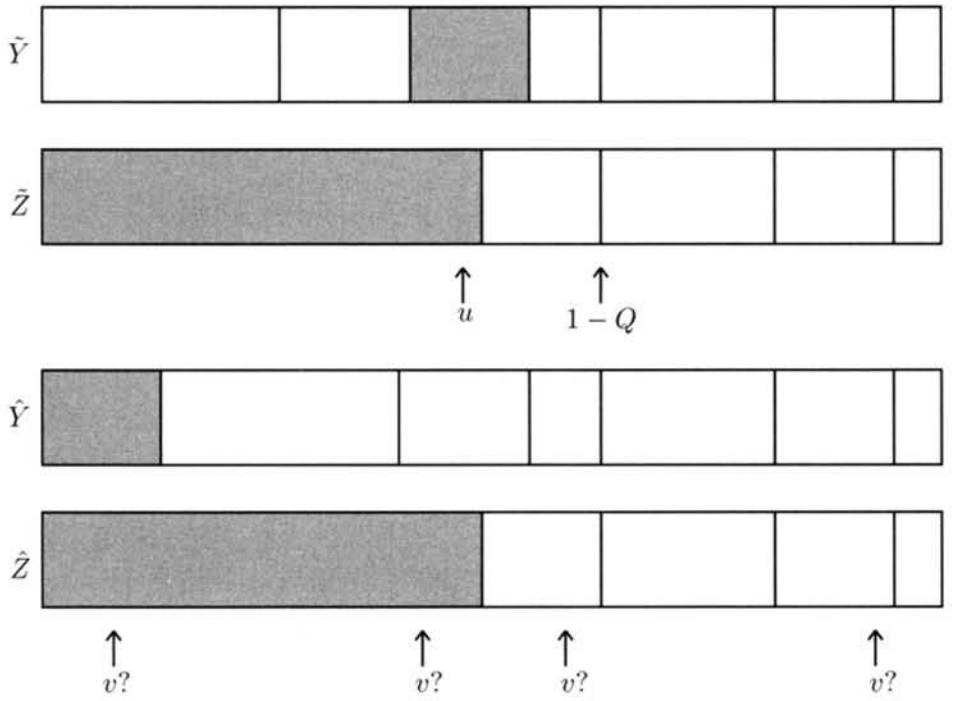


Figure 3.1. The random variable u chooses a segment in $\tilde{Y}$ and a segment in $\tilde{Z}$. The different illustrated choices for the random variable v yield a split in Y and in Z, a merge in Y and a split in Z, a merge in both where a matched segment is not involved, and merges involving matched segments, respectively.

We first informally describe the general behaviour of $\tilde{M}$, postponing the exact statements and proofs. When there are several unmatched reasonably large entries in Y^t and in Z^t, these merge and become few quite quickly. However, when they are very few, it is hard for them to dissappear completely. Suppose that there is one large unmatched entry in Y^t and two unmatched entries in Z^t. When the two unmatched entries in Z^t are merged, the single unmatched entry in Y^t is likely to be split. Thus, the situation does not improve so quickly. There is a parity phenomenon here: if the number of positive entries in Y^t is finite, then its parity either stays the same as that of t, or is opposite to that of t. Even if the number of positive entries is infinite, if it takes a long time for the smaller entries to be hit, the larger entries appear to follow this parity periodicity. One way to handle the parity issue would be to introduce

a delay to either Y^t or Z^t, but not both, in order to match up their parities. However, another phenomenon will be used instead. An unmatched entry in Y^t often splits into one matched entry and one unmatched entry. With any luck, the unmatched entry might be rather small. Thus, large unmatched entries are replaced by small unmatched entries. In effect, there is a diffusion of unmatched entries between different scales. Because of this, it is eventually unlikely to find a large unmatched entry, which is what we want to prove. However, this latter process is much slower than the first stage where large unmatched entries merge and become fewer. Thus, in time t the largest unmatched entry one can expect to find is of order roughly $1/\log t$.

Define

$$N_\epsilon(Y,Z) := |\{i \in \mathbb{N}_+ \setminus I(Y,Z) : Y_i > \epsilon\}|.$$

This is the number of entries in Y that are not matched by entries in Z and have size larger than ϵ.

LEMMA 3.1: *Let $\epsilon > 0$, and let $Y^0, Z^0 \in \Omega$. Let (Y^t, Z^t) be the Markov chain given by $\tilde{M}$ starting at (Y^0, Z^0). To abbreviate notations, set $N^t := N_\epsilon(Y^t, Z^t) + N_\epsilon(Z^t, Y^t)$, $Q^t = Q(Y^t, Z^t) = Q(Z^t, Y^t)$, $I^t := I(Y^t, Z^t)$ and $J^t := I(Z^t, Y^t)$. Also define*

$$\bar{\epsilon} := \epsilon + \sum\{Y_i^0 : Y_i^0 < \epsilon\} + \sum\{Z_i^0 : Z_i^0 < \epsilon\}.$$

Let $y_1^t := \max\{Y_i^t : i \notin I^t\}$ be the size of the largest unmatched entry of Y^t (set $y_1^t = 0$ if all entries are matched), and let z_1^t be the size of the largest unmatched entry of Z^t. Let q be a random variable with values in $\mathbb{N}$ which is independent from the evolution of the chain (Y^t, Z^t). Set

$$\eta := \max\{\mathbf{P}[q = t] : t \in \mathbb{N}\}.$$

Then

$$\mathbf{E}[(1 - Q^q)(1 - Q^q - \max\{y_1^q, z_1^q\})] \leq \frac{\eta}{2} N^0 + 4\bar{\epsilon}\mathbf{E}[q+1]. \tag{3.1}$$

When the right hand side in (3.1) is small, we know that with high probability either the sum of the unmatched entries in Y^q is only slightly larger than the largest unmatched entry, or this is true for Z^q.

Proof: Let $\mathcal{A}_s$ be the event that up to time s in every merging occurring both merged pieces are of size at least ϵ and in every splitting both resulting pieces are of size at least ϵ. Let $\mathcal{F}_s$ be the σ-field generated by $((Y^t, Z^t) : t = 0, 1, \ldots, s)$.

Conditioned on $\mathcal{F}_{t-1}$, the probability that at time t there is a split in any Y_i^{t-1} and one of the pieces is of size less than ϵ is at most 2ϵ. Conditioned on $\mathcal{F}_{t-1}$, the probability that there is any Y_i^{t-1} with $Y_i^{t-1} < \epsilon$ that is merged at time t with some other Y_j^{t-1} is at most $2\sum\{Y_i^{t-1} : Y_i^{t-1} < \epsilon\}$. Similar considerations apply to Z^t. Consequently, for $t \in \mathbb{N}_+$,

$$\mathbf{P}[\neg \mathcal{A}_t | \mathcal{A}_{t-1}, \mathcal{F}_{t-1}] \leq 4\bar{\epsilon}. \tag{3.2}$$

We now study the evolution of the quantity N^t, and consider several different cases for the transition from (Y^t, Z^t) to (Y^{t+1}, Z^{t+1}). In each case we assume that $\mathcal{A}_{t+1}$ holds.

1. The transition involves splitting in Y^t and merging in Z^t. Suppose that Y_i^t is split and Z_j^t is merged with $Z_{j'}^t$. Then necessarily $i \notin I^t$ and $j, j' \notin J^t$. Since $\mathcal{A}_{t+1}$ is assumed to hold, it follows that $N_\epsilon(Y^{t+1}, Z^{t+1}) \leq N_\epsilon(Y^t, Z^t) + 1$ and $N_\epsilon(Z^{t+1}, Y^{t+1}) \leq N_\epsilon(Z^t, Y^t) - 1$. Thus, in this case, $N^{t+1} \leq N^t$.
2. The transition involves splitting in Z^t and merging in Y^t. By symmetry, also in this case we have $N^{t+1} \leq N^t$.
3. The transition involves splitting in Y^t and splitting in Z^t. Note that by construction the size of one of the newly created split entries is the same for Y as for Z. Suppose that Y_i^t and Z_j^t are split. If $i \in I^t$ then also $j \in J^t$ and $Y_i^t = Z_j^t$. In that case, both new entries for Y are the same as the new entries for Z, and hence $N^{t+1} = N^t$. The same conclusion is obtained if $j \in J^t$. If $i \notin I^t$ and $j \notin J^t$, then in both Y^t and Z^t an unmatched entry is replaced by two entries at least one of which is matched. Thus $N^{t+1} \leq N^t$.
4. The transition involves merging in Z^t and merging in Y^t. Suppose that Y_i^t is merged with $Y_{i'}^t$. It is easy to verify, as above, that in this case also $N^{t+1} \leq N^t$. However, if $i, i' \notin I^t$, then the corresponding statement is also true for the merged entries in Z^t, and we actually have $N^{t+1} \leq N^t - 2$.

In summary, we see that on the event $\mathcal{A}_{t+1}$ we have $N^{t+1} \leq N^t$ and $N^{t+1} \leq N^t - 2$ when there is merging in both Y^t and Z^t and the merging does not involve matched entries.

Since $N^t \geq 0$, we obviously have

$$\sum_{t=0}^{\infty} (N^t - N^{t+1}) 1_{\mathcal{A}_{t+1}} \leq N^0,$$

and we have seen that all the summands are nonnegative. Since q is independent

from $(N^t - N^{t+1})1_{\mathcal{A}_{t+1}}$,

$$\begin{aligned}\mathbf{E}[(N^q - N^{q+1})1_{\mathcal{A}_{q+1}}] &= \sum_t \mathbf{E}[(N^t - N^{t+1})1_{\mathcal{A}_{t+1}}1_{q=t}] \\ &= \sum_t \mathbf{E}[(N^t - N^{t+1})1_{\mathcal{A}_{t+1}}]\mathbf{P}[q=t] \le \eta N^0.\end{aligned} \tag{3.3}$$

Set $a^t = 1 - Q^t - \max\{y_1^t, z_1^t\}$. Recall the random variables u and v used in the transition kernel $\tilde{M}$. If in the transition from (Y^t, Z^t) to (Y^{t+1}, Z^{t+1}) we have $u < 1 - Q^t$ and $\max\{y_1^t, z_1^t\} < v < 1 - Q^t$, then in both Y and Z we have merging of unmatched entries. Thus,

$$\mathbf{P}[N^t - N^{t+1} \ge 2 \text{ or } \neg\mathcal{A}_{t+1}|\mathcal{F}_t] \ge (1 - Q^t)a^t.$$

By applying this at time $t = q$ and taking expectations, we get

$$\begin{aligned}\mathbf{E}[(1 - Q^q)a^q] &\le \mathbf{P}[N^q - N^{q+1} \ge 2 \text{ or } \neg\mathcal{A}_{q+1}] \\ &\le \frac{1}{2}\mathbf{E}[(N^q - N^{q+1})1_{\mathcal{A}_{q+1}}] + \mathbf{P}[\neg\mathcal{A}_{q+1}].\end{aligned}$$

Consequently, (3.2) and (3.3) complete the proof of the lemma. ∎

Assuming that we can make the right hand side of (3.1) small, Lemma 3.1 tells us that with high probability either $1 - Q^q - y_1^q$ or $1 - Q^q - z_1^q$ is small. If we knew that both are small, it would follow that also $y_1^q - z_1^q$ is rather small, since $\sum_i Y_i^q = \sum_j Z_j^q = 1$. However, it might be the case that $1 - Q^q - z_1^q$ is small but $1 - Q^q - y_1^q$ is not. The next lemma tells us that in such a situation, with high probability, Y^q does not have more than two significant unmatched entries.

LEMMA 3.2: *With the setting and notations of Lemma 3.1, let y_2^t be the second largest unmatched entry in Y^t. For every $\rho \in (0,1)$*

$$\mathbf{P}[1 - Q^q - y_1^q - y_2^q > \rho] < 2^6\rho^{-4}\eta N^0 + 2^9\bar{\epsilon}\rho^{-4}\mathbf{E}[q+2]. \tag{3.4}$$

Proof: Let $\mathcal{D}$ be the event $\{1 - Q^q - y_1^q - y_2^q > \rho\}$ and let $\mathcal{R}$ be the event $\{1 - Q^q - z_1^q < \rho/4\}$. Assume that $\mathcal{D} \cap \mathcal{R}$ holds. Then $z_1^q \ge 3\rho/4 + y_1^q + y_2^q$. Let $\mathcal{U}$ be the event that the random variables u and v used in the transition from (Y^q, Z^q) to (Y^{q+1}, Z^{q+1}) satisfy $u < 3\rho/4$ and $z_1^q - \rho/2 < v < z_1^q - \rho/4$. On $\mathcal{D} \cap \mathcal{R} \cap \mathcal{U}$, the largest unmatched entry in Z^q will be split and the transition from Y^q to Y^{q+1} would involve a merge (of unmatched entries), because $z_1^q - \rho/2 > y_1$. Consequently, a.s. on $\mathcal{D} \cap \mathcal{R} \cap \mathcal{U}$ the two new entries of Z^{q+1} will be unmatched in Y^{q+1}, and in particular, $1 - Q^{q+1} \ge z_1^q$. Moreover, each of the new entries

of Z^{q+1} would be larger than $\rho/4$. Clearly, $y_1^{q+1} \le y_1^q + y_2^q$. Consequently, $1 - Q^{q+1} - y_1^{q+1} \ge z_1^q - y_1^q - y_2^q \ge 3\rho/4$ and $1 - Q^{q+1} - z_1^{q+1} \ge \rho/4$. Thus, on $\mathcal{D} \cap \mathcal{R} \cap \mathcal{U}$, we have $1 - Q^{q+1} - \max\{y_1^{q+1}, z_1^{q+1}\} \ge \rho/4$. Now,

$$\begin{aligned}
&\mathbf{E}[(1 - Q^{q+1})(1 - Q^{q+1} - \max\{y_1^{q+1}, z_1^{q+1}\})] \\
&\quad \ge \mathbf{E}[(1 - Q^{q+1} - \max\{y_1^{q+1}, z_1^{q+1}\})^2 | \mathcal{D}, \mathcal{R}, \mathcal{U}] \mathbf{P}[\mathcal{D}, \mathcal{R}, \mathcal{U}] \\
&\quad \ge (\rho^2/16) \mathbf{P}[\mathcal{D}, \mathcal{R}, \mathcal{U}].
\end{aligned}$$

Lemma 3.1 with q replaced by $q+1$ therefore gives

$$\mathbf{P}[\mathcal{D}, \mathcal{R}, \mathcal{U}] \le 8\rho^{-2}\eta N^0 + 2^6 \bar{\epsilon} \rho^{-2} \mathbf{E}[q+2].$$

Clearly, $\mathbf{P}[\mathcal{U} | \mathcal{D}, \mathcal{R}] = 3\rho^2/16$, and hence

$$\mathbf{P}[\mathcal{D}, \mathcal{R}] \le (16/3)\rho^{-2} \mathbf{P}[\mathcal{D}, \mathcal{R}, \mathcal{U}].$$

On the other hand, on $\mathcal{D} \setminus \mathcal{R}$ we have $(1 - Q^q)(1 - Q^q - \max\{y_1^q, z_1^q\}) > \rho^2/4$. Thus, applying Lemma 3.1 again gives

$$\mathbf{P}[\mathcal{D} \setminus \mathcal{R}] \le 2\rho^{-2}\eta N^0 + 16\bar{\epsilon}\rho^{-2} \mathbf{E}[q+1].$$

Since $\mathbf{P}[\mathcal{D}] = \mathbf{P}[\mathcal{D}, \mathcal{R}] + \mathbf{P}[\mathcal{D} \setminus \mathcal{R}]$, the above estimates combine to give (3.4), and complete the proof. ∎

LEMMA 3.3: *With the setting and notations of Lemma 3.1, let $\rho \in (0, 1/8)$ and assume that $0 < \epsilon < \rho$. Then for each $t \in \mathbb{N}_+$ and for every $n \in \mathbb{N}_+$ satisfying $2^n \le t\rho$*

$$t^{-1} \sum_{\tau=0}^{t-1} \mathbf{P}[y_1^\tau \ge \rho] \le O(\rho^{-1} n^{-1}) + O(2^{4n}/\rho^5)(N^0/t + \bar{\epsilon}t). \tag{3.5}$$

The basic idea of the proof of the lemma is to use the fact that conditioned on $y_1^\tau \ge \rho$ there is a significant enough probability that at a later time σ there will be some unmatched $Y_{i'}^\sigma \in [2^{-k}\rho, 2^{-k+1}\rho]$, since the unmatched piece at time τ of size $\ge \rho$ may be split immediately. Lemma 3.2 is then used to show that when we fix σ, with high probability the latter event occurs for at most three different k in the range $\{1, \dots, n\}$, if n is not too large. An appropriate summation over k and σ completes the proof.

Proof: For $\sigma > \tau$, $\sigma, \tau \in \mathbb{N}$, $k \in \mathbb{N}_+$, let $\mathcal{X}(\tau, \sigma, k)$ be the event that the transition between time τ and $\tau + 1$ produces a splitting in Y^τ and one of the

split pieces is unmatched, has size in the range $[2^{-k-1}\rho, 2^{-k}\rho)$, and this split piece is not modified up to time σ. Suppose that $y_1^\tau \geq \rho$ and that $Y_i^\tau = y_1^\tau$. If in the transition from τ to $\tau+1$ we have $u \in \tilde{Y}_i^\tau$ and $v \in (Y_i^\tau - 2^{-k}\rho, Y_i^\tau - 2^{-k-1}\rho)$, then Y_i^τ is indeed split, and it is easy to see that the resulting piece $Y_i^\tau - v$ is unmatched a.s. If that happens, the conditioned probability that up to time σ this piece is modified is bounded by $2(\sigma-\tau)2^{-k}\rho$, since the size of this piece is at most $2^{-k}\rho$. Thus,

$$\mathbf{P}[\mathcal{X}(\tau,\sigma,k)|y_1^\tau \geq \rho] \geq 2^{-k-1}\rho^2(1-2(\sigma-\tau)2^{-k}\rho),$$

which implies

$$\mathbf{P}[y_1^\tau \geq \rho] \leq 2^{k+2}\rho^{-2}\mathbf{P}[\mathcal{X}(\tau,\sigma,k)], \quad \text{if } \tau < \sigma \leq \tau + 2^{k-2}/\rho. \tag{3.6}$$

Let $\mathcal{V}^\sigma(n)$ be the event that there are at least 3 distinct $k \in \{0,1,\ldots,n-1\}$ such that there is an unmatched Y_i^σ in the range $[2^{-k-1}\rho, 2^{-k}\rho)$. We now apply Lemma 3.2 with q chosen uniformly in $\{0,1,\ldots,2t-1\}$ and with ρ replaced by $2^{-n}\rho$ to get

$$\sum_{\sigma=0}^{2t-1} \mathbf{P}[\mathcal{V}^\sigma(n)] \leq 2^{4n+11}\rho^{-4}(N^0 + \bar{\epsilon}(t+2)^2). \tag{3.7}$$

Now, observe that

$$\sum_{\tau=0}^{\sigma-1}\sum_{k=0}^{n-1} 1_{\mathcal{X}(\tau,\sigma,k)} < 3 + 1_{\mathcal{V}^\sigma(n)}n.$$

Therefore, by taking expectations and applying (3.7) we get

$$\sum_{\sigma=0}^{2t-1}\sum_{\tau=0}^{\sigma-1}\sum_{k=0}^{n-1}\mathbf{P}[\mathcal{X}(\tau,\sigma,k)] \leq 6t + n\sum_{\sigma=0}^{2t-1}\mathbf{P}[\mathcal{V}^\sigma(n)] \leq O(t) + O(1)2^{4n}n\rho^{-4}(N^0+\bar{\epsilon}t^2).$$

We now assume that $2^n \leq t\rho$. Then the inequalities (3.6) may be applied to the above, giving

$$\sum_{k=0}^{n-1}\sum_{\tau=0}^{t-1}\sum_{\sigma=\tau+1}^{\tau+\lfloor 2^{k-2}/\rho\rfloor} 2^{-k-2}\rho^2\mathbf{P}[y_1^\tau \geq \rho] \leq O(t) + O(1)2^{4n}n\rho^{-4}(N^0+\bar{\epsilon}t^2).$$

This implies (3.5), and completes the proof. ∎

COROLLARY 3.4: *Let $\gamma \in (0,1/2)$. Let q be a random variable with values in $\mathbb{N}$ which is independent from the Markov chain (Y^t, Z^t). Set $\eta := \max\{\mathbf{P}[q = t] : t \in \mathbb{N}\}$, and suppose that $(\bar{\epsilon})^{1-\gamma} \leq \eta \leq (\bar{\epsilon})^{\gamma}/\max\{N^0, 1\}$. Then for all $\lambda \geq 1$ and $\rho > 0$*

$$\mathbf{P}[y_1^q \geq \rho] \leq \mathbf{P}[q > \lambda\eta^{-1}] + C(\lambda/\rho)|\log\bar{\epsilon}|^{-1},$$

where C is a constant depending only on γ.

Proof: Let $s := \lfloor \lambda\eta^{-1} \rfloor$. We have

$$\begin{aligned}\mathbf{P}[y_1^q \geq \rho] &\leq \mathbf{P}[q > \lambda\eta^{-1}] + \sum_{\tau=0}^{s} \mathbf{P}[q = t]\mathbf{P}[y_1^\tau \geq \rho] \\ &\leq \mathbf{P}[q > \lambda\eta^{-1}] + \eta \sum_{\tau=0}^{s} \mathbf{P}[y_1^\tau \geq \rho].\end{aligned}$$

Thus, the proof is completed by applying (3.5) with $t = s + 1$ and $n := \lfloor |\log\bar{\epsilon}|/C \rfloor$, provided that with sufficiently large C we have

$$2^{4n}\rho^{-5}(N^0/t + \bar{\epsilon}t) \leq C\rho^{-1}|\log\bar{\epsilon}|^{-1} \tag{3.8}$$

and $2^n \leq t\rho$. First, note that we may assume that $\lambda, \rho^{-1} < |\log\bar{\epsilon}|$ and $\bar{\epsilon} < 1/10$. Then $2^n \leq t\rho$ holds by the assumptions on η. It is also easy to verify that with an appropriate choice of C, (3.8) follows from our inequalities for η and assumptions about ρ and λ. ∎

THEOREM 3.5: *Let Y^0 and Z^0 be independent random samples from $PD(1)$, and let (Y^t, Z^t) denote their evolution under $\tilde{M}$. Let $t_0 \in \mathbb{N}_+$, and let $q \in \{0, 1, \ldots, t_0 - 1\}$ be chosen uniformly and independently from the evolution of the chain (Y^t, Z^t). Then for each $\rho > 0$,*

$$\mathbf{P}[\max\{y_1^q, z_1^q\} > \rho] \leq O(1)\rho^{-1}(\log t_0)^{-1}.$$

Proof: Set $\epsilon := (t_0)^{-2}$, and define $\bar{\epsilon}$ as in Lemma 3.1. Recall that a size biased sample from the $PD(1)$ sample Y^0 gives the uniform distribution on $[0,1]$ (this is well-known, but also easy to verify from the definition). Consequently, $\mathbf{E}[\bar{\epsilon}] = 3\epsilon$. Let $\mathcal{A}_1$ be the event that $\bar{\epsilon} \leq \epsilon^{3/4}$. Then $\mathbf{P}[\neg\mathcal{A}_1] \leq 3\epsilon^{1/4}$. Let $\mathcal{A}_2$ be the event that $N^0 \leq \epsilon^{-1/4}$. It is easy to see (e.g., using the description of $PD(1)$ from the introduction) that $\mathbf{P}[\neg\mathcal{A}_2] \leq O(\epsilon)$. (In fact, $N^0/|\log\epsilon|$ is very unlikely to be large.) Define η as in Corollary 3.4. Then $\eta = \epsilon^{1/2}$ and on $\mathcal{A}_1 \cap \mathcal{A}_2$ we have $(\bar{\epsilon})^{1-\gamma} \leq \eta \leq (\bar{\epsilon})^{\gamma}/\max\{N^0, 1\}$ with $\gamma = 1/5$, for example. On the event

$\mathcal{A}_1 \cap \mathcal{A}_2$, apply the corollary with $\lambda = 1$ and the corresponding statement with the roles of Y and Z switched, to get

$$\mathbf{P}[\max\{y_1^q, z_1^q\} > \rho | \mathcal{A}_1 \cap \mathcal{A}_2] \leq O(\rho^{-1})|\log\epsilon|^{-1}.$$

Now our estimates for $\mathbf{P}[\neg\mathcal{A}_1]$ and $\mathbf{P}[\neg\mathcal{A}_2]$ complete the proof. ∎

Proof of Theorem 1.2: The proof is similar to the proof of Theorem 3.5 Let μ be a measure that is invariant under M, and let Y^0 be a sample from μ. Let Z^0 be a sample from $PD(1)$ (which we may take to be independent from Y^0, though this is not important). Let $t_0 \in \mathbb{N}_+$, $t_0 > 5$, and let q be as in Theorem 3.5. As in the proof of that theorem, choose $\epsilon = t_0^{-2}$.

Note that for every $t \in \mathbb{N}$, Y^t is also a sample from μ, because μ is t invariant. The same also holds for Y^q, since q is independent from the chain (Y^t).

We now explain how to get bounds on the distributions of $\bar{\epsilon}$ and N^0 using continuous analogs of Lemmas 2.3 and 2.4. Let $\beta(s, Y) := \sum\{Y_i : Y_i \leq s\}$. Since $\lim_{s\searrow 0}\beta(s, Y^0) = 0$ a.s., we may choose $k = k(\epsilon) > 0$ sufficiently large so that $\mathbf{P}[\beta(2^{-k}, Y^0) > \epsilon] < \epsilon$ and $2^{-k} < \epsilon(1-\epsilon)/8$. Set $\delta = 1 - 2^{-k}$ and $t_1 = \lceil 2^6\delta^{-1}k2^k\rceil$. The proof of Lemma 2.3 applied to the continuous setting gives

$$\mathbf{E}[\beta(\epsilon, Y^{t_1})|\beta(2^{-k}, Y^0) \leq \epsilon] \leq O(1)\epsilon|\log\epsilon|.$$

By our choice of k this implies $\mathbf{E}[\beta(\epsilon, Y^{t_1})] \leq O(\epsilon)|\log\epsilon|$. Since Y^{t_1} and Y^0 have the same distribution, this gives

$$\mathbf{E}[\beta(\epsilon, Y^0)] \leq O(\epsilon)|\log\epsilon|, \tag{3.9}$$

and since $\mathbf{E}[\beta(\epsilon, Z^0)] = \epsilon$, as in the proof of Theorem 3.5, we conclude that $\mathbf{E}[\bar{\epsilon}] \leq O(\epsilon)|\log\epsilon|$.

We now adapt the latter part of the proof of Lemma 2.4. Let $m_0 := \lceil|\log_2\epsilon|\rceil$. On the one hand

$$\sum_{m=0}^{m_0} 2^m\beta(2^{-m}, Y^0) = \sum_{m=0}^{m_0}\sum\{2^mY_i : Y_i \leq 2^{-m}\} \leq 2|\{i \in \mathbb{N}_+ : Y_i^0 \geq \epsilon\}|.$$

On the other hand, (3.9) gives

$$\mathbf{E}\left[\sum_{m=0}^{m_0} 2^m\beta(2^{-m}, Y^0)\right] \leq O(1)\sum_{m=0}^{m_0} m = O(1)|\log\epsilon|^2.$$

Consequently, we have $\mathbf{E}[N^0] = O(1)|\log\epsilon|^2$. Now, the proof of Theorem 3.5 applies, and gives for all $\rho > 0$

$$\mathbf{P}[\max\{y_1^q, z_1^q\} > \rho] \leq O(1)\rho^{-1}(\log t_0)^{-1}.$$

We conclude that for every $\rho > 0$ there is a coupling of Y^0 and Z^0 so that $\mathbf{P}[\max\{y_1^0, z_1^0\} > \rho] < \rho$. This implies that $\mu = PD(1)$. ∎

4. Conclusion

Proof of Theorem 1.1: The proof is similar to the proof of Theorem 3.5. Let $\epsilon > 0$. Let q be uniformly chosen in $(2\mathbb{Z}) \cap [0, \epsilon^{-1/2}]$. Set $z = z(2t/n)$, as in (1.1). Let Z^0 be chosen according to $PD(1)$, and let $Y^\tau = \mathfrak{X}(\pi_{t+\tau})/(nz)$.

We now apply a coupling of Z^τ and Y^τ similar to the coupling $\tilde{M}$ given in Section 3. There are a few minor necessary modifications in the definition of the coupling. First, note that the entries of Y^τ do not sum to 1 but to $1/z$. Thus, the random variables u and v needed in the transition kernel for $\tilde{M}$ should be uniform in $[0, 1/z]$. In $\tilde{Y}^\tau$ and $\hat{Y}^\tau$ we put those segments corresponding to cycles that do not intersect V_G^t in the very end; that is, roughly in the interval $[1, 1/z]$. When u or v turn out to be outside of $[0, 1]$, we make no transition to Z; that is, $Z^{\tau+1} = Z^\tau$, in this case.

Another modification is necessary because the transitions of Y^τ are discrete. Thus, the actual size of the splits occurring in the transitions of Y would be determined with $\lceil nzv \rceil/(nz)$. The definition of the matching between entries in Y^τ and entries in Z^τ needs to be modified as well. When a split is made in both Y^τ and Z^τ, pieces which would have been exactly the same may differ slightly now, because of the discretization in the transition of Y. This difference is of order $1/n$, and may be safely ignored. Although these errors may accumulate over time, when matched pieces are merged and split, the total discrepancy would still be small, since we take n much larger than $\epsilon^{-1/2}$, which bounds q.

Lemma 2.4 gives us good control on N^0 while (2.4) gives a bound on the probability that $\bar{\epsilon}$ is large. Consequently, the proof of Theorem 3.5 shows that for all $\rho > 0$ if n is large

$$\mathbf{P}[\|Y^q - Z^q\|_\infty > \rho] \leq O(1)\rho^{-1}|\log \epsilon|^{-1}. \tag{4.1}$$

Thus, the statement of Theorem 1.1 is obtained with t replaced by $t + q$. If we consider ϵ and ρ as fixed, then q is bounded. Since q is even, the following lemma completes the proof. ∎

LEMMA 4.1: *Let $t \geq cn$, $c > 1/2$. As $n \to \infty$, the total variation distance between the law of $\mathfrak{X}(\pi_t)$ and the law of $\mathfrak{X}(\pi_{t+2})$ tends to zero.*

First, we give a slightly informal proof. Note that when the largest entry in $\mathfrak{X}(\pi_\tau)$ is not too small, there is probability bounded away from 0 and 1 that

$\mathfrak{X}(\pi_\tau) = \mathfrak{X}(\pi_{\tau+2})$, because that entry may split and then recombine. We know that for many $\tau \in [n/2, t]$ the largest entry is not small. Consequently, there is a random "delay", which implies the statement of the lemma. For readers who are not convinced yet, we offer a proof with more details.

Proof: Set $W^\tau = \mathfrak{X}(\pi_\tau)$. To prove that W^t and W^{t+2} have close distributions, we couple the chain (W^τ) with a chain (U^τ) which has the same distribution as (W^τ). In essence, the two chains will be the same; the significant difference involves a random shift in time. Set $\tau_0 = \tau_0' = 0$. Inductively, suppose that τ_i and τ_i' have been defined such that $W^{\tau_i} = U^{\tau_i'}$. Let $m = m_i$ be the largest integer such that $W^{\tau_i+2j} = W^{\tau_i}$ for all $j = 1, 2, \ldots, m$. The distribution of m conditioned on W^{τ_i} is geometric; that is, $\mathbf{P}[m = k|W^{\tau_i}] = (1-p)p^k$, where $p = p_i = \mathbf{P}[m > 0|W^{\tau_i}]$. Similarly, the largest integer m' such that $U^{\tau_i'+2j} = U^{\tau_i'}$ for all $j = 1, 2, \ldots, m'$ has the same conditioned distribution: $\mathbf{P}[m' = k|W^{\tau_i}] = \mathbf{P}[m' = k|U^{\tau_i'}] = (1-p)p^k$. We now couple m and m'. If $\tau_i' = \tau_i + 2$, take $m' = m$. Otherwise we couple m and m' so that $|m - m'| \leq 1$, but $m \neq m'$ happens quite frequently. For example, for all $k \in \mathbb{N}$ take $(m, m') = (k, k+1)$ with probability $p^{k+1}(1-p)/(1+p)$, $(m, m') = (k+1, k)$ with the same probability, and $(m, m') = (0, 0)$ with probability $(1-p)/(1+p)$, all conditioned on W^{τ_i}. In this case, the conditioned probability that $m' - m = \pm 1$ is $2p/(1+p)$. In either case, take $U^{\tau_i'+j} = W^{\tau_i+j}$ for $j = 1, 2, \ldots, \min\{2m, 2m'\}$. If $m' > m$ let $U^{\tau_i'+2m+1}$ be independent from the chain (W^τ) given (W^{τ_i}, m, m'), and similarly if $m > m'$. Clearly, $W^{\tau_i+2m} = U^{\tau_i'+2m'}$. Take $U^{\tau_i'+2m'+j} = W^{\tau_i+2m+j}$ for $j = 1, 2$, $\tau_{i+1} := \tau_i + 2m + 2$ and $\tau_{i+1}' := \tau_i' + 2m' + 2$. Then continue inductively. This completes the specification of the coupling.

It clearly suffices to prove that with probability tending to 1 as $n \to \infty$, we have $\tau_i' = \tau_i + 2$ with some $\tau_i < t$. First, observe that the p_i are bounded away from 1. This guarantees that a.a.s. $\tau_i = O(i)$. Now note that p_i is bounded away from zero by some positive function of $W_1^{\tau_i}/n$ (the largest entry normalized), because that largest entry may split in the next step, and then the same two parts may merge in the step after that. We know, for example from (2.4), that with high probability for most values of i such that $t \geq \tau_i \geq (cn + n/2)/2$, the largest entry of W^{τ_i} is not too much smaller than n. Consequently, a.a.s. we have p_i bounded away from zero for many values of i satisfying $\tau_i < t$. Similarly, $m_i \neq m_i'$ for many values of i. Note that $(\tau_i - \tau_i')/2$ is a martingale, and its increments are $\{-1, 0, 1\}$. By removing the 0 increment steps, the martingale may be coupled with a simple random walk on $\mathbb{Z}$. The martingale starts at 0. Thus, the probability that many ± 1 steps are performed and it never gets to 1

tends to 0. This completes the proof. ∎

ACKNOWLEDGEMENT: We have had the pleasure to benefit from conversations with Rick Durrett, Michael Larsen, Russ Lyons, David Wilson and Ofer Zeitouni in connection with this work.

References

[Ang03] O. Angel, *Random infinite permutations and the cyclic time random walk,* in *Random Walks and Discrete Potential Theory* (C. Banderier and C. Krattenthaler, eds.), Discrete Mathematics and Theoretical Computer Science, 2003, pp. 9–16, http://dmtcs.loria.fr/proceedings/html/dmAC0101.abs.html.

[AS00] N. Alon and J. H. Spencer, *The Probabilistic Method,* second edition, Wiley-Interscience Series in Discrete Mathematics and Optimization, Wiley-Interscience, New York, 2000. With an appendix on the life and work of Paul Erdős.

[BD] N. Berestycki and R. Durrett, *A phase transition in the random transposition random walk,* arXiv:math.PR/0403259.

[DMP95] P. Diaconis, M. McGrath and J. Pitman, *Riffle shuffles, cycles, and descents,* Combinatorica **15** (1995), 11–29.

[DMWZZ] P. Diaconis, E. Mayer-Wolf, O. Zeitouni and M. P. Zerner, *The Poisson–Dirichlet law is the unique invariant distribution for uniform split-merge transformations,* The Annals of Probability **32** (2004), 915–938.

[DS81] P. Diaconis and M. Shahshahani, *Generating a random permutation with random transpositions,* Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **57** (1981), 159–179.

[Dud89] R. M. Dudley, *Real Analysis and Probability,* Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1989.

[Hol01] L. Holst, *The Poisson–Dirichlet distribution and its relatives revisited,* 2001, http://www.math.kth.se/matstat/fofu/reports/PoiDir.pdf, preprint.

[JLR00] S. Janson, T. Łuczak and A. Rucinski, *Random Graphs,* Wiley-Interscience Series in Discrete Mathematics and Optimization, Wiley-Interscience, New York, 2000.

[Spe94] J. Spencer, *Ten lectures on the probabilistic method,* Volume 64 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, second edition, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994.

[Tót93] B. Tóth, *Improved lower bound on the thermodynamic pressure of the spin 1/2 Heisenberg ferromagnet*, Letters in Mathematical Physics **28** (1993), 75–84.

[Wat76] G. A. Watterson, *The stationary distribution of the infinitely-many neutral alleles diffusion model*, Journal of Applied Probability **13** (1976), 639–651.

JOURNAL OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 22, Number 1, January 2009, Pages 167–210
S 0894-0347(08)00606-1
Article electronically published on July 28, 2008

TUG-OF-WAR AND THE INFINITY LAPLACIAN

YUVAL PERES, ODED SCHRAMM, SCOTT SHEFFIELD, AND DAVID B. WILSON

1. Introduction and preliminaries

1.1. **Overview.** We consider a class of zero-sum two-player stochastic games called **tug-of-war** and use them to prove that every bounded real-valued Lipschitz function F on a subset Y of a length space X admits a unique **absolutely minimal (AM)** extension to X, i.e., a unique Lipschitz extension $u : X \to \mathbb{R}$ for which $\mathrm{Lip}_U u = \mathrm{Lip}_{\partial U} u$ for all open $U \subset X \setminus Y$. We present an example that shows this is not generally true when F is merely Lipschitz and positive. (Recall that a metric space (X, d) is a **length space** if for all $x, y \in X$, the distance $d(x, y)$ is the infimum of the lengths of continuous paths in X that connect x to y. Length spaces are more general than geodesic spaces, where the infima need to be achieved.)

When X is the closure of a bounded domain $U \subset \mathbb{R}^n$ and Y is its boundary, a Lipschitz extension u of F is AM if and only it is **infinity harmonic** in the interior of $X \setminus Y$; i.e., it is a **viscosity solution** (defined below) to $\Delta_\infty u = 0$, where Δ_∞ is the so-called **infinity Laplacian**

$$\Delta_\infty u = |\nabla u|^{-2} \sum_{i,j} u_{x_i} u_{x_i x_j} u_{x_j} \tag{1.1}$$

(informally, this is the second derivative of u in the direction of the gradient of u). Aronsson proved this equivalence for smooth u in 1967, and Jensen proved the general statement in 1993 [1, 13]. Our analysis of tug-of-war also shows that in this setting $\Delta_\infty u = g$ has a unique viscosity solution (extending F) when $g : \overline{U} \to \mathbb{R}$ is continuous and $\inf g > 0$ or $\sup g < 0$, but not necessarily when g assumes values of both signs. We note that in the study of the homogenous equation $\Delta_\infty u = 0$, the normalizing factor $|\nabla u|^{-2}$ in (1.1) is sometimes omitted; however, it is important to include it in the non-homogenous equation. Observe that with the normalization, Δ_∞ coincides with the ordinary Laplacian Δ in the one-dimensional case.

Unlike the ordinary Laplacian or the p-Laplacian for $p < \infty$, the infinity Laplacian can be defined on **any** length space with no additional structure (such as a measure or a canonical Markov semigroup)—that is, we will see that viscosity solutions to $\Delta_\infty u = g$ are well defined in this generality. We will establish the above stated uniqueness of solutions u to $\Delta_\infty u = g$ in the setting of length spaces.

Originally, we were motivated not by the infinity Laplacian but by random turn Hex [22] and its generalizations, which led us to consider the tug-of-war game. As

Received by the editors July 11, 2006.
2000 *Mathematics Subject Classification.* Primary 91A15, 91A24, 35J70, 54E35, 49N70.
Key words and phrases. Infinity Laplacian, absolutely minimal Lipschitz extension, tug-of-war.
Research of the first and third authors was supported in part by NSF grants DMS-0244479 and DMS-0104073.

©2008 by the authors. This paper or any part thereof may be reproduced for non-commercial purposes.

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_18, © Springer Science+Business Media, LLC 2011

we later learned, tug-of-war games have been considered by Lazarus, Loeb, Propp and Ullman in [16] (see also [15]).

Tug-of-war on a metric space is very natural and conceivably applicable (like differential game theory) to economic and political modeling.

The intuition provided by thinking of strategies for tug-of-war yields new results even in the classical setting of domains in $\mathbb{R}^n$. For instance, in Section 4 we show that if u is infinity harmonic in the unit disk and its boundary values are in $[0,1]$ and supported on a δ-neighborhood of the ternary Cantor set on the unit circle, then $u(0) < \delta^\beta$ for some $\beta > 0$.

Before precisely stating our main results, we need several definitions.

1.2. **Random turn games and values.** We consider two-player, zero-sum **random-turn games**, which are defined by the following parameters: a set X of **states** of the game, two directed **transition graphs** $E_{\mathrm{I}}, E_{\mathrm{II}}$ with vertex set X, a non-empty set $Y \subset X$ of **terminal states** (a.k.a. **absorbing states**), a **terminal payoff function** $F : Y \to \mathbb{R}$, a **running payoff function** $f : X \setminus Y \to \mathbb{R}$, and an **initial state** $x_0 \in X$.

The game play is as follows: a token is initially placed at position x_0. At the kth step of the game, a fair coin is tossed, and the player who wins the toss may move the token to any x_k for which (x_{k-1}, x_k) is a directed edge in her transition graph. The game ends the first time $x_k \in Y$, and player I's payoff is $F(x_k) + \sum_{i=0}^{k-1} f(x_i)$. Player I seeks to maximize this payoff, and since the game is zero-sum, player II seeks to minimize it.

We will use the term **tug-of-war (on the graph with edges E)** to describe the game in which $E := E_{\mathrm{I}} = E_{\mathrm{II}}$ (i.e., players have identical move options) and E is undirected (i.e., all moves are reversible). Generally, our results pertain only to the undirected setting. Occasionally, we will also mention some counterexamples showing that the corresponding results do not hold in the directed case.

In the most conventional version of tug-of-war on a graph, Y is a union of "target sets" Y^{I} and Y^{II}, there is no running payoff ($f = 0$), and F is identically 1 on Y^{I} and identically 0 on Y^{II}. Players then try to "tug" the game token to their respective targets (and away from their opponent's targets), and the game ends when a target is reached.

A **strategy** for a player is a way of choosing the player's next move as a function of all previously played moves and all previous coin tosses. It is a map from the set of partially played games to moves (or in the case of a **random strategy**, a probability distribution on moves). Normally, one would think of a good strategy as being Markovian, i.e., as a map from the current state to the next move, but it is useful to allow more general strategies that take into account the history.

Given two strategies $\mathcal{S}_{\mathrm{I}}, \mathcal{S}_{\mathrm{II}}$, let $F_-(\mathcal{S}_{\mathrm{I}}, \mathcal{S}_{\mathrm{II}})$ and $F_+(\mathcal{S}_{\mathrm{I}}, \mathcal{S}_{\mathrm{II}})$ be the expected total payoff (including the running payoffs received) at the termination of the game, if the game terminates with probability one and this expectation exists in $[-\infty, \infty]$; otherwise, let $F_-(\mathcal{S}_{\mathrm{I}}, \mathcal{S}_{\mathrm{II}}) = -\infty$ and $F_+(\mathcal{S}_{\mathrm{I}}, \mathcal{S}_{\mathrm{II}}) = +\infty$.

The **value of the game for player** I is defined as $\sup_{\mathcal{S}_{\mathrm{I}}} \inf_{\mathcal{S}_{\mathrm{II}}} F_-(\mathcal{S}_{\mathrm{I}}, \mathcal{S}_{\mathrm{II}})$. The **value for player** II is $\inf_{\mathcal{S}_{\mathrm{II}}} \sup_{\mathcal{S}_{\mathrm{I}}} F_+(\mathcal{S}_{\mathrm{I}}, \mathcal{S}_{\mathrm{II}})$. We use the expressions $u_{\mathrm{I}}(x)$ and $u_{\mathrm{II}}(x)$ to denote the values for players I and II, respectively, as a function of the starting state x of the game.

Note that if player I cannot force the game to end almost surely, then $u_{\mathrm{I}} = -\infty$, and if player II cannot force the game to end almost surely, then $u_{\mathrm{II}} = \infty$. Clearly,

$u_{\mathrm{I}}(x) \le u_{\mathrm{II}}(x)$. When $u_{\mathrm{I}}(x) = u_{\mathrm{II}}(x)$, we say that the game has a **value**, given by $u(x) := u_{\mathrm{I}}(x) = u_{\mathrm{II}}(x)$.

Our definition of value for player I penalizes player I severely for not forcing the game to terminate with probability one, awarding $-\infty$ in this case.

(As an alternative definition, one could define F_-, and hence player I's value, by assigning payoffs to all of the non-terminating sequences $x_0, x_1, x_2, \ldots$. If the payoff function for the non-terminating games is a zero-sum Borel-measurable function of the infinite sequence, then player I's value is equal to player II's value in great generality [17]; see also [20] for more on stochastic games. The existence of a value by our strong definition implies the existence and equality of the values defined by these alternative definitions.)

Considering the two possibilities for the first coin toss yields the following lemma, a variant of which appears in [16].

Lemma 1.1. *The function $u = u_{\mathrm{I}}$ satisfies the equation*

$$u(x) = \frac{1}{2}\left(\sup_{y:(x,y)\in E_1} u(y) + \inf_{y:(x,y)\in E_2} u(y)\right) + f(x) \tag{1.2}$$

for every non-terminal state $x \in X \setminus Y$ for which the right-hand side is well defined, and $u_{\mathrm{I}}(x) = -\infty$ when the right-hand side is of the form $\frac{1}{2}(\infty + (-\infty)) + f(x)$. The analogous statement holds for u_{II}, except that $u_{\mathrm{II}}(x) = +\infty$ when the right-hand side of (1.2) is of the form $\frac{1}{2}(\infty + (-\infty)) + f(x)$.

When $E = E_1 = E_2$, the operator

$$\Delta_\infty u(x) := \sup_{y:(x,y)\in E} u(y) + \inf_{y:(x,y)\in E} u(y) - 2u(x)$$

is called the **(discrete) infinity Laplacian**. A function u is **infinity harmonic** if (1.2) holds and $f(x) = 0$ at all non-terminal $x \in X \setminus Y$. When u is finite, this is equivalent to $\Delta_\infty u = 0$. However, it will be convenient to adopt the convention that u is infinity harmonic at x if $u(x) = +\infty$ (resp. $-\infty$) and the right-hand side in (1.2) is also $+\infty$ (resp. $-\infty$). Similarly, it will be convenient to say "u is a solution to $\Delta_\infty u = -2\,f$" at x if $u(x) = +\infty$ (resp. $-\infty$) and the right-hand side in (1.2) is also $+\infty$ (resp. $-\infty$).

In a tug-of-war game, it is natural to guess that the value $u = u_{\mathrm{I}} = u_{\mathrm{II}}$ exists and is the *unique* solution to

$$\Delta_\infty u(x) = -2f,$$

and also that (at least when E is locally finite) player I's optimal strategy will be to always move to the vertex that maximizes $u(x)$ and that player II's optimal strategy will be to always move to the vertex that minimizes $u(x)$. This is easy to prove when E is undirected and finite and f is everywhere positive or everywhere negative. Subtleties arise in more general cases (X infinite, E directed, F unbounded, f having values of both signs, etc.).

Our first theorem addresses the question of the existence of a value.

Theorem 1.2. *A tug-of-war game with parameters X, E, Y, F, f has a value whenever the following hold:*

(1) *Either $f = 0$ everywhere or $\inf f > 0$.*
(2) *$\inf F > -\infty$.*
(3) *E is undirected.*

Counterexamples exist when any one of the three criteria is removed. In Section 5 (a section devoted to counterexamples) we give an example of a tug-of-war game without a value, where E is undirected, $F = 0$, and the running payoff satisfies $f > 0$ but $\inf f = 0$.

The case where $f = 0$, F is bounded, and (X, E) is locally finite was proved earlier in [15]. That paper discusses an (essentially non-random) game in which the two players bid for the right to choose the next move. That game, called the Richman game, has the same value as tug-of-war with $f = 0$, where F takes the values 0 and 1. Additionally, a simple and efficient algorithm for calculating the value when $f = 0$ and (X, E) is finite is presented there.

1.3. **Tug-of-war on a metric space.** Now consider the special case where (X, d) is a metric space, $Y \subset X$, and Lipschitz functions $F : Y \to \mathbb{R}$ and $f : X \setminus Y \to \mathbb{R}$ are given. Let E_ε be the edge-set in which $x \sim y$ if and only if $d(x, y) < \varepsilon$, and let u^ε be the value (if it exists) of the game played on E_ε with terminal payoff F and running payoff normalized to be $\varepsilon^2 f$.

In other words, $u^\varepsilon(x)$ is the value of the following two-player zero-sum game, called **ε-tug-of-war**: fix $x_0 = x \in X \setminus Y$. At the kth turn, the players toss a coin and the winner chooses an x_k with $d(x_k, x_{k-1}) < \varepsilon$. The game ends when $x_k \in Y$, and player I's payoff is $F(x_k) + \varepsilon^2 \sum_{i=0}^{k-1} f(x_i)$.

When the limit $u := \lim_{\varepsilon \to 0} u^\varepsilon$ exists pointwise, we call u the **continuum value** (or just "value") of the quintuple (X, d, Y, F, f). We define the **continuum value for player** I (or II) analogously.

The reader may wonder why we have chosen not to put an edge in E_ε between x and y when $d(x, y) = \varepsilon$ exactly. This choice has some technical implications. Specifically, we will compare the ε-game with the $2\,\varepsilon$-game. If x, z are such that $d(x, z) \le 2\,\varepsilon$, then in a length space it does not follow that there is a y such that $d(x, y) \le \varepsilon$ and $d(y, z) \le \varepsilon$. However, it does follow if you replace the weak inequalities with strong inequalities throughout.

We prove the following:

Theorem 1.3. *Suppose X is a length space, $Y \subset X$ is non-empty, $F : Y \to \mathbb{R}$ is bounded below and has an extension to a uniformly continuous function on X, and either $f : X \setminus Y \to \mathbb{R}$ satisfies $f = 0$ or all three of the following hold: $\inf |f| > 0$, f is uniformly continuous, and X has finite diameter. Then the continuum value u exists and is a uniformly continuous function extending F. Furthermore, $\|u - u^\varepsilon\|_\infty \to 0$ as $\varepsilon \searrow 0$. If F is Lipschitz, then so is u. If F and f are Lipschitz, then $\|u - u^\varepsilon\|_\infty = O(\varepsilon)$.*

The above condition that $F : Y \to \mathbb{R}$ extends to a uniformly continuous function on X is equivalent to having F uniformly continuous on Y and "Lipschitz on large scales," as we prove in Lemma 3.9 below.

We will see in Section 5.1 that this fails in general when $f > 0$ but $\inf f = 0$. When f assumes values of both signs, it fails even when X is a closed disk in $\mathbb{R}^2$, Y is its boundary and $F = 0$. In Section 5.3 we show by means of an example that in such circumstances it may happen that $u_{\text{I}}^\varepsilon \ne u_{\text{II}}^\varepsilon$ and moreover, $\liminf_{\varepsilon \searrow 0} \|u_{\text{I}}^\varepsilon - u_{\text{II}}^\varepsilon\|_\infty > 0$.

1.4. **Absolutely minimal Lipschitz extensions.** Given a metric space (X, d), a subset $Y \subset X$ and a function $u : X \to \mathbb{R}$, we write

$$\operatorname{Lip}_Y u = \sup_{x,y \in Y} |u(y) - u(x)|/d(x, y)$$

and $\operatorname{Lip} u = \operatorname{Lip}_X u$. Thus u is Lipschitz iff $\operatorname{Lip} u < \infty$. Given $F : Y \to \mathbb{R}$, we say that $u : X \to \mathbb{R}$ is a **minimal extension of** F if $\operatorname{Lip}_X u = \operatorname{Lip}_Y F$ and $u(y) = F(y)$ for all $y \in Y$.

It is well known that for any metric space X, any Lipschitz F on a subset Y of X admits a minimal extension. The largest and smallest minimal extensions (introduced by McShane [18] and Whitney [24] in the 1930's) are respectively

$$\inf_{y \in Y} [F(y) + \operatorname{Lip}_Y F \, d(x, y)] \qquad \text{and} \qquad \sup_{y \in Y} [F(y) - \operatorname{Lip}_Y F \, d(x, y)] .$$

We say u is an **absolutely minimal (AM) extension of** F if $\operatorname{Lip} u < \infty$ and $\operatorname{Lip}_U u = \operatorname{Lip}_{\partial U} u$ for every open set $U \subset X \setminus Y$. We say that u is AM on U if it is defined on $\overline{U}$ and is an AM extension of its restriction to ∂U. AM extensions were first introduced by Aronsson in 1967 [1] and have applications in engineering and image processing (see [4] for a recent survey).

We prove the following:

Theorem 1.4. *Let X be a length space and let $F : Y \to \mathbb{R}$ be Lipschitz, where $\emptyset \neq Y \subset X$. If $\inf F > -\infty$, then the continuum value function u described in Theorem* 1.3 *(with $f = 0$) is an AM extension of F. If F is also bounded, then u is the unique AM extension of F.*

We present in the counterexample section, Section 5, an example in which F is Lipschitz, non-negative, and unbounded, and although the continuum value is an AM extension, it is not the only AM extension.

Prior to our work, the existence of AM extensions in the above settings was known only for *separable* length spaces [14] (see also [19]). The uniqueness in Theorem 1.4 was known only in the case that X is the closure of a bounded domain $U \subset \mathbb{R}^n$ and $Y = \partial U$. (To deduce this case from Theorem 1.4, one needs to replace X by the smallest closed ball containing $\overline{U}$, say.) Three uniqueness proofs in this setting have been published, by Jensen [13], by Barles and Busca [6], and by Aronsson, Crandall, and Juutinen [4]. The third proof generalizes from the Euclidean norm to uniformly convex norms.

Our proof applies to more general spaces because it invokes no outside theorems from analysis (which assume existence of a local Euclidean geometry, a measure, a notion of twice differentiability, etc.), and relies only on the structure of X as a length space.

As noted in [4], AM extensions do not generally exist on metric spaces that are not length spaces. (For example, if X is the L-shaped region $\{0\} \times [0, 1] \cup [0, 1] \times \{0\} \subset \mathbb{R}^2$ with the Euclidean metric, and $Y = \{(0, 1), (1, 0)\}$, then no non-constant $F : Y \to \mathbb{R}$ has an AM extension. Indeed, suppose that $u : X \to \mathbb{R}$ is an AM extension of $F : Y \to \mathbb{R}$. Let $a := u(0, 0)$, $b := u(0, 1)$ and $c := u(1, 0)$. Then, considering $U = \{0\} \times (0, 1)$, it follows that $u(0, s) = a + s\,(b - a)$. Likewise, $u(s, 0) = a + s\,(c - a)$. Now taking $U_\epsilon := \{0\} \times [0, 1) \cup [0, \epsilon) \times \{0\}$, we see that $\lim_{\varepsilon \searrow 0} \operatorname{Lip}_{U_\varepsilon} u = |b - a|$. Hence $|c - a| \leq |b - a|$. By symmetry, $|c - a| = |b - a|$. Since F is assumed to be non-constant, $b \neq c$, and hence $c - a = a - b$. Then $\big|u(0, s) - u(s, 0)\big|/\big(\sqrt{2}s\big) = \sqrt{2}\,|b - a|$, which contradicts $\lim_{\varepsilon \searrow 0} \operatorname{Lip}_{U_\varepsilon} u = |b - a|$.)

One property that makes length spaces special is the fact that the Lipschitz norm is determined locally. More precisely, if $W \subset X$ is closed, then either $\mathrm{Lip}_W u = \mathrm{Lip}_{\partial W} u$ or for every $\delta > 0$

$$\sup\left\{ \frac{|u(x) - u(y)|}{d(x,y)} : x, y \in W,\, 0 < d(x,y) < \delta \right\} = \mathrm{Lip}_W u\,.$$

The definition of AM was inspired by the notion that if u is the "tautest possible" Lipschitz extension of F, it should be tautest possible on any open $V \subset X \setminus Y$, given the values of u on ∂V and ignoring the rest of the metric space. Without locality, the rest of the metric space cannot be ignored (since long-distance effects may change the global Lipschitz constant), and the definition of AM is less natural. Another important property of length spaces is the fact that the graph distance metric on E_ε scaled by ε approximates the original metric, namely, it is within ε of $d(\cdot,\cdot)$.

1.5. **Infinity Laplacian on $\mathbb{R}^n$.** The continuum version of the **infinity Laplacian** is defined for C^2 functions u on domains $U \subset \mathbb{R}^n$ by

$$\Delta_\infty u = |\nabla u|^{-2} \sum_{i,j} u_{x_i} u_{x_i x_j} u_{x_j}.$$

This is the same as $\eta^T H \eta$, where H is the Hessian of u and $\eta = \nabla u / |\nabla u|$. Informally, $\Delta_\infty u$ is the second derivative of u in the direction of the gradient of u. If $\nabla u(x) = 0$, then $\Delta_\infty u(x)$ is undefined; however, we adopt the convention that if the second derivative of $u(x)$ happens to be the same in *every* direction (i.e., the matrix $\{u_{x_i x_j}\}$ is λ times the identity), then $\Delta_\infty u(x) = \lambda$, which is the second derivative in any direction. (As mentioned above, some texts on infinity harmonic functions define Δ_∞ without the normalizing factor $|\nabla u|^{-2}$. When discussing viscosity solutions to $\Delta_\infty u = 0$, the two definitions are equivalent. The fact that the normalized version is sometimes undefined when $\nabla u = 0$ does not matter because it *is* always well defined at x when φ is a **cone function**, i.e., when $\varphi(z)$ has the form $a|x - z| + b$ for $a, b \in \mathbb{R}$ and $z \in \mathbb{R}^n$ with $z \neq x$, and viscosity solutions can be defined via comparison with cones; see Section 1.6.) As in the discrete setting, u is **infinity harmonic** if $\Delta_\infty u = 0$.

While discrete infinity harmonic functions are a recent concept, introduced in finite-difference schemes for approximating continuous infinity harmonic functions [21], related notions of value for stochastic games are of course much older. The continuous infinity Laplacian first appeared in the work of Aronsson [1] and has been very thoroughly studied [4]. Key motivations for studying this operator are the following:

(1) **AM extensions:** Aronsson proved that C^2 extensions u on domains $U \subset \mathbb{R}^n$ (of functions F on ∂U) are infinity harmonic if and only if they are AM.

(2) **p-harmonic functions:** As noted by Aronsson [1], the infinity Laplacian is the formal limit, as $p \to \infty$ of the (properly normalized) p-Laplacians. Recall that **p-harmonic** functions, i.e., minimizers u of $\int |\nabla u(x)|^p \, dx$ subject to boundary conditions, solve the **Euler-Lagrange equation**

$$\nabla \cdot (|\nabla u|^{p-2} \nabla u) = 0,$$

which can be rewritten

$$|\nabla u|^{p-2} \left(\Delta u + (p-2)\Delta_\infty u\right) = 0,$$

where Δ is the ordinary Laplacian. Dividing by $|\nabla u|^{p-2}$, we see that (at least when $|\nabla u| \neq 0$) p-harmonic functions satisfy $\Delta_p u = 0$, where $\Delta_p := \Delta_\infty + (p-2)^{-1}\Delta$; the second term vanishes in the large p limit. It is not too hard to see that as p tends to infinity, the Lipschitz norm of any limit of the p-harmonic functions extending F will be $\mathrm{Lip}_{\partial U} F$. So it is natural to guess (and was proved in [7]) that as p tends to infinity the p-harmonic extensions of F converge to a limit that is both absolutely minimal and a viscosity solution to $\Delta_\infty u = 0$.

In the above setting, Aronsson also proved that there always exists an AM extension, and that in the planar case $U \subset \mathbb{R}^2$, there exists at most one C^2 infinity harmonic extension; however C^2 infinity harmonic extensions do not always exist [2].

To define the infinity Laplacian in the non-C^2 setting requires us to consider weak solutions; the right notion here is that of *viscosity solution*, as introduced by Crandall and Lions (1983) [11]. Start by observing that if u and v are C^2 functions, $u(x) = v(x)$, and $v \geq u$ in a neighborhood of x, then $v - u$ has a local minimum at x, whence $\Delta_\infty v(x) \geq \Delta_\infty u(x)$ (if both sides of this inequality are defined). This comparison principle (which has analogs for more general degenerate elliptic PDEs [5]) suggests that if u is not C^2, in order to define $\Delta_\infty u(x)$ we want to compare it to C^2 functions φ for which $\Delta_\infty \varphi(x)$ is defined. Let $S(x)$ be the set of real valued functions φ defined and C^2 in a neighborhood of x for which $\Delta_\infty \varphi(x)$ has been defined; that is, either $\nabla\varphi(x) \neq 0$, or $\nabla\varphi(x) = 0$ and the limit $\Delta_\infty \varphi(x) := \lim_{x' \to x} 2\,\frac{\varphi(x') - \varphi(x)}{|x'-x|^2}$ exists.

Definition. Let X be a domain in $\mathbb{R}^n$ and let $u : X \to \mathbb{R}$ be continuous. Set

$$\Delta_\infty^+ u(x) = \inf\{\Delta_\infty \varphi(x) \,:\, \varphi \in S(x) \text{ and } x \text{ is a local minimum of } \varphi - u\}\,. \tag{1.3}$$

Thus u satisfies $\Delta_\infty^+(u) \geq g$ in a domain X, iff every $\varphi \in C^2$ such that $\varphi - u$ has a local minimum at some $x \in X$ satisfies $\Delta_\infty^+ \varphi(x) \geq g(x)$. In this case u is called a **viscosity subsolution** of $\Delta_\infty(\cdot) = g$. Note that if $\varphi \in C^2$, then $\Delta_\infty^+ \varphi = \Delta_\infty \varphi$ wherever $\nabla\varphi \neq 0$.

Similarly, let

$$\Delta_\infty^- u(x) = \sup\{\Delta_\infty \varphi(x) \,:\, \varphi \in S(x) \text{ and } x \text{ is a local maximum of } \varphi - u\}\,, \tag{1.4}$$

and call u a **viscosity supersolution** of $\Delta_\infty(\cdot) = g$ iff $\Delta_\infty^- u \leq g$ in X.

Finally, u is a **viscosity solution** of $\Delta_\infty(\cdot) = g$ if $\Delta_\infty^- u \leq g \leq \Delta_\infty^+ u$ in X (i.e., u is both a supersolution and a subsolution).

Here is a little caveat. At present, we do not know how to show that $\Delta_\infty u = g$ in the viscosity sense determines g. For example, if u is Lipschitz, g_1 and g_2 are continuous, and $\Delta_\infty u = g_j$ holds for $j = 1, 2$ (in the viscosity sense), how does one prove that $g_1 = g_2$?

The following result of Jensen (alluded to above) is now well known [1, 13, 4]: if X is a domain in $\mathbb{R}^n$ and $u : X \to \mathbb{R}$ is continuous, then $\mathrm{Lip}_U u = \mathrm{Lip}_{\partial U} u < \infty$ for every bounded open set $U \subset \overline{U} \subset X$ (i.e., u is AM) if and only if u is a viscosity solution to $\Delta_\infty u = 0$ in X.

Let $A \subset Y \subset X$, where A is closed, $Y \neq \emptyset$ and X is a length space. If $x \in X$, one can define the **∞-harmonic measure of A from x** as the infimum of $u(x)$ over all functions $u : X \to [0, \infty)$ that are Lipschitz on X, AM in $X \setminus Y$ and satisfy

$u \geq 1$ on A. This quantity will be denoted by $\omega_\infty(A) = \omega_\infty^{(x,Y,X)}(A)$. In Section 4 we prove

Theorem 1.5. *Let X be the unit ball in $\mathbb{R}^n$, $n > 1$, let $Y = \partial X$, let x be the center of the ball, and for each $\delta > 0$ let $A_\delta \subset Y$ be a spherical cap of radius δ (of dimension $n-1$). Then*

$$c\,\delta^{1/3} \leq \omega_\infty(A_\delta) \leq C\,\delta^{1/3},$$

where $c, C > 0$ are absolute constants (which do not depend on n).

Numerical calculations [21] had suggested that in the setting of the theorem $\omega_\infty(A_\delta)$ tends to 0 as $\delta \to 0$, but this was only recently proved [12], and the proof did not yield any quantitative information on the rate of decay. In contrast to our other theorems in the paper, the proof of this theorem does not use tug-of-war. The primary tool is the comparison of a specific AM function in $\mathbb{R}^2 \setminus \{0\}$ with decay $r^{-1/3}$ discovered by Aronsson [3].

1.6. **Quadratic comparison on length spaces.** To motivate the next definition, observe that for continuous functions $u : \mathbb{R} \to \mathbb{R}$, the inequality $\Delta_\infty^+ u \geq 0$ reduces to convexity. The definition of convexity requiring a function to lie below its chords has an analog, *comparison with cones*, which characterizes infinity harmonic functions. Call the function $\varphi(y) = b\,|y-z| + c$ a **cone** based at $z \in \mathbb{R}^n$. For an open $U \subset \mathbb{R}^n$, say that a continuous $u : U \to \mathbb{R}$ satisfies **comparison with cones from above** on U if for every open $W \subset \overline{W} \subset U$ for every $z \in \mathbb{R}^n \setminus W$, and for every cone φ based at z such that the inequality $u \leq \varphi$ holds on ∂W, the same inequality is valid throughout W. **Comparison with cones from below** is defined similarly using the inequality $u \geq \varphi$.

Jensen [13] proved that viscosity solutions to $\Delta_\infty u = 0$ for domains in $\mathbb{R}^n$ satisfy comparison with cones (from above and below), and Crandall, Evans, and Gariepy [9] proved that a function on $\mathbb{R}^n$ is absolutely minimal in a bounded domain U if and only if it satisfies comparison with cones in U.

Champion and De Pascale [8] adapted this definition to length spaces, where cones are replaced by functions of the form $\varphi(x) = b\,d(x,z) + c$, where $b > 0$. Their precise definition is as follows. Let U be an open subset of a length-space X and let $u : \overline{U} \to \mathbb{R}$ be continuous. Then u is said to satisfy **comparison with distance functions from above** on U if for every open $W \subset U$, for every $z \in X \setminus W$, for every $b \geq 0$ and for every $c \in \mathbb{R}$, if $u(x) \leq b\,d(x,z) + c$ holds on ∂W, then it also holds in W. The function u is said to satisfy **comparison with distance functions from below** if $-u$ satisfies comparison with distance functions from above. Finally, u satisfies **comparison with distance functions** if it satisfies comparison with distance functions from above and from below.

The following result from [8] will be used in the proof of Theorem 1.4:

Lemma 1.6 ([8])**.** *Let U be an open subset of a length space. A continuous $u : \overline{U} \to \mathbb{R}$ satisfies comparison with distance functions in U if and only if it is AM in U.*

To study the inhomogenous equation $\Delta_\infty u = g$, because u will in general have a non-zero second derivative (in its gradient direction), it is natural to extend these definitions to comparison with functions that have a quadratic term.

Definitions. Let $Q(r) = ar^2 + br + c$ with $r, a, b, c \in \mathbb{R}$ and let X be a length space.

- Let $z \in X$. We call the function $\varphi(x) = Q(d(x,z))$ a **quadratic distance function** (centered at z).
- We say that a quadratic distance function $\varphi(x) = Q(d(x,z))$ is **$\star$-increasing** (in distance from z) on an open set $V \subset X$ if either (1) $z \notin V$ and for every $x \in V$, we have $Q'(d(x,z)) > 0$, *or* (2) $z \in V$ and $b = 0$ and $a > 0$. Similarly, we say that a quadratic distance function φ is **$\star$-decreasing** on V if $-\varphi$ is $\star$-increasing on V.
- If $u : U \to \mathbb{R}$ is a continuous function defined on an open set U in a length space X, we say that u satisfies **g-quadratic comparison** on U if the following two conditions hold:
  (1) g-QUADRATIC COMPARISON FROM ABOVE: For every open $V \subset \overline{V} \subset U$ and $\star$-increasing quadratic distance function φ on V with quadratic term $a \leq \inf_{y \in V} \frac{g(y)}{2}$, the inequality $\varphi \geq u$ on ∂V implies $\varphi \geq u$ on V.
  (2) g-QUADRATIC COMPARISON FROM BELOW: For every open $V \subset \overline{V} \subset U$ and $\star$-decreasing quadratic distance function φ on V with quadratic term $a \geq \sup_{y \in V} \frac{g(y)}{2}$, the inequality $\varphi \leq u$ on ∂V implies $\varphi \leq u$ on V.

The following theorem is proved in Section 6.

Theorem 1.7. *Let u be a real-valued continuous function on a bounded domain U in $\mathbb{R}^n$, and suppose that g is a continuous function on U. Then u satisfies g-quadratic comparison on U if and only if u is a viscosity solution to $\Delta_\infty u = g$ in U.*

This equivalence motivates the study of functions satisfying quadratic comparison. Note that satisfying $\Delta_\infty u(x) = g(x)$ in the viscosity sense depends only on the local behavior of u near x. We may use Theorem 1.7 to extend the definition of Δ_∞ to length spaces, saying that $\Delta_\infty u = g$ on an open subset U of a length space if and only if every $x \in U$ has a neighborhood $V \subset U$ on which u satisfies g-quadratic comparison. We warn the reader, however, that even for length spaces X contained within $\mathbb{R}$, there can be solutions to $\Delta_\infty u = 0$ that do not satisfy comparison with distance functions (or 0-quadratic comparison, for that matter): for example, $X = U = (0,1)$ and $u(x) = x$. The point here is that when we take $V \subset (0,1)$ and compare u with some function φ on ∂V, the "appropriate" notion of the boundary ∂V is the boundary in $\mathbb{R}$, not in X.

The continuum value of the tug-of-war game sometimes gives a construction of a function satisfying g-quadratic comparison. We prove:

Theorem 1.8. *Suppose X is a length space, $Y \subset X$ is non-empty, $F : Y \to \mathbb{R}$ is uniformly continuous and bounded, and either $f : X \setminus Y \to \mathbb{R}$ satisfies $f = 0$ or all three of the following hold: $\inf |f| > 0$, f is uniformly continuous, and X has finite diameter. Then the continuum value u is the unique continuous function satisfying $(-2f)$-quadratic comparison on $X \setminus Y$ and $u = F$ on Y. Moreover, if $\tilde{u} : X \to \mathbb{R}$ is continuous, satisfies $\tilde{u} \geq F$ on Y and $(-2f)$-quadratic comparison from below on $X \setminus Y$, then $\tilde{u} \geq u$ throughout X.*

Putting these last two theorems together, we obtain

Corollary 1.9. *Suppose $U \subset \mathbb{R}^n$ is a bounded open set, $F : \partial U \to \mathbb{R}$ is uniformly continuous, and $f : U \to \mathbb{R}$ satisfies either $f = 0$ or $\inf f > 0$ and f is uniformly continuous. Then there is a unique continuous function $u : \overline{U} \to \mathbb{R}$ that is a viscosity solution to $\Delta_\infty u = -2f$ on U and satisfies $u = F$ on ∂U. This unique solution is the continuum value of tug-of-war on $(\overline{U}, d, \partial U, F, f)$.*

It is easy to verify that F indeed satisfies the assumptions in Theorem 1.8. In order to deduce the corollary, we may take the length space X as a ball in $\mathbb{R}^n$ which contains U and extend F to $X \setminus U$, say. Alternatively, we may consider U with its intrinsic metric and lift F to the completion of U.

We present in Section 5 an example showing that the corollary may fail if f is permitted to take values of both signs. Specifically, the example describes two functions u_1, u_2 defined in the closed unit disk in $\mathbb{R}^2$ and having boundary values identically zero on the unit circle such that with some Lipschitz function g we have $\Delta_\infty u_j = g$ for $j = 1, 2$ (in the viscosity sense), while $u_1 \neq u_2$.

The plan of the paper is as follows. Section 2 discusses the discrete tug-of-war on graphs and proves Theorem 1.2, and Section 3 deals with tug-of-war on length spaces and proves Theorems 1.3, 1.4 and 1.8. Section 4 is devoted to estimates of ∞-harmonic measure. In Section 5, we present a few counterexamples showing that some of the assumptions in the theorems we prove are necessary. Section 6 is devoted to the proof of Theorem 1.7. Section 7 presents some heuristic arguments describing what the limiting trajectories of some ϵ-tug-of-war games on domains in $\mathbb{R}^n$ may look like and states a question regarding the length of the game. We conclude with additional open problems in Section 8.

2. Discrete game value existence

2.1. **Tug-of-war on graphs without running payoffs.** In this section, we will generally assume that $E = E_{\mathrm{I}} = E_{\mathrm{II}}$ is undirected and connected and that $f = 0$.

Though we will not use this fact, it is interesting to point out that in the case where E is finite, there is a simple algorithm from [15] which calculates the value u of the game and proceeds as follows. Assuming the value $u(v)$ is already calculated at some set $V' \supset Y$ of vertices, find a path $v_0, v_1, \ldots, v_k$, $k > 1$, with interior vertices $v_1, \ldots, v_{k-1} \in X \setminus V'$ and endpoints $v_0, v_k \in V'$ which maximizes $\big(u(v_k) - u(v_0)\big)/k$ and set $u(v_i) = u(v_0) + i\,\big(u(v_k) - u(v_0)\big)/k$ for $i = 1, 2, \ldots, k-1$. Repeat this as long as $V' \neq X$.

Recall that u_{I} is the value function for player I.

Lemma 2.1. *Suppose that $\inf_Y F > -\infty$ and $f = 0$. Then u_{I} is the smallest ∞-harmonic function bounded from below on X that extends F. More generally, if v is an ∞-harmonic function which is bounded from below on X and $v \geq F$ on Y, then $v \geq u_{\mathrm{I}}$ on X.*

Similarly, if F is bounded from above on Y, then u_{II} is the largest ∞-harmonic function bounded from above on X that extends F.

Proof. Player I could always try to move closer to some specific point $y \in Y$. Since in almost every infinite sequence of fair coin tosses there will be a time when the number of tails exceeds the number of heads by $d(x_0, y)$, this ensures that the game terminates a.s., and we have $u_{\mathrm{I}} \geq \inf_Y F > -\infty$. Suppose that $v \geq F$ on Y and

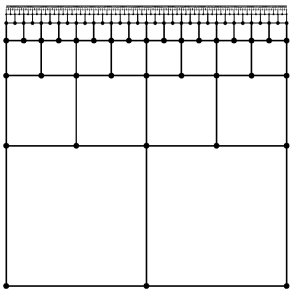


FIGURE 1. A graph for which the obvious tug-of-war strategy of maximizing/minimizing u does poorly.

v is ∞-harmonic on X. Given $\delta > 0$, consider an arbitrary strategy for player I and let II play a strategy that at step k (if II wins the coin toss) II selects a state where $v(\cdot)$ is within $\delta 2^{-k}$ of its infimum among the states to which II could move. We will show that the expected payoff for player I is at most $v(x_0) + \delta$. We may assume the game terminates a.s. at a time $\tau < \infty$. Let $\{x_j\}_{j \geq 0}$ denote the random sequence of states encountered in the game. Since v is ∞-harmonic, the sequence $M_k = v(x_{k \wedge \tau}) + \delta 2^{-k}$ is a supermartingale. Optional sampling and Fatou's lemma imply that $v(x_0) + \delta = M_0 \geq \mathbb{E}[M_\tau] \geq \mathbb{E}[F(x_\tau)]$. Thus $u_{\mathrm{I}} \leq v$. By Lemma 1.1, this completes the proof. □

Now, we prove the first part of Theorem 1.2. When the graph (X, E) is locally finite and Y is finite, this was proven in [15, Thm. 17].

Theorem 2.2. *Suppose that (X, E) is connected, and $Y \neq \emptyset$. If F is bounded below (or above) on Y and $f = 0$, then $u_{\mathrm{I}} = u_{\mathrm{II}}$, so the game has a value.*

Before we prove this, we discuss several counterexamples that occur when the conditions of the theorem are not met. First, this theorem can fail if E is directed. A trivial counterexample is when X is finite and there is no directed path from the initial state to Y.

If X is infinite and E is directed, then there are counterexamples to Theorem 2.2, even when every vertex lies on a directed path towards a terminal state. For example, suppose $X = \mathbb{N}$ and $Y = \{0\}$, with $F(0) = 0$. If E consists of directed edges of the form $(n, n-1)$ and $(n, n+2)$, then II may play so that with positive probability the game never terminates, and hence the value for player I is by definition $-\infty$.

Even in the undirected case, a game may not have a value if F is not bounded either from above or below. The reader may check that if X is the integer lattice $\mathbb{Z}^2$ and the terminal states are the x-axis with $F((x, 0)) = x$, then the players' value functions are given by $u_{\mathrm{I}}((x, y)) = x - |y|$ and $u_{\mathrm{II}}((x, y)) = x + |y|$. (Roughly speaking, this is because, in order to force the game to end in a finite amount of time, player I has to "give up" $|y|$ opportunities to move to the right.) Observe also that in this case any linear function which agrees with F on the x-axis is ∞-harmonic.

As a final remark before proving this theorem, let us consider the obvious strategy for the two players, namely for player I to always maximize u on her turn and for II to always minimize u on her turn. Even when E is (undirected and) locally finite and the payoff function satisfies $0 \leq F \leq 1$, these obvious strategies need not be

optimal. Consider, e.g., the game shown in Figure 1, where $X \subset \mathbb{R}^2$ is given by $X = \{v(k,j) : j = 1,2,\ldots; k = 0,1,\ldots,2^j\}$, $v(k,j) = (2^{-j}k, 1-2^{-j+1})$, and E consists of edges of the form $\{v(k,j), v(k+1,j)\}$ and $\{v(k,j), v(2k,j+1)\}$. The terminal states are on the left and right edges of the square, and the payoff is 1 on the left and 0 on the right. Clearly the function $u(v(k,j)) = 1-k/2^j$ (the Euclidean distance from the right edge of the square) is infinity harmonic, and by Corollary 2.3 below, we have $u_{\mathrm{I}} = u = u_{\mathrm{II}}$. The obvious strategy for player I is to always move left, and for II it is to always move right. Suppose however that player I always pulls left and player II always pulls up. It is easy to check that the probability that the game ever terminates when starting from $v(k,j)$ is at most $2/(k+2)$ (this function is a supermartingale under the corresponding Markov chain). Therefore the game continues forever with positive probability, resulting in a payoff of $-\infty$ to player I. Thus, a near-optimal strategy for player I must be able to force the game to end, and must be prepared to make moves which do not maximize u. (This is a well-known phenomenon, not particular to tug-of-war.)

Proof of Theorem 2.2. If F is bounded above but not below, then we may exchange the roles of players I and II and negate F to reduce to the case where F is bounded from below. Since $u_{\mathrm{I}} \le u_{\mathrm{II}}$ always holds, we just need to show that $u_{\mathrm{II}} \le u_{\mathrm{I}}$. Since player I could always pull towards a point in Y and thereby ensure that the game terminates, we have $u_{\mathrm{I}} \ge \inf_Y F > -\infty$. Let $u = u_{\mathrm{I}}$ and write $\delta(x) = \sup_{y:y\sim x} |u(y) - u(x)|$. Let $x_0, x_1, \ldots$ be the sequence of positions of the game. For ease of exposition, we begin by assuming that E is locally finite (so that the suprema and infima in the definition of the ∞-Laplacian definition are achieved) and that $\delta(x_0) > 0$; later we will remove these assumptions.

To motivate the following argument, we make a few observations. In order to prove that $u_{\mathrm{II}} \le u_{\mathrm{I}}$, we need to show that player II can guarantee that the game terminates while also making sure that the expected payoff is not much larger than $u(x_0)$. These are two different goals, and it is not *a priori* clear how to combine them. To resolve this difficulty, observe that $\delta(x_j)$ is non-decreasing in j if players I and II adopt the strategies of maximizing (respectively, minimizing) u at every step. As we will later see, this implies that the game terminates a.s. under these strategies. On the other hand, if player I deviates from this strategy and thereby reduces $\delta(x_j)$, then perhaps player II can spend some turns playing suboptimally with respect to u in order to increase δ. Let $X_0 := \{x \in X : \delta(x) \ge \delta(x_0)\} \cup Y$. For $n = 0,1,2,\ldots$ let $j_n = \max\{j \le n : x_j \in X_0\}$ and $v_n = x_{j_n}$, which is the last position in X_0 up to time n. We will shortly describe a strategy for II based on the idea of backtracking to X_0 when not in X_0. If $v_n \ne x_n$, we may define a backtracking move from x_n as any move to a neighbor y_n of x_n that is closer to v_n than x_n in the subgraph $G_n \subset (X,E)$ spanned by the vertices $x_{j_n}, x_{j_n+1}, \ldots, x_n$. Here, "closer" refers to the graph metric of G_n. When II plays the backtracking strategy, she backtracks whenever not in X_0 and plays to a neighbor minimizing u_{I} when in X_0.

Now consider the game evolving under any strategy for player I and the backtracking strategy for II. Let d_n be the distance from x_n to v_n in the subgraph G_n. Set

$$m_n := u(v_n) + \delta(x_0)\, d_n\,.$$

It is clear that $u(x_n) \le m_n$, because there is a path of length d_n from x_n to v_n in G_n and the change in u across any edge in this path is less than $\delta(x_0)$, by the definition of X_0. It is easy to verify that m_n is a supermartingale, as follows. If $x_n \in X_0$, and

player I plays, then $m_{n+1} \le u(x_n) + \delta(x_n) = m_n + \delta(x_n)$, while if II gets the turn, then $m_{n+1} = u(x_n) - \delta(x_n)$. If $x_n \notin X_0$ and II plays, then $m_{n+1} = m_n - \delta(x_0)$. If $x_n \notin X_0$ and player I does not play into X_0, then $m_{n+1} \le m_n + \delta(x_0)$. The last case to consider is that $x_n \notin X_0$ and player I plays into a vertex in X_0. In such a situation,

$$m_{n+1} = u(x_{n+1}) \le u(x_n) + \delta(x_n) \le m_n + \delta(x_0)\,.$$

Thus, indeed, m_n is a supermartingale (bounded from below). Let τ denote the first time a terminal state is reached (so $\tau = \infty$ if the game does not terminate). By the martingale convergence theorem, the limit $\lim_{n\to\infty} m_{n\wedge\tau}$ exists. But when player II plays we have $m_{n+1} \le m_n - \delta(x_0)$. Therefore, the game must terminate with probability 1. The expected outcome of the game thus played is at most $m_0 = u(x_0)$. Consequently, $u_{\mathrm{II}} \le u$, which completes the proof in the case where E is locally finite and $\delta(x_0) > 0$.

Next, what if E is not locally finite, so that suprema and infima might not be achieved? In this case, we fix a small $\eta > 0$, and use the same strategy as above, except that if $x_n \in X_0$ and II gets the turn, she moves to a neighbor at which u is at most $\eta 2^{-n-1}$ larger than its infimum value among neighbors of x_n. In this case, $m_n + \eta\, 2^{-n}$ is a supermartingale, and hence the expected payoff is at most $u(x_0) + \eta$. Since this can be done for any $\eta > 0$, we again have that $u_{\mathrm{II}} \le u$.

Finally, suppose that $\delta(x_0) = 0$. Let $y \in Y$, and let player II pull toward y until the first time a vertex x_0^* with $\delta(x_0^*) > 0$ or $x_0^* \in Y$ is reached. After that, II continues as above. Since $u(x_0) = u(x_0^*)$, this completes the proof. □

Corollary 2.3. *If E is connected, $Y \neq \emptyset$ and $\sup|F| < \infty$, then $u = u_{\mathrm{I}} = u_{\mathrm{II}}$ is the unique bounded ∞-harmonic function agreeing with F on Y.*

Proof. This is an immediate consequence of Lemma 2.1, the remark that follows it, and Theorem 2.2. □

If $E = \{(n, n+1) : n = 0, 1, 2, \dots\}$, $Y = \{0\}$ and $F(0) = 0$, then $\tilde{u}(n) = n$ is an example of an (unbounded) ∞-harmonic function that is different from u.

2.2. **Tug-of-war on graphs with running payoffs.** Suppose now that $f \neq 0$. Then the analog of Theorem 2.2 does not hold without additional assumptions. For a simple counterexample, suppose that E is a triangle with self-loops at its vertices (i.e., a player may opt to remain in the same position), that the vertex v_0 is a terminal vertex with final payoff $F(v_0) = 0$, and the running payoff is given by $f(v_1) = -1$ and $f(v_2) = 1$. Then the function given by $u(v_0) = 0$, $u(v_1) = a - 1$, $u(v_2) = a + 1$ is a solution to $\Delta_\infty u = -2f$, provided $-1 \le a \le 1$. The reader may check that u_{I} is the smallest of these functions and u_{II} is the largest. The gap of 2 between u_{I} and u_{II} appears because a player would have to give up a move (sacrificing one) in order to force the game to end. This is analogous to the $\mathbb{Z}^2$ example given in Section 2.1. Both players are earning payoffs in the interior of the game, and moving to a terminal vertex costs a player a turn.

One way around this is to assume that f is either uniformly positive or uniformly negative, as in the following analog of Theorem 2.2. We now prove the second half of Theorem 1.2:

Theorem 2.4. *Suppose that E is connected and $Y \neq \emptyset$. Assume that F is bounded from below and $\inf f > 0$. Then $u_{\mathrm{I}} = u_{\mathrm{II}}$. If, additionally, f and F are bounded*

from above, then any bounded solution $\tilde{u}$ to $\Delta_\infty \tilde{u} = -2f$ with the given boundary conditions is equal to u.

Proof. By considering a strategy for player I that always pulls toward a specific terminal state y, we see that $\inf_X u_{\mathrm{I}} \geq \inf_Y F$. By Lemma 1.1, $\Delta_\infty u_{\mathrm{I}} = -2f$ on $X \setminus Y$. Let $\tilde{u}$ be any solution to $\Delta_\infty \tilde{u} = -2f$ on $X \setminus Y$ that is bounded from below on X and has the given boundary values on Y.

Claim. $u_{\mathrm{II}} \leq \tilde{u}$. In proving this, we may assume without loss of generality that $G \setminus Y$ is connected, where G is the graph (X, E). Then if $\tilde{u} = \infty$ at some vertex in $X \setminus Y$, we also have $\tilde{u} = \infty$ throughout $X \setminus Y$, in which case the Claim is obvious. Thus, assume that $\tilde{u}$ is finite on $X \setminus Y$. Fix $\delta \in (0, \inf f)$ and let II use the strategy that at step k, if the current state is x_{k-1} and II wins the coin toss, selects a state x_k with $\tilde{u}(x_k) < \inf_{z : z \sim x_{k-1}} \tilde{u}(z) + 2^{-k}\delta$. Then for any strategy chosen by player I, the sequence $M_k = \tilde{u}(x_k) + 2^{-k}\delta + \sum_{j=0}^{k-1} f(x_j)$ is a supermartingale bounded from below, which must converge a.s. to a finite limit. Since $\inf f > 0$, this also forces the game to terminate a.s. Let τ denote the termination time. Then

$$\tilde{u}(x_0) + \delta = M_0 \geq \mathbb{E}(M_\tau) \geq \mathbb{E}\Big(\tilde{u}(x_\tau) + \sum_{j=0}^{\tau-1} f(x_j)\Big).$$

Thus this strategy for II shows that $u_{\mathrm{II}}(x_0) \leq \tilde{u}(x_0) + \delta$. Since $\delta > 0$ is arbitrary, this verifies the claim. In particular, $u_{\mathrm{II}} \leq u_I$ in X, so $u_{\mathrm{II}} = u_{\mathrm{I}}$.

Now suppose that $\sup F < \infty$ and $\sup f < \infty$, and $\tilde{u}$ is a *bounded* solution to $\Delta_\infty \tilde{u} = -2f$ with the given boundary values. By the claim above, $\tilde{u} \geq u_{\mathrm{II}} = u_{\mathrm{I}}$. On the other hand, player I can play to maximize or nearly maximize $\tilde{u}$ in every move. Under such a strategy, she guarantees that by turn k the expected payoff is at least $\tilde{u}(x_0) - \mathbb{E}[Q(k)] - \varepsilon$, where $Q(k)$ is 0 if the game has terminated by time k, and $\tilde{u}(x_k)$ otherwise. If the expected number of moves played is infinite, the expected payoff is infinite. Otherwise, $\lim_k \mathbb{E}[Q(k)] = 0$, since $\tilde{u}$ is bounded. Thus, $u_{\mathrm{II}} = u_{\mathrm{I}} \geq \tilde{u}$ in any case. □

3. Continuum value of tug-of-war on a length space

3.1. **Preliminaries and outline.** In this section, we will prove Theorem 1.3, Theorem 1.8 and Theorem 1.4. Throughout this section, we assume (X, d, Y, F, f) denotes a tug-of-war game, i.e., X is a length space with distance function d, $Y \subset X$ is a non-empty set of terminal states, $F : Y \to \mathbb{R}$ is the final payoff function, and $f : X \setminus Y \to \mathbb{R}$ is the running payoff function. We let x_k denote the game state at time k in ε-tug-of-war.

It is natural to ask for a continuous-time version of tug-of-war on a length space. Precisely and rigorously defining such a game (which would presumably involve replacing coin tosses with white noise, making sense of what a continuum no-look-ahead strategy means, etc.) is a technical challenge we will not undertake in this paper (though we include some discussion of the small-ε limiting trajectory of ε-tug-of-war in the finite-dimensional Euclidean case in Section 7). But we can make sense of the continuum game's *value function* u^0 by showing that the value function u^ε for the ε-step tug-of-war game converges as $\varepsilon \to 0$.

The value functions u^ε do not satisfy any nice monotonicity properties as $\varepsilon \to 0$. In the next subsection we define two modified versions of tug-of-war whose

values closely approximate u^ε, and which do satisfy a monotonicity property along sequences of the form $\varepsilon 2^{-n}$, allowing us to conclude that $\lim_n u^{\varepsilon 2^{-n}}$ exists. Then we show that any such limit is a bounded from below viscosity solution to $\Delta_\infty u = -2f$, and that any viscosity solution bounded from below is an upper bound on such a limit, so that any two such limits must be equal, which will allow us to prove that the continuum limit $u^0 = \lim_\varepsilon u^\varepsilon$ exists.

Because the players can move the game state almost as far as ε, either player can ensure that $d(x_k, y)$ is "almost a supermartingale" up until the time that $x_k = y$. When doing calculations it is more convenient to instead work with a related metric d^ε defined by

$$\begin{aligned} d^\varepsilon(x,y) &:= \varepsilon \times (\min \#\text{ steps from } x \text{ to } y \text{ using steps of length} < \varepsilon) \\ &= \begin{cases} 0, & x = y, \\ \varepsilon + \varepsilon \lfloor d(x,y)/\varepsilon \rfloor, & x \neq y. \end{cases} \end{aligned}$$

Since d^ε is the graph distance scaled by ε, it is in fact a metric, and either player may choose to make $d^\varepsilon(x_k, y)$ a supermartingale up until the time $x_k = y$.

3.2. **II-favored tug-of-war and dyadic limits.** We define a game called **II-favored ε-tug-of-war** that is designed to give a lower bound on player I's expected payoff. It is related to ordinary ε-step tug-of-war, but II is given additional options, and player I's running payoffs are slightly smaller. At the $(i+1)^{\text{st}}$ step, player I chooses a point z in $B_\varepsilon(x_i)$ and a coin is tossed. If player I wins the coin toss, the game position moves to a point, of player II's choice, in $(B_{2\varepsilon}(z) \cap Y) \cup \{z\}$. (If $d(z, Y) \geq 2\varepsilon$, this means simply moving to z.) If II wins, then the game position moves to a point in $B_{2\varepsilon}(z)$ of II's choice. The game ends at the first time τ for which $x_\tau \in Y$. Player I's payoff is then $-\infty$ if the game never terminates, and otherwise it is

$$\varepsilon^2 \sum_{i=1}^{\tau} \inf_{y \in B_{2\varepsilon}(z_i)} f(y) + F(x_\tau), \tag{3.1}$$

where z_i is the point that player I targets on the ith turn, and $f(y)$ is defined to be zero if $y \in Y$. We let v^ε be the value for player I for this game. Given a strategy for player II in the ordinary ε-game, player II can easily mimic this strategy in the II-favored ε-game and do at least as well, so $v^\varepsilon \leq u_{\text{I}}^\varepsilon$.

Let w^ε be the value for player II of I-favored ε-tug-of-war, defined analogously but with the roles of player I and player II reversed (i.e., at each move, II selects the target less than ε units away, instead of player I, etc., and the inf in the running payoff term in equation (3.1) is replaced with a sup, and games that never terminate have payoff $+\infty$). For any $\varepsilon > 0$ we have

$$v^\varepsilon \leq u_{\text{I}}^\varepsilon \leq u_{\text{II}}^\varepsilon \leq w^\varepsilon.$$

Lemma 3.1. *For any $\varepsilon > 0$,*

$$v^{2\varepsilon} \leq v^\varepsilon \leq u_{\text{I}}^\varepsilon \leq u_{\text{II}}^\varepsilon \leq w^\varepsilon \leq w^{2\varepsilon}.$$

Proof. We have already noted that $v^\varepsilon \leq u_{\text{I}}^\varepsilon \leq u_{\text{II}}^\varepsilon \leq w^\varepsilon$. We will prove that $v^{2\varepsilon} \leq v^\varepsilon$; the inequality $w^\varepsilon \leq w^{2\varepsilon}$ follows by symmetry. Consider a strategy $S_{\text{I}}^{2\varepsilon}$ for a II-favored 2ε-tug-of-war. We define a strategy S_{I}^ε for player I for the II-favored ε-game that mimics $S_{\text{I}}^{2\varepsilon}$ as follows. Whenever strategy $S_{\text{I}}^{2\varepsilon}$ would choose a target point z, player I "aims" for z for one "round," which we define to be the time until

one of the players has won the coin toss two more times than the other player. By "aiming for z" we mean that player I picks a target point that, in the metric d^ε, is ε units closer to z than the current point. With probability 1/2, player I gets two surplus moves before II, and then the game position reaches z (or a point in $Y \cap B_{4\varepsilon}(z)$) before the game position exits $B_{4\varepsilon}(z)$. If player II gets two surplus moves before player I, then the game position will be in $B_{4\varepsilon}(z)$. (See Figure 2.) The expected number of moves in this round of the II-favored ε-game is 4; and the

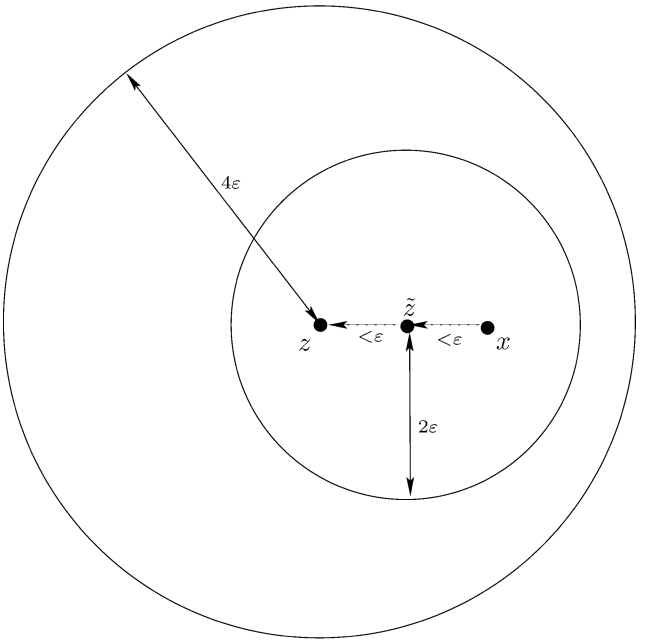


FIGURE 2. At the end of the round, player I reaches the target z (or a point in $Y \cap B_{4\varepsilon}(z)$) with probability at least 1/2, and otherwise the state still remains within $B_{4\varepsilon}(z)$. During the first step of the round, when player I has target $\tilde{z}$, the running payoff is the infimum of f over $B_{2\varepsilon}(\tilde{z}) \subset B_{4\varepsilon}(z)$, and during any step of the round the infimum is over a subset of $B_{4\varepsilon}(z)$.

running payoff at each move is an ε^2 times the infimum over a ball of radius 2ε that is a subset of $B_{4\varepsilon}(z)$ (as opposed to $(2\varepsilon)^2$ times the infimum over the whole ball). Hence strategy $S_{\mathrm{I}}^{\varepsilon}$ guarantees for the II-favored ε-game an expected total payoff that is at least as large as what $S_{\mathrm{I}}^{2\varepsilon}$ guarantees for the II-favored 2ε-game. □

Thus v^ε converges along dyadic sequences: $v^{\varepsilon/2^\infty} := \lim_{n\to\infty} v^{\varepsilon 2^{-n}}$ exists. *A priori* the subsequential limit could depend upon the choice of the dyadic sequence, i.e., the initial ε.

The same argument can be used to show that $v^{k\varepsilon} \le v^\varepsilon$ for positive integers k.

3.3. **Comparing favored and ordinary tug-of-war.** We continue with a preliminary bound on how far apart v^ε and w^ε can be. Let $\mathrm{Lip}_Y^\varepsilon F$ denote the Lipschitz constant of F with respect to the restriction of the metric d^ε to Y. Since $d \le d^\varepsilon$, the Lipschitz constant $\mathrm{Lip}_Y F$ of F with respect to d upper bounds $\mathrm{Lip}_Y^\varepsilon F$.

Lemma 3.2. *Let $\varepsilon > 0$. Suppose that $\operatorname{Lip}_Y^\varepsilon F < \infty$ and either*

(1) *$f = 0$ everywhere, or*
(2) *$|f|$ is bounded above and X has finite diameter.*

Then for each $x \in X$ and $y \in Y$,

$$v^\varepsilon(x) \geq F(y) - 2\,\varepsilon \operatorname{Lip}_Y^\varepsilon F - \left(\operatorname{Lip}_Y^\varepsilon F + 2\,(\varepsilon + \operatorname{diam} X) \sup |f|\right) d^\varepsilon(x,y).$$

Such an expected payoff is guaranteed for player I *if she adopts a pull towards y strategy, which at each move attempts to reduce $d^\varepsilon(x_t, y)$. Similarly a pull towards y strategy for* II *gives*

$$w^\varepsilon(x) \leq F(y) + 2\,\varepsilon \operatorname{Lip}_Y^\varepsilon F + \left(\operatorname{Lip}_Y^\varepsilon F + 2\,(\varepsilon + \operatorname{diam} X) \sup |f|\right) d^\varepsilon(x,y).$$

Proof. Let player I use a pull towards y strategy. Let τ be the time at which Y is reached, which will be finite a.s. The distance $d^\varepsilon(x_k, y)$ is a supermartingale, except possibly at the last step, where player II may have moved the game state to a terminal point up to a distance of 2ε from the target, even if player I wins the coin toss. Thus $\mathbb{E}[d^\varepsilon(x_\tau, y)] < d^\varepsilon(x,y) + 2\varepsilon$, whence

$$\mathbb{E}[F(x_\tau)] \geq F(y) - \left(d^\varepsilon(x,y) + 2\varepsilon\right) \operatorname{Lip}_Y^\varepsilon F.$$

If $f = 0$, then this implies $v^\varepsilon(x) \geq F(y) - (d^\varepsilon(x,y) + 2\varepsilon)\operatorname{Lip}_Y^\varepsilon F$. If $f \neq 0$ and X has finite diameter, then the expected number of steps before the game terminates is at most the expected time that a simple random walk on the interval of integers $[0, 1 + \lfloor \operatorname{diam}(X)/\varepsilon \rfloor]$ (with a self-loop added at the right endpoint) takes to reach 0 when started at $j := d^\varepsilon(x,y)/\varepsilon$, i.e., at most $j\left(3 + 2\lfloor \operatorname{diam}(X)/\varepsilon \rfloor - j\right)$. Hence

$$\begin{aligned}
&\mathbb{E}\Big[F(x_\tau) + \sum_{i=0}^{\tau-1} \varepsilon^2 f(x_i)\Big] \\
&\quad \geq F(y) - \left(d^\varepsilon(x,y) + 2\,\varepsilon\right) \operatorname{Lip}_Y^\varepsilon F + j\left(3 + 2\lfloor \operatorname{diam}(X)/\varepsilon \rfloor - j\right) \varepsilon^2 \min(0, \inf f) \\
&\quad \geq F(y) - \left(d^\varepsilon(x,y) + 2\,\varepsilon\right) \operatorname{Lip}_Y^\varepsilon F - d^\varepsilon(x,y)\left(2\,\varepsilon + 2\operatorname{diam}(X)\right) \sup |f|.
\end{aligned}$$

This gives the desired lower bound on $v^\varepsilon(x)$. The symmetric argument gives the upper bound for $w^\varepsilon(x)$. □

Next, we show that the lower bound v^ε on $u_{\mathrm{I}}^\varepsilon$ is a good lower bound.

Lemma 3.3. *Suppose F is Lipschitz, and either*

(1) *$f = 0$ everywhere, or*
(2) *f is uniformly continuous, X has finite diameter and $\inf |f| > 0$.*

Then $\|u_{\mathrm{I}}^\varepsilon - v^\varepsilon\|_\infty \to 0$ as $\varepsilon \to 0$. If f is also Lipschitz, then $\|u_{\mathrm{I}}^\varepsilon - v^\varepsilon\|_\infty = O(\varepsilon)$.

Note that since X is assumed to be a length space, the assumptions imply that $\operatorname{sign}(f)$ is constant and $\sup |f| < \infty$.

Proof. In order to prove that $u_{\mathrm{I}}^\varepsilon$ is not much larger than v^ε, consider a strategy S_{I} for player I in ordinary ε-tug-of-war, which achieves an expected payoff of at least $u_{\mathrm{I}}^\varepsilon - \varepsilon$ against any strategy for player II. We shortly describe a modified strategy $S_{\mathrm{I}}^{\mathfrak{F}}$ for player I playing the II-favored game, which does almost as well as S_{I} does in the ordinary game. To motivate $S_{\mathrm{I}}^{\mathfrak{F}}$, observe that a turn in the II-favored ε-tug-of-war can alternatively be described as follows. Suppose that the position at the end of the previous turn is x. First, player I gets to make a move to an arbitrary point z satisfying $d(x,z) < \varepsilon$. Then a coin is tossed. If player II wins the toss, she gets

to make two steps from z, each of distance less than ε. Otherwise, II gets to move to an arbitrary point in $B_{2\varepsilon}(z) \cap Y$, but only if the latter set is non-empty. This completes the turn. The strategy $S_{\mathrm{I}}^{\mathfrak{F}}$ is based on the idea that after a win in the coin toss by II, player I may use her move to reverse one of the two steps executed by II (provided Y has not been reached).

As strategy $S_{\mathrm{I}}^{\mathfrak{F}}$ is playing the II-favored game against player II, it keeps track of a virtual ordinary game. At the outset, the II-favored game is in state $x_0^{\mathfrak{F}}$, as is the virtual game. As long as the virtual and the favored game have not ended, each turn in the favored game corresponds to a turn in the virtual game, and the virtual game uses the same coin tosses as the favored game. In each such turn t, the target $z_t^{\mathfrak{F}}$ for player I in the favored game is the current state x_t in the ordinary game. If player I wins the coin toss, then the new game state $x_{t+1}^{\mathfrak{F}}$ in the favored game is his current target $z_t^{\mathfrak{F}}$, which is the state of the virtual game x_t, and the new state of the virtual game x_{t+1} is chosen according to strategy S_{I} applied to the history of the virtual game. If player II wins the toss and chooses the new state of the favored game to be $x_{t+1}^{\mathfrak{F}}$, where necessarily $d(x_{t+1}^{\mathfrak{F}}, x_t) < 2\,\varepsilon$, then in the virtual game the virtual player II chooses the new state x_{t+1} as some point satisfying $d(x_{t+1}, x_t) < \varepsilon$ and $d(x_{t+1}, x_{t+1}^{\mathfrak{F}}) < \varepsilon$. Induction shows that $d(x_t^{\mathfrak{F}}, x_t) < \varepsilon$ as long as both games are running, and thus the described moves are all legal.

If at some time the virtual game has terminated, but the favored game has not, we let player I continue playing the favored game by always pulling towards the final state of the virtual game. If the favored game has terminated, for the sake of comparison, we continue the virtual game, but this time let player II pull towards the final state of the favored game and let player I continue using strategy S_{I}.

Let $\tau^{\mathfrak{F}}$ be the time at which the favored game has ended, and let τ be the time at which the ordinary virtual game has ended. By Lemma 3.2, if $\tau < \tau^{\mathfrak{F}}$, then the conditioned expectation of the remaining running payoffs and final payoff to player I in the favored game after time τ, given what happened up to time τ, is at least $F(x_\tau) - O(\varepsilon)$ (here the implicit constant may depend on $\mathrm{diam}X$, $\mathrm{Lip}_Y F$ and $\sup|f|$). Likewise, $u_I^\varepsilon \le w^\varepsilon$ from Lemma 3.1 and the second inequality from Lemma 3.2 show that if $\tau^{\mathfrak{F}} < \tau$, then the conditioned expectation of the remaining running and final payoffs in the virtual game is at most $F(x_{\tau^{\mathfrak{F}}}^{\mathfrak{F}}) + O(\varepsilon)$.

Set $\lambda = \lambda_\varepsilon := \sup\bigl\{|f(x) - f(x')| : x, x' \in X \setminus Y, d(x, x') < 2\,\varepsilon\bigr\}$. Then $\lambda = O(\varepsilon)$ if f is Lipschitz and $\lim_{\varepsilon\to 0}\lambda_\varepsilon = 0$ if f is uniformly continuous. At each time $t < \tau \wedge \tau^{\mathfrak{F}}$, since $z_t^{\mathfrak{F}} = x_t$, the running payoff in the virtual game and in the favored game differ by at most $\varepsilon^2\,\lambda$. Lemmas 3.1 and 3.2 show that there is a constant C, which may depend on X, Y, f and F, but not on ε, such that $-C \le v^\varepsilon \le u_{\mathrm{I}}^\varepsilon \le C$. Thus, in the case where $\sup f < 0$, since S_I guarantees a payoff of at least $u_{\mathrm{I}}^\varepsilon(x_0) - \varepsilon$, we have $\mathbb{E}[\tau^{\mathfrak{F}} \wedge \tau] = O(\varepsilon^{-2})$. Assume that player II plays the II-favored game (up to time $\tau \wedge \tau^{\mathfrak{F}}$) using a strategy $S_{\mathrm{II}}^{\mathfrak{F}}$ such that the expected payoff to player I who uses $S_{\mathrm{I}}^{\mathfrak{F}}$ is at most $v_\varepsilon + \varepsilon$. Then, in the case where $\inf f > 0$, we will have $\mathbb{E}[\tau \wedge \tau^{\mathfrak{F}}] = O(\varepsilon^{-2})$, again. There is a strategy S_{II} for player II in the ordinary game, which corresponds to the play of player II in the virtual game, when player I uses S_I and $S_I^{\mathfrak{F}}$ and player II uses $S_{\mathrm{II}}^{\mathfrak{F}}$ in the favored game. (The description of the virtual game defines S_{II} for some game histories, and we may take an arbitrary extension of this partial strategy to all possible game histories.) The above shows that the expected payoff for player I in the ordinary game when player I uses S_{I} and player II uses S_{II} differs from the expected payoff for player I in the favored game when

player I uses $S_{\mathrm{I}}^{\mathfrak{F}}$ and player II uses $S_{\mathrm{II}}^{\mathfrak{F}}$ by at most $O(\varepsilon)+\lambda\,\varepsilon^2\,\mathbb{E}[\tau\wedge\tau^{\mathfrak{F}}]\le O(\varepsilon+\lambda)$. Thus $u_{\mathrm{I}}^{\varepsilon}\le v^{\varepsilon}+O(\varepsilon+\lambda)$. Since $u_{\mathrm{I}}^{\varepsilon}\ge v^{\varepsilon}$, the proof is now complete. □

Hence under the assumption of Lemma 3.3, $\lim_{n\to\infty} u_{\mathrm{I}}^{\varepsilon 2^{-n}} = v^{\varepsilon/2^{\infty}}$.

3.4. **Dyadic limits satisfy quadratic comparison.** We start by showing that $u_{\mathrm{I}}^{\varepsilon}$ almost satisfies $(-2f)$-quadratic comparison from above.

Lemma 3.4. *Let $\varepsilon>0$, let V be an open subset of $X\setminus Y$ and write $V_\varepsilon=\{x:\overline{B_\varepsilon(x)}\subset V\}$. Suppose that $\varphi(x)=Q(d(x,z))$ is a quadratic distance function that is $\star$-increasing on V, where $Q(r)=ar^2+br+c$ satisfies*

$$a\le -\sup_{x\in V_\varepsilon} f(x). \tag{3.2}$$

Also suppose that $\sup_{V_\varepsilon} f\ge 0$ or $\operatorname{diam}(V)<\infty$. If the value function $u_{\mathrm{I}}^{\varepsilon}$ for player I *in ε-tug-of-war satisfies $u_{\mathrm{I}}^{\varepsilon}\le\varphi$ on $V\setminus V_\varepsilon$, then $u_{\mathrm{I}}^{\varepsilon}\le\varphi$ on V_ε.*

Proof. Fix some $\delta>0$. Consider the strategy for player II that from a state $x_{k-1}\in V_\varepsilon$ at distance $r=d(x_{k-1},z)$ from z pulls to state z (if $r<\varepsilon$) or else moves to reduce the distance to z by "almost" ε units, enough to ensure that $Q(d(x_k,z))<Q(r-\varepsilon)+\delta 2^{-k}$. If $r<\varepsilon$, then $z\in V$, whence $Q(t)=at^2+c$ with $a\ge 0$. In this case, if II wins the toss, then $x_k=z$, whence $\varphi(x_k)=Q(0)\le Q(r-\varepsilon)$. Thus for all $r\ge 0$, regardless of what strategy player I adopts,

$$\mathbb{E}[\varphi(x_k)\mid x_{k-1}]-\delta 2^{-k-1}\le\frac{Q(r+\varepsilon)+Q(r-\varepsilon)}{2}=Q(r)+a\,\varepsilon^2=\varphi(x_{k-1})+a\,\varepsilon^2\,.$$

Setting $\tau:=\inf\{k:x_k\notin V_\varepsilon\}$, we conclude that $M_k:=\varphi(x_{k\wedge\tau})-a\,\varepsilon^2(k\wedge\tau)+\delta\,2^{-k}$ is a supermartingale.

Suppose that player I uses a strategy with expected payoff larger than $-\infty$. (If there is no such strategy, the assertion of the lemma is obvious.) Then $\tau<\infty$ a.s. We claim that

$$\mathbb{E}[M_\tau]\le M_0\,. \tag{3.3}$$

Clearly this holds if τ is replaced by $\tau\wedge k$. To pass to the limit as $k\to\infty$, consider two cases:

- If $a\le 0$, then since φ is $\star$-increasing on V, it is also bounded from below on V. Consequently, M_k is a supermartingale bounded from below, so (3.3) holds.
- If $a>0$, then $\sup_{V_\varepsilon} f<0$, by (3.2). By assumption therefore $\operatorname{diam} V<\infty$, which implies $\sup_V|\varphi|<\infty$. If $\mathbb{E}[\tau]=\infty$, we get $\mathbb{E}[M_\tau]=-\infty$, and hence (3.3) holds. On the other hand, if $\mathbb{E}[\tau]<\infty$, then dominated convergence gives (3.3).

Since $u_{\mathrm{I}}^{\varepsilon}(x_\tau)\le\varphi(x_\tau)$, we deduce that

$$u_{\mathrm{I}}^{\varepsilon}(x_0)\le\sup_{S_{\mathrm{I}}}\mathbb{E}\Big[\varphi(x_\tau)+\sum_{t=0}^{\tau-1}f(x_t)\Big]\overset{(3.2)}{\le}\sup_{S_{\mathrm{I}}}\mathbb{E}[M_\tau]\le M_0=\varphi(x_0)+\delta\,,$$

where S_{I} runs over all possible strategies for player I with expected payoff larger than $-\infty$. Since $\delta>0$ was arbitrary, the proof is now complete. □

In order for $u_{\mathrm{I}}^{\varepsilon}$ to satisfy $(-2f)$-quadratic comparison (from above), we would like to know that if $u_{\mathrm{I}}^{\varepsilon} \leq \varphi$ on the boundary of an open set, then this (almost) holds in a neighborhood of the boundary, so that we can apply the above lemma. To do this we prove a uniform Lipschitz lemma:

Lemma 3.5 (Uniform Lipschitz)**.** *Suppose that F is Lipschitz, and either*

(1) *$f = 0$ everywhere, or*
(2) *$|f|$ is bounded from above and X has finite diameter.*

Then for each $\varepsilon \in (0, \mathrm{diam}X)$, $u_{\mathrm{I}}^{\varepsilon}$ and $u_{\mathrm{II}}^{\varepsilon}$ are Lipschitz on X w.r.t. the metric d^{ε} with the Lipschitz constant depending only on $\mathrm{diam}X, \sup|f|$ and $\mathrm{Lip}_Y F$.

Proof. By symmetry, it suffices to prove this for $u_{\mathrm{I}}^{\varepsilon}$. Set

$$L := 3\,\mathrm{Lip}_Y^{\varepsilon} F + 4\,\mathrm{diam}X \sup|f|.$$

Let $x, y \in X$ be distinct. If $x, y \in Y$, then $\left|u_{\mathrm{I}}^{\varepsilon}(x) - u_{\mathrm{I}}^{\varepsilon}(y)\right| = \left|F(x) - F(y)\right| \leq \mathrm{Lip}_Y^{\varepsilon} F\, d^{\varepsilon}(x, y)$. If $x \in X \setminus Y$ and $y \in Y$, Lemmas 3.2 and 3.1 give

$$\left|u_{\mathrm{I}}^{\varepsilon}(x) - u_{\mathrm{I}}^{\varepsilon}(y)\right| = \left|u_{\mathrm{I}}^{\varepsilon}(x) - F(y)\right| \leq L\, d^{\varepsilon}(x, y)\,, \qquad x \in X \setminus Y, \quad y \in Y\,. \tag{3.4}$$

Now suppose $x, y \in X \setminus Y$. Set $Y^* = Y \cup \{y\}$, $F^* = F$ on Y and $F^*(y) = u_{\mathrm{I}}^{\varepsilon}(y)$. Then, clearly, the value of $u_{\mathrm{I}}^{\varepsilon}(x)$ for the game where Y is replaced by Y^* and F is replaced by F^* is the same as for the original game. By (3.4), we have $\mathrm{Lip}_{Y^*}^{\varepsilon}(F^*) \leq L$. Consequently, (3.4) gives

$$\begin{aligned}\left|u_{\mathrm{I}}^{\varepsilon}(x) - u_{\mathrm{I}}^{\varepsilon}(y)\right| &= \left|u_{\mathrm{I}}^{\varepsilon}(x) - F^*(y)\right| \leq \left(3\,L + 2\,(\varepsilon + \mathrm{diam}X)\,\sup|f|\right) d^{\varepsilon}(x, y) \\ &\leq 4\,L\, d^{\varepsilon}(x, y)\,,\end{aligned}$$

which completes the proof. □

Lemma 3.6. *Suppose F is Lipschitz and $\inf F > -\infty$, and either (1) $f = 0$ identically or else (2) $\inf f > 0$, f is uniformly continuous, and X has finite diameter. Then the subsequential limit $v^{\varepsilon/2^{\infty}} = \lim_{n\to\infty} v^{\varepsilon 2^{-n}}$ satisfies $(-2f)$-quadratic comparison on $X \setminus Y$.*

Proof. By Theorems 2.2 and 2.4, $u_{\mathrm{I}}^{\varepsilon} = u_{\mathrm{II}}^{\varepsilon}$, so from Lemma 3.3 we have $\|w^{\varepsilon} - v^{\varepsilon}\|_{\infty} \to 0$ as $\varepsilon \to 0$. Thus $\lim_{n\to\infty} w^{\varepsilon 2^{-n}} = v^{\varepsilon/2^{\infty}}$.

Note that the hypotheses imply that $|f|$ is bounded. Consider an open $V \subset X \setminus Y$ and an $\star$-increasing quadratic distance function φ on V with quadratic term $a \leq -\sup_{y\in V} f(y)$, such that $\varphi \geq v^{\varepsilon/2^{\infty}}$ on ∂V. We must show that $\varphi \geq v^{\varepsilon/2^{\infty}}$ on V. Since $\|w^{\varepsilon} - v^{\varepsilon}\|_{\infty} \to 0$, we have by Lemma 3.1 that $v^{\varepsilon 2^{-n}}$ converges uniformly to $v^{\varepsilon/2^{\infty}}$. So for any $\delta > 0$, if $n \in \mathbb{N}$ is large enough, then $v^{\varepsilon 2^{-n}} \leq \varphi + \delta$ on ∂V. Note also that φ is necessarily uniformly continuous on $\overline{V}$. (If $\mathrm{diam}V < \infty$ or $a = 0$, this is clear. Otherwise, $f = 0$ and $a < 0$. However, $a < 0$ implies that $\mathrm{diam}V < \infty$, since φ is $\star$-increasing on V.) Hence, by the uniform Lipschitz lemma, Lemma 3.5, $u_{\mathrm{I}}^{\varepsilon 2^{-n}} \leq \varphi + 2\,\delta$ on $V \setminus V_{\varepsilon 2^{-n}}$ for all sufficiently large $n \in \mathbb{N}$, where we use the notation of Lemma 3.4. By that lemma $u_{\mathrm{I}}^{\varepsilon 2^{-n}} \leq \varphi + 2\,\delta$ on all of V. Letting $n \to \infty$ and $\delta \to 0$ shows that $v^{\varepsilon/2^{\infty}}$ satisfies $(-2f)$-quadratic comparison from above.

To prove quadratic comparison from below, note that the only assumptions which are not symmetric under exchanging the roles of the players are $\inf F > -\infty$ and $\inf f > 0$. However, we only used these assumptions to prove $w^{\varepsilon/2^{\infty}} = v^{\varepsilon/2^{\infty}}$. Consequently, comparison from below follows by symmetry. □

3.5. **Convergence.**

Lemma 3.7. *Suppose that v is continuous and satisfies $(-2f)$-quadratic comparison from below on $X \setminus Y$, and that f is locally bounded from below. Let $\delta > 0$. Then in* II*-favored ε-tug-of-war, when player* I *(using any strategy) targets point z_i on step i, player* II *may play to make $M_{t\wedge\tau_\varepsilon}$ a supermartingale, where*

$$M_t := v(x_t) + \varepsilon^2 \sum_{i=1}^{t} \inf_{y\in B_{2\varepsilon}(z_i)} f(y) + \delta 2^{-t}$$

and $\tau_\varepsilon := \inf\{t : d(x_t, Y) < 3\,\varepsilon\}$.

Proof. Let $z = z_t$ be the point that player I has targeted at time t. Assume that $t < \tau_\varepsilon$. We define the following:

(1) $\alpha := \inf\{f(x) : x \in B_{2\varepsilon}(z)\}$,
(2) $A := \inf\{v(x) : x \in B_{2\varepsilon}(z)\}$ (the infimum value of v that II can guarantee if II wins the coin toss),
(3) $\beta := \frac{v(z)+A}{2} + \alpha\,\varepsilon^2$, and
(4) $Q(r) := -\alpha r^2 + \frac{A-v(z)+4\alpha\varepsilon^2}{2\varepsilon} r + v(z)$ $(Q(0) = v(z)$, $Q(2\varepsilon) = A$, $Q(\varepsilon) = \beta$, and $Q'' = -2\alpha)$.

Player II can play so that $\mathbb{E}\big[v(x_t) \;\big|\; z_t \text{ and all prior events}\big] \le (v(z) + A)/2 + \delta 2^{-t}$, i.e., so that $\mathbb{E}\big[M_t \;\big|\; z_t \text{ and prior events}\big] - M_{t-1} \le \beta - v(x_{t-1})$. We will show that whenever $d(x, z) < \varepsilon$ we have $v(x) \ge \beta$, and then it will follow that M is a supermartingale. There are two cases to check, depending on whether $\alpha > 0$ or $\alpha \le 0$:

Suppose $\alpha \le 0$. Note that $A \le v(z)$. If $v(z) = A$, the inequality $v(x) \ge \beta$ on $B_\varepsilon(z)$ follows from $\alpha \le 0$ and the definition of β. Assume therefore that $v(z) > A$. Then $Q'(\varepsilon) = (A - v(z))/(2\varepsilon) < 0$. Thus Q is decreasing on $[0, \varepsilon]$. Let $r_0 \in [\varepsilon, 2\,\varepsilon]$ be the point where Q attains its minimum in $[\varepsilon, 2\,\varepsilon]$. Then Q is decreasing on $[0, r_0]$. Set $V := B_{r_0}(z) \setminus \{z\}$. We have $v(z) = Q(0) = Q\big(d(z, z)\big)$, and for $x \in \partial B_{r_0}(z)$ we have $v(x) \ge A = Q(2\,\varepsilon) \ge Q(r_0) = Q\big(d(x, z)\big)$. Thus, $v(x) \ge Q\big(d(x, z)\big)$ for $x \in \partial V$. Since v satisfies $(-2f)$-quadratic comparison from below, and $Q'' = -2\,\alpha \ge \sup_{x\in V} -2\,f(x)$, we get $v(x) \ge Q\big(d(x, z)\big)$ for $x \in V$. In particular, for $x \in B_\varepsilon(z)$ one has $v(x) \ge Q\big(d(x, z)\big) \ge Q(\varepsilon) = \beta$.

Now suppose $\alpha > 0$. The function $L_0(x) = -\alpha\, d(x, z)^2 + \alpha(2\,\varepsilon)^2 + A$ is a lower bound for v on $\partial B_{2\varepsilon}(z)$, and hence applying $(-2f)$-quadratic comparison in $B_{2\varepsilon}(z)$ with $L_0(\cdot)$ gives $v(z) \ge L_0(z) = A + 4\,\alpha\,\varepsilon^2$. Therefore, $Q'(0) = (A - v(z) + 4\alpha\varepsilon^2)/(2\varepsilon) \le 0$, which together with $Q'' < 0$ implies that Q is decreasing on $[0, 2\varepsilon]$. By applying $(-2f)$-quadratic comparison on $B_{2\varepsilon}(z) \setminus \{z\}$ we see that for $x \in B_\varepsilon(z)$ we have $v(x) \ge Q(d(x, z)) \ge Q(\varepsilon) = \beta$. □

Lemma 3.8. *Suppose that F is Lipschitz, and either*

(1) *$f = 0$ everywhere, or*
(2) *$|f|$ is bounded from above and X has finite diameter.*

Also suppose that $v : X \to \mathbb{R}$ is continuous, satisfies $(-2f)$-quadratic comparison from below on $X \setminus Y$, $v \ge F$ on Y, and $\inf v > -\infty$. Then $v^\varepsilon \le v$ for all ε.

Proof. The idea is for player II to make the M defined in Lemma 3.7 a supermartingale, but we need to pick a stopping time τ such that $\mathbb{E}[M_\tau] \le M_0$, while $v^\varepsilon(x_\tau)$ is unlikely to be much larger than $v(x_\tau)$. Let $W := \{x \in X : d(x, Y) \ge 3\,\varepsilon\}$

and let τ_ε be defined as in Lemma 3.7; that is, $\tau_\varepsilon := \inf\{t \in \mathbb{N} : x_t \notin W\}$. Set $\lambda_\varepsilon := \sup_{X \setminus W}(v^\varepsilon - v)$, and let $\delta > 0$. We first show

$$v^\varepsilon \le v + \lambda_\varepsilon + \delta\,. \tag{3.5}$$

In the case that $f = 0$, the supermartingale M is bounded from below, so we can choose $\tau = \tau_\varepsilon$. Player I is compelled to ensure $\tau < \infty$. Conditional on the game up to time τ, player I cannot guarantee a conditional expected payoff better than $v^\varepsilon(x_\tau) + \delta\, 2^{-\tau} \le M_\tau + \lambda_\varepsilon$. Since $\mathbb{E}[M_\tau] \le M_0 = v(x_0) + \delta$, we get (3.5), as desired.

In the case $f \not\equiv 0$, we let $\tau_n := \tau_\varepsilon \wedge n$, where $n \in \mathbb{N}$. Then $\mathbb{E}[M_{\tau_n}] \le M_0$. Note that $\sup v^\varepsilon < \infty$ follows from $\operatorname{diam} X < \infty$, $v^\varepsilon \le u_{\mathrm{I}}^\varepsilon$ and Lemma 3.5. Suppose player II makes M a supermartingale up until time τ_n. Given play until time τ_n, player II may make sure that the conditional expected payoff to player I is at most

$$\delta\, 2^{-\tau_n} + v^\varepsilon(x_{\tau_n}) + \varepsilon^2 \sum_{i=1}^{\tau_n} \inf_{y \in B_{2\varepsilon}(z_i)} f(y)\,.$$

Taking expectation and separating into cases in which $x_{\tau_n} \in W$ or not, we get

$$\begin{aligned} v^\varepsilon(x_0) \le \mathbb{E}\Big[v(x_{\tau_n}) + \varepsilon^2 \sum_{i=1}^{\tau_n} \inf_{y \in B_{2\varepsilon}(z_i)} f(y)\Big] + \lambda_\varepsilon + \delta\, 2^{-\tau_n} \\ + \Pr[x_{\tau_n} \in W] \sup_W \big(v^\varepsilon(x) - v(x)\big)\,. \end{aligned}$$

Since $\mathbb{E}[M_{\tau_n}] \le M_0 = v(x_0) + \delta$, this gives

$$v^\varepsilon(x_0) \le v(x_0) + \delta + \lambda_\varepsilon + \Pr[x_{\tau_n} \in W] \sup_W \big(v^\varepsilon(x) - v(x)\big)\,.$$

The first term above is $< \mathbb{E}[M_{\tau_n}] \le M_0 = v(x_0) + \delta$, independent of n. Player I is compelled to play a strategy that ensures τ_∞ is finite a.s., since otherwise the payoff is $-\infty < v(x)$. With such a strategy, $\Pr[x_{\tau_n} \in W] \to 0$ as $n \to \infty$, and since v is bounded from below and $\sup v^\varepsilon < \infty$, the last summand tends to 0 as $n \to \infty$. Thus, we get (3.5) in this case as well.

Next, we show that $\limsup_{\varepsilon \searrow 0} \lambda_\varepsilon \le 0$. Let $y \in Y$, and set $Q(r) = a\, r^2 + b\, r + c$, where $a := \sup|f|$, $b < b^* := -2 \sup|f| \operatorname{diam} X - \operatorname{Lip} F$, and $c := F(y)$. Let $\varphi(x) := Q\big(d(x,y)\big)$. Then $\varphi(y') \le F(y') \le v(y')$ for $y' \in Y$. Since v satisfies comparison from below and φ is $\star$-decreasing, we get $v(x) \ge \varphi(x)$ on X. Thus, if $x \in \overline{B_{3\varepsilon}(y)}$, then $v(x) \ge F(y) + b^*\, d(x,y) \ge F(y) + 3\, b^*\, \varepsilon$. In conjunction with Lemma 3.5, this implies that $\limsup_{\varepsilon \searrow 0} \lambda_\varepsilon \le 0$. Choosing $\delta = \varepsilon$ and taking ε to 0 therefore gives in (3.5) $\limsup_{\varepsilon \searrow 0} v^\varepsilon \le v$. However, the inequality $v^{2\varepsilon} \le v^\varepsilon$ from Lemma 3.1 implies $v^\varepsilon \le \limsup_{\varepsilon' \searrow 0} v^{\varepsilon'} \le v$, completing the proof. □

Proof of Theorem 1.3. First, suppose that F is Lipschitz. Let $\varepsilon, \varepsilon' > 0$. Let $v := \lim_{n \to \infty} v^{\varepsilon 2^{-n}}$ and $v' := \lim_{n \to \infty} v^{\varepsilon' 2^{-n}}$. We know that these limits exist from Lemma 3.1. Lemma 3.3 tells us that $\|u_I^\varepsilon - v^\varepsilon\|_\infty \to 0$ as $\varepsilon \to 0$. Since the assumptions of that lemma are player-symmetric, we likewise get $\|u_{\mathrm{II}}^\varepsilon - w^\varepsilon\|_\infty \to 0$. From Theorems 2.2 and 2.4 (possibly with the roles of the players interchanged) we know that $u_{\mathrm{I}}^\varepsilon = u_{\mathrm{II}}^\varepsilon$, and hence the above gives $\|v^\varepsilon - w^\varepsilon\|_\infty \to 0$. By Lemma 3.1, $v^\varepsilon \le v \le w^\varepsilon$ and $v^\varepsilon \le u_{\mathrm{I}}^\varepsilon \le w^\varepsilon$, and so we conclude that $\|v^\varepsilon - v\|_\infty \to 0$ and $\|u_{\mathrm{I}}^\varepsilon - v\|_\infty \to 0$. Note that the assumptions imply that $\sup|f| < \infty$. Therefore, from Lemma 3.5 and $\|u_{\mathrm{I}}^\varepsilon - v\|_\infty \to 0$ we conclude that v is Lipschitz. Precisely the same argument gives $\|v^\varepsilon - v\|_\infty = O(\varepsilon)$ if f is also assumed to be Lipschitz,

and similar estimates also hold for $\|v^{\varepsilon'} - v'\|_\infty$. Note that the assumptions imply that F is bounded if $f \neq 0$. Lemma 3.6 (applied possibly with the roles of the players reversed) tells us that v' satisfies $(-2f)$-quadratic comparison on $X \setminus Y$. Clearly $\inf v' > -\infty$. (If $f = 0$ identically, then $\inf v' \geq \inf F$, while if $\operatorname{diam} X < \infty$, we may use the fact that v' is Lipschitz.) Thus, Lemma 3.8 implies that $v^\varepsilon \leq v'$. Consequently, $v \leq v'$. By symmetry $v' \leq v$, and hence $v = v'$. This completes the proof in the case where F is Lipschitz.

Using the result of Lemma 3.9 below, we know that for every $\delta > 0$ there is a Lipschitz $F_\delta : Y \to \mathbb{R}$ such that $\|F - F_\delta\|_\infty < \delta$. Then the Lipschitz case applies to the functions $F_\delta \pm \delta$ in place of F. Since the game value u^ε for F is bounded between the corresponding value with $F_\delta + \delta$ and $F_\delta - \delta$, and the latter two values differ by 2δ, the result easily follows. □

Lemma 3.9. *Let X be a length space, and let $F : Y \to \mathbb{R}$ be defined on a non-empty subset $Y \subset X$. The following conditions are equivalent:*

(1) *F is uniformly continuous and* $\sup\{(F(y)-F(y'))/\max\{1, d(y,y')\} : y, y' \in Y\} < \infty$.
(2) *F extends to a uniformly continuous function on X.*
(3) *There is a sequence of Lipschitz functions on Y tending to F in $\|\cdot\|_\infty$.*

Proof. We start by assuming (1) and proving (2). Let $\varphi(\delta) := \sup\{F(y) - F(y') : d(y,y') \leq \delta,\, y, y' \in Y\}$ and $\tilde{\varphi}(t) := \sup\{t\,\varphi(\delta)/\delta : \delta \geq t\}$. We now show that $\lim_{t \searrow 0} \tilde{\varphi}(t) = 0$. Let $\varepsilon > 0$, and let $\delta_\varepsilon > 0$ satisfy $\varphi(\delta_\varepsilon) < \varepsilon$. Such a δ_ε exists because F is uniformly continuous. Condition (1) implies that $M := \sup\{\varphi(\delta)/\delta : \delta \geq \delta_\varepsilon\} < \infty$. For $t < (\varepsilon/M) \wedge \delta_\varepsilon$, we have

$$\tilde{\varphi}(t) = \sup\{t\,\varphi(\delta)/\delta : \delta_\varepsilon \geq \delta \geq t\} \vee \sup\{t\,\varphi(\delta)/\delta : \delta > \delta_\varepsilon\} \leq \varphi(\delta_\varepsilon) \vee (t\,M) \leq \varepsilon\,,$$

which proves that $\lim_{t \searrow 0} \tilde{\varphi}(t) = 0$. Two other immediate properties of $\tilde{\varphi}$ which we will use are that $F(y) - F(y') \leq \varphi\big(d(y,y')\big) \leq \tilde{\varphi}\big(d(y,y')\big)$ holds for $y, y' \in Y$ and $\tilde{\varphi}(s\,t) \geq s\,\tilde{\varphi}(t)$ when $s \in [0,1]$ and $t \geq 0$. It follows that $\tilde{\varphi}$ is subadditive: $\tilde{\varphi}(a) + \tilde{\varphi}(b) \geq a\,\tilde{\varphi}(a+b)/(a+b) + b\,\tilde{\varphi}(a+b)/(a+b) = \tilde{\varphi}(a+b)$ for $a, b \geq 0$.

Now for $x \in X$ set $u(x) := \inf\{F(y) + \tilde{\varphi}(d(y,x)) : y \in Y\}$. Then $u = F$ on Y. We now prove

$$u(x) - u(x') \leq \tilde{\varphi}\big(d(x,x')\big) \tag{3.6}$$

for $x, x' \in X$. Indeed, let $y' \in Y$. Then

$$\begin{aligned} u(x) - F(y') - \tilde{\varphi}\big(d(y',x')\big) &\leq F(y') + \tilde{\varphi}\big(d(y',x)\big) - F(y') - \tilde{\varphi}\big(d(y',x')\big) \\ &= \tilde{\varphi}\big(d(y',x)\big) - \tilde{\varphi}\big(d(y',x')\big). \end{aligned}$$

Consequently, subadditivity and monotonicity of $\tilde{\varphi}$ gives

$$u(x) - F(y') - \tilde{\varphi}\big(d(y',x')\big) \leq \tilde{\varphi}\big(\big|d(x,y') - d(x',y')\big|\big) \leq \tilde{\varphi}\big(d(x,x')\big).$$

Taking the supremum over all $y' \in Y$ then implies (3.6). Therefore, u is uniformly continuous and (2) holds.

We now assume (2) and prove (3). Let $u : X \to \mathbb{R}$ be a uniformly continuous extension of F to X. It clearly suffices to approximate u by Lipschitz functions on X in $\|\cdot\|_\infty$. For $x \in X$ let $u_L(x) := \inf\{u(x') + L\,d(x,x') : x' \in X\}$. The same argument which was used above to prove (3.6) now shows that $\operatorname{Lip}(u_L) \leq L$.

Clearly, $u_L(x) \le u(x)$. Let $\varphi(t) := \sup\{u(x) - u(x') : d(x,x') \le t\}$ for $t \ge 0$. Since X is a length space, φ is subadditive. Let $t > 0$ and $k := \lfloor d(x,x')/t \rfloor$. Then

$$u(x) - u(x') - L\,d(x,x') \le \varphi\big(d(x,x')\big) - L\,d(x,x') \le \varphi\big((k+1)\,t\big) - L\,k\,t\,.$$

Taking the supremum over all x' and using the subadditivity of φ therefore gives

$$u(x) - u_L(x) \le \sup_{k\in\mathbb{N}}\Big(\varphi\big((k+1)\,t\big) - L\,k\,t\Big) \le \varphi(t) + \sup_{k\in\mathbb{N}} k\,\big(\varphi(t) - L\,t\big).$$

Therefore, $u(x) - u_L(x) \le \varphi(t)$ once $L > \varphi(t)/t$. Since $\inf_{t>0}\varphi(t) = 0$ and $u_L \le u(x)$, this proves (3).

The passage from (3) to (1) is standard, and therefore omitted. This concludes the proof. □

Proof of Theorem 1.8. Lemma 3.9 tells us that F extends to a uniformly continuous function on X and that it can be approximated in $\|\cdot\|_\infty$ by Lipschitz functions. We know from Theorem 1.3 that the continuum value u exists and is uniformly continuous. Suppose first that F is Lipschitz. Lemma 3.6 (applied possibly with the roles of the players reversed) says that u satisfies $(-2f)$-quadratic comparison on $X \setminus Y$. If F is not Lipschitz, we may deduce the same result by approximating F from below by Lipschitz functions and observing that a monotone non-decreasing $\|\cdot\|_\infty$-limit of functions satisfying $(-2f)$-quadratic comparison from above also satisfies $(-2f)$-quadratic comparison from above, and making the symmetric argument for comparison from below. Now suppose that $\tilde{u}$ is as in the second part of the theorem. This clearly implies $\inf \tilde{u} > -\infty$. Now Lemma 3.8 gives $u \le \tilde{u}$ (again, we may need to first approximate F by Lipschitz functions). The uniqueness follows directly. □

Proof of Theorem 1.4. If $x \in X \setminus Y$ and $y \in Y$, then from Lemma 3.2 it follows that $\big|u(x) - u(y)\big| \le \mathrm{Lip}_Y F\, d(x,y)$. Let $U \subset X \setminus Y$ be open and define $F^*(x) = u(x)$ for $x \in Y \cup \partial U$. It is clear that the continuum value u^* of $(X, d, Y \cup \partial U, F^*, 0)$ is the same as u, since any player may first play a strategy that is appropriate for the $(X, d, Y \cup \partial U, F^*, 0)$ game, and once $Y \cup \partial U$ is hit, start playing a strategy that is appropriate for the original game. The above argument shows that $\big|u(x) - u(y)\big| \le d(x,y)\,\mathrm{Lip}_{\partial U} u$ if $y \in \partial U$ and $x \in U$. In particular, $\mathrm{Lip}_{\partial U \cup \{x\}} u = \mathrm{Lip}_{\partial U} u$, which implies $\mathrm{Lip}_U u = \mathrm{Lip}_{\partial U} u$; that is, u is AM in $X \setminus Y$.

Now suppose that $u^* : X \to \mathbb{R}$ is an AM extension of F and $\sup|F| < \infty$. Lemma 1.6 tells us that u^* satisfies comparison with distance functions. Observe that the proof of Lemma 3.8 shows that in the case $f = 0$ we may replace the hypothesis that v satisfies 0-quadratic comparison from below on $X \setminus Y$ by the hypothesis that v satisfies comparison with distance functions from below (since only comparisons with distance functions are used in this case). Thus, $u \le u^*$. Similarly, $u \ge u^*$, which implies the required uniqueness statement and completes the proof. □

4. Harmonic measure for Δ_∞

Here, we present a few estimates of the ∞-harmonic measure ω_∞. Before proving Theorem 1.5, we consider the ∞-harmonic of porous sets. Recall that a set S in a metric space Z is α-**porous** if for every $r \in (0, \mathrm{diam} Z)$ every ball of radius r contains a ball of radius $\alpha\, r$ that is disjoint from S. An example of a porous set

is the ternary Cantor set in $[0,1]$. We start with a general lemma in the setting of length spaces.

Lemma 4.1. *Suppose X is a length space and $0 \le F \le 1$ on the terminal states Y, where $F : Y \to \mathbb{R}$ is continuous. Let $S := \operatorname{supp} F$. Suppose that for some integer k and positive constants $\varepsilon_0, d_{\min}$ and $\gamma \in (0,1)$, for every $x \in X$ with $d(x,S) \ge d_{\min}$, there is a sequence of points $x = z_0, \dots, z_k$ such that $z_k \in Y$, $F(z_k) = 0$, and*

$$d(z_i, S) \ge 2\, d(z_i, z_{i-1}) + 2\,\varepsilon_0 + \gamma\, d(z_0, S)$$

for $i = 1, \dots, k$. Then the AM extension u of F satisfies

$$0 \le u(x) \le (1 - 2^{-k})^{\log_\gamma (d_{\min}/d(x,S))}.$$

Proof. We will obtain bounds on $u^\varepsilon(x)$ that are independent of ε (as long as $\varepsilon \in (0, \varepsilon_0)$), and these yield bounds on $u(x)$. Since $u^\varepsilon \ge 0$, we need only give a good strategy for player II to obtain an upper bound on $u^\varepsilon(x)$. The idea is for player II to always have a "plan" for reaching a terminal state at which F is 0, while staying away from the terminal states at which F is non-zero. A plan consists of a sequence of points $z_0, z_1, \dots, z_k$, where z_0 is the game state when the plan was formed, $z_k \in Y$, and $F(z_k) = 0$. As soon as the game state reaches z_i, player II starts pulling towards z_{i+1}. If player I is lucky and gets many moves, player II may have to give up on the plan and form a new plan. When tugging towards z_i, we suppose that player II gives up and forms a new plan as soon as $d^\varepsilon(x_t, z_i) = 2d^\varepsilon(z_i, z_{i-1})$, and otherwise plays to ensure that $d^\varepsilon(x_t, z_i)$ is a supermartingale. While tugging towards z_i, the plan will be aborted at that stage with probability at most $1/2$, so with probability at least 2^{-k} the plan is never aborted and succeeds in reaching z_k.

Suppose player II aborts the plan at time t while tugging towards z_i, and forms a new plan starting at $z_0' = x_t$. Then $d(x_t, z_i) \le d^\varepsilon(x_t, z_i) = 2\, d^\varepsilon(z_i, z_{i-1}) \le 2\, d(z_i, z_{i-1}) + 2\,\varepsilon$, so

$$d(x_t, S) \ge d(z_i, S) - d(x_t, z_i) \ge d(z_i, S) - [2d(z_i, z_{i-1}) + 2\varepsilon] \ge \gamma\, d(z_0, S),$$

so the game state remains far from S. Since player II can always find a short plan (length at most k) when the distance from S is at least $d_{\min}$, player I gets to S with probability at most $(1-2^{-k})^{\lceil \log_\gamma (d_{\min}/d(x_0,S))\rceil}$, which yields the desired upper bound. □

The reader may wish to check that the lemma implies $\lim_{\delta \searrow 0} \omega_\infty(A_\delta) = 0$ in the setting of Theorem 1.5 (i.e., when X is the unit ball in $\mathbb{R}^n$, $n > 1$, $Y = \partial X$ and A_δ is a spherical cap of radius δ).

We are now ready to state and prove an upper bound on the ∞-harmonic measure of neighborhoods of porous sets.

Theorem 4.2. *Let $X \subset \mathbb{R}^n$, $n > 1$, be the closed unit ball and let Y be its boundary, the unit sphere. Let $\alpha \in (0, 1/2)$ and let $\delta > 0$. Let S be an α-porous subset of Y, and let S_δ be the closure of the δ-neighborhood of S. Then*

$$\omega_\infty^{(0,Y,X)}(S_\delta) \le \delta^{\alpha^{O(1)}}.$$

Of course, in the above, S is α-porous as a subset of Y. (Every subset of Y is $(1/3)$-porous as a subset of X.)

Proof. The plan is to use the lemma, of course. Let $d_{\min} := 2\,\delta/\alpha$. Let $z_0 \in X \setminus Y$, and suppose that $d_0 := d(z_0, S) \ge d_{\min}$. Let $y_0 \in Y$ be a closest point to z_0 on Y.

Inside $B_{d_0}(y_0) \cap Y$ there is a point y_1 such that $B_{\alpha d_0}(y_1) \cap S = \emptyset$. (See Figure 3.) Therefore, $d(y_1, S_\delta) \geq \alpha\, d_0 - \delta \geq \alpha\, d_0/2$. We define the sequence z_j inductively, as follows. If $d(z_j, S_\delta) \geq 3\, d(z_j, Y)$, then we take z_{j+1} to be any closest point to z_j on Y. Otherwise, let z_{j+1} be the point on the line segment from z_j to y_1 whose distance from z_j is $d(z_j, y_1)/10$. It can be checked that after $k = O(-\log \alpha)$ steps the sequence hits $Y \setminus S$ and that the assumptions of the lemma hold with this k, with $\gamma = \alpha/10$ and with some $\varepsilon_0 > 0$, independent of z_0. The theorem now easily follows from the lemma. □

We now proceed to study the ∞-harmonic measure of spherical caps.

Proof of Theorem 1.5. It turns out to be more convenient to work with $-A_\delta = \{-y : y \in A_\delta\}$ in place of A_δ. By an obvious comparison argument, it is sufficient

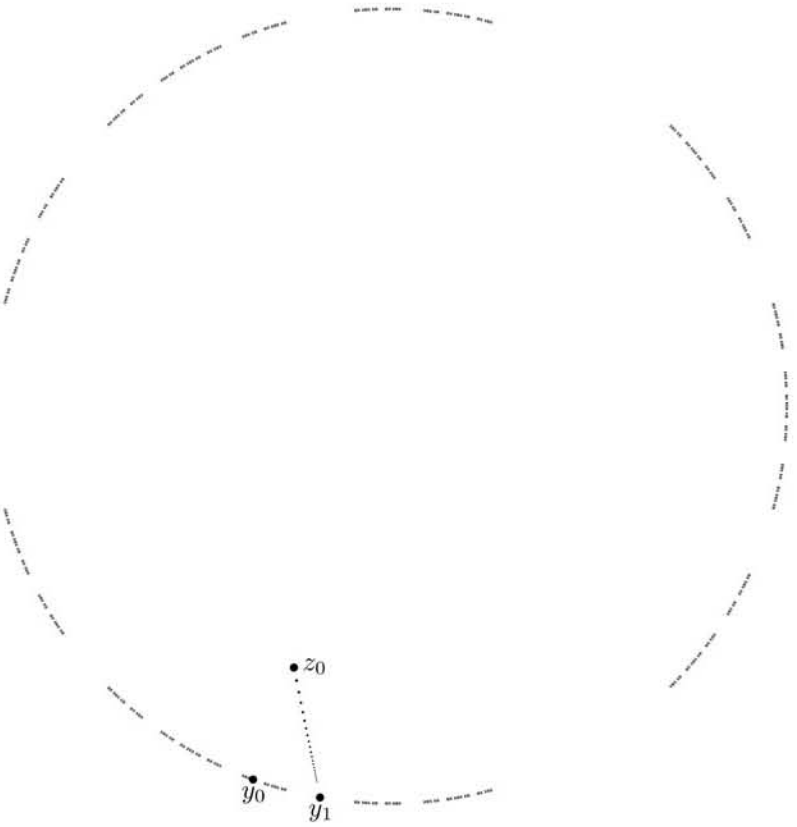


FIGURE 3. The terminal states are the unit sphere, and the support of F is S_δ, the δ-neighborhood of a porous set on the unit sphere. From a starting point z_0, player II finds a point y_1 on the sphere that is near the closest point y_0 on the sphere but far from S_δ. Player II then tugs towards y_1 along a sequence of points, and then when it gets much closer to the sphere than to S_δ, it tugs straight to the sphere.

to estimate $u(0)$, where $u = u_\delta$ is the AM function in $X \setminus Y$ with boundary values

$$F(y) := \begin{cases} 1\,, & y \in -A_\delta\,, \\ 0\,, & d(y, -A_\delta) > \delta\,, \\ 1 - d(y, -A_\delta)\,\delta^{-1}\,, & 0 < d(y, -A_\delta) \le \delta\,. \end{cases}$$

The function u is invariant under rotations of X preserving $(1, 0, \dots, 0) \in \mathbb{R}^n$. Therefore, u is also AM in $\mathbb{R}^2 \cap X$. Thus, we henceforth restrict ourselves to the case $n = 2$, with no loss of generality.

Aronsson [3] constructed a family of viscosity solutions G_m to $\Delta_\infty u = 0$ in $\mathbb{R}^2 \setminus \{0\}$ that are separable in polar coordinates: $G_m(r, \theta) = r^{m^2/(2m-1)} h_m(\theta)$ (for $m \in \mathbb{Z}$). We are interested in the $m = -1$ solution, which may be written as

$$G = \left[\frac{\cos\theta\,(1 - |\tan(\theta/2)|^{4/3})^2}{1 + |\tan(\theta/2)|^{4/3} + |\tan(\theta/2)|^{8/3}} \right]^{1/3} r^{-1/3}. \tag{4.1}$$

Observe that when $-\pi/2 \le \theta \le \pi/2$ this solution is non-negative, and $G \ge c\, r^{-1/3}$ when $\theta \in [-\pi/4, \pi/4]$, say, where $c > 0$ is some fixed constant. Then $F \le O(1)\,\tilde{G}_\delta$, where

$$\tilde{G}_\delta(x, y) = \delta^{1/3}\, G(x + 1 + 2\,\delta, y)\,.$$

Since $u(0)$ is monotone in F, and $\tilde{G}_\delta$ satisfies $\Delta_\infty \tilde{G}_\delta = 0$, it follows that $u(0) \le O(1)\,\tilde{G}_\delta(0) = O(\delta^{1/3})$.

We now show that the bound $u(0) = O(\delta^{1/3})$ is tight. It is easy to see, using comparison with a cone centered at $(-1, 0)$, say, that $u(-1+\delta/10, 0) > c > 0$, where c is a constant that does not depend on δ. Comparison with a cone centered at any point z in the unit disk shows that $u(z') \ge u(z)/2$ if $|z - z'| \le (1 - |z|)/2$. Using such estimates, it is easy to see that there is a constant $c' > 0$ such that $u \ge c'$ on the disk B of radius δ centered at $q := (-1 + 2\delta, 0)$, provided that $\delta < 1/4$, say. Now consider the function

$$G^*_\delta(x, y) = c'\,\delta^{1/3}\, G\big((x, y) - q\big)\,.$$

By the choice of the constant c', we have $G^*_\delta \le u$ on ∂B, since $G \le r^{-1/3}$ in $\mathbb{R}^2 \setminus \{0\}$. If (r', θ') denote the polar coordinates centered at q, the center of B, and (r, θ) denote the standard polar coordinates centered at 0, then $2\,|\theta'| \ge \pi - |\theta - \pi|$ when $r = 1$, and we choose $\theta' \in [-\pi, \pi)$. Also, r' is bounded from below by a constant times $|\theta - \pi|$ when $\theta \in [0, 2\pi]$ and $r = 1$. Since $|G| \le O(1)\,\big((\pi/2) - |\theta|\big)\, r^{-1/3}$ when $\theta \in [-\pi, \pi)$, it follows that on the unit circle

$$G^*_\delta \le O(1)\,\big((\pi/2) - |\theta'|\big)(r')^{-1/3} \le O(1)\,\delta^{1/3}\,|\theta - \pi|^{2/3} \le O(\delta^{1/3})\,.$$

Therefore $G^*_\delta - O(\delta^{1/3}) \le u$ on the boundary of the unit circle, as well as on ∂B. Consequently, $u \ge G^*_\delta - O(\delta^{1/3})$ in the complement of B in the unit disk. This implies that there is some $r \in (0, 1)$ such that $u_\delta(-r, 0) \ge \delta^{-1/3}$ for all sufficiently small δ. The required estimate $u_\delta(0) \ge \Theta(\delta^{-1/3})$ now follows by several applications of comparison with cones (the number of which depends on r), similar to the above argument estimating a lower bound for u on B. □

One could also try to use Aronsson's other solutions G_m to bound ∞-harmonic measure of certain sets in other domains. Also, Aronsson has some explicit solutions for $\Delta_p u = 0$, which may serve a similar purpose.

5. Counterexamples

5.1. **Tug-of-war games with positive payoffs and no value.** Here we give the promised counterexample showing that the hypothesis $\inf f > 0$ in Theorem 1.2 cannot be relaxed to $f > 0$. As we later point out, a similar construction works in the continuum setting of length spaces.

The comb game is tug-of-war with running payoffs on an infinite graph shaped like a comb, as shown in Figure 4. The comb is defined by an infinite sequence of positive integers $\ell_0, \ell_1, \ell_2 \ldots$, and has states (vertices) $\{(x, y) \in \mathbb{Z}^2 : 0 \le x \text{ and } 0 \le y \le \ell_x\}$. The edges of the comb are of the form $(x, y) \sim (x, y + 1)$ and $(x, 0) \sim (x + 1, 0)$, giving the graph the shape of a comb, where the xth tooth of the comb has length ℓ_x. The running payoff $f(x, y)$ is $1/\ell_x$ if $y = 0$ and zero otherwise. (Later we will consider a variation where $f > 0$ everywhere.) The terminal states are the states of the form (x, ℓ_x), and the terminal payoff F is zero on all terminal states.

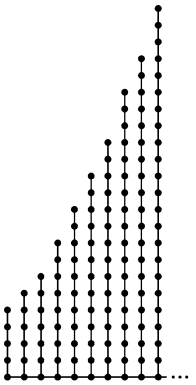


Figure 4. A comb.

Lemma 5.1. *For any comb (choice of the sequence $\{\ell_x\}$), we have $u_{\mathrm{I}}(0, 0) = 2$.*

Proof. First we argue that $u_{\mathrm{I}}(0, 0) \ge 2$. Denote by $\psi(t) := \sum_{i=0}^{t-1} f(x_i, y_i)$ the accumulated running payoff. Fix a large $B > 0$ and equip player I with the following strategy: at all times $t < \tau$ for which $\psi(t) < B$, player I pulls down if $y_t \ne 0$, left if $y_t = 0$ and $x_t \ne 0$, and right if $(x_t, y_t) = (0, 0)$; if $\psi(x_t) \ge B$, then player I pulls toward the closest terminal state. Define the termination time $\tau = \inf\{t : y_t = \ell_{x_t}\}$ and also the stopping time $\sigma_0 = \inf\{t : x_t = 0\} \wedge \tau$. Fix any strategy for player II and observe that $\{x_{t \wedge \sigma_0}\}_{t \ge 0}$ is a non-negative supermartingale, which must converge a.s.; therefore, from any initial state, $\sigma_0 < \infty$ a.s.; therefore, the stopping time $\sigma = \inf\{t : \psi(t) \ge B\} \wedge \tau$ is almost surely finite, whence $\tau < \infty$ a.s. as well.

Consider the process

$$M_t = 2(1 - y_t/\ell_{x_t}) + \psi(t)\,.$$

Then $M_{t \wedge \sigma}$ is a submartingale. Note that if $y_\sigma = 0$, then $M_\sigma = \psi(\sigma) + 2$ and $\psi(\sigma) \ge B$, while if $y_\sigma > 0$, then $\sigma = \tau$ and $M_\sigma = \psi(\sigma)$. In any case, $M_\sigma \le \psi(\sigma)(1 + 2/B)$. Thus by optional stopping,

$$2 = M_0 \le \mathbb{E} M_\sigma \le (1 + \frac{2}{B})\mathbb{E}\psi(\sigma) \le (1 + \frac{2}{B}) u_{\mathrm{I}}(0, 0).$$

Next, to show that $u_{\mathrm{I}}(0,0) \le 2$, suppose that player II adopts the strategy of always pulling up, and player I knows this and seeks to maximize her payoff. Because player II always pulls up, M_t is a positive supermartingale, so M_t a.s. converges to M_∞, and by optional stopping,

$$2(1 - y_0/\ell_{x_0}) = M_0 \ge \mathbb{E}[M_\infty] \ge \mathbb{E}\big[\sum_{t'=0}^{\infty} f(x_{t'}, y_{t'})\big].$$

This shows that by always pulling up, player II can force player I's expected payoff from $(x_0, y_0) = (0,0)$ to be no more than $2(1 - y_0/\ell_{x_0}) = 2$. □

Remark 5.2. Note that the strategy for player II of always pulling up does not necessarily force the game to end with probability one; this strategy always gives an upper bound on $u_{\mathrm{I}}(0,0)$, but it only yields an upper bound on $u_{\mathrm{II}}(0,0)$ if $\sum_{x\ge 0} \ell_x^{-1} = \infty$, when the Borel-Cantelli Lemma ensures termination.

Suppose that the teeth of the comb are long, i.e., $\sum_{x=0}^{\infty} \ell_x^{-1} < \infty$, and that player I plays the strategy of always pulling down if $y_t > 0$ and always pulling right if $y_t = 0$. If player II plays the strategy of always pulling up, then we may calculate the probability that the game terminates when started in state $(x,0)$. Either the state goes to the terminal state of the tooth or goes to the base of the next tooth, and the latter probability is $\ell_x/(\ell_x + 1)$. When the teeth of the comb are long, the game lasts forever with probability $\prod_{x\ge 0} \ell_x/(\ell_x + 1) > 0$. If player II needs to ensure that the game terminates, she will need to be prepared to sometimes pull left instead of up, and this necessity will be costly for player II.

Lemma 5.3. *For every $s > 0$ there is a suitable comb (choice of the sequence $\{\ell_x\}$) such that $u_{\mathrm{II}}(0,0) \ge s$.*

Proof. Let player I play the strategy of pulling down when $y > 0$ and pulling to the right when $y = 0$. We will estimate from below the expected payoff when player II plays any strategy which guarantees that the game terminates in finite time a.s. against the above strategy for player I. Let a_x denote the probability that the game terminates at the terminal state (x, ℓ_x). Then $\sum_{x=0}^{\infty} a_x = 1$. For every $x \in \mathbb{N}$ let L_x, U_x and R_x denote the expected number of game transitions from $(x,0)$ to the left, upwards and to the right, respectively (that is, to $(x-1,0)$, to $(x,1)$ and to $(x+1,0)$, respectively).

First, observe that

$$U_x \ge a_x\, \ell_x\,,$$

because every time that the game state arrives at $(x,1)$, with conditional probability at most $1/\ell_x$, the game terminates at (x,ℓ_x) before returning to $(x,0)$. Next, observe that

$$L_{x+1} \ge R_x - 1\,,$$

because the number of transitions from $(x,0)$ to $(x+1,0)$ can exceed the number of transitions from $(x+1,0)$ to $(x,0)$ by at most one. (More precisely, we have $L_{x+1} = R_x - \sum_{j>x} a_j$, but this will not be needed.) Finally, note that

$$R_x \ge L_x + U_x\,,$$

because player I always pulls right at $(x,0)$. (We set $L_0 := 0$.)

An easy induction shows that the above relations imply

$$R_x \geq \sum_{j=0}^{x} a_j \, \ell_j - x \, .$$

The expected payoff $\mathcal{E}$ satisfies

$$\mathcal{E} = \sum_{x=0}^{\infty} \frac{R_x + L_x + U_x}{\ell_x} \geq \sum_{x=0}^{\infty} \frac{R_x}{\ell_x} \, .$$

We plug in the above lower bound on R_x to obtain

$$\mathcal{E} \geq \sum_{j=0}^{\infty} a_j \left(\sum_{x=j}^{\infty} \frac{\ell_j}{\ell_x} \right) - \sum_{x=0}^{\infty} \frac{x}{\ell_x} \, .$$

Since $\sum_{x \geq 0} a_x = 1$, the lemma follows if we choose $\ell_x = (c + x)^3$ with $c > 0$ sufficiently large. □

Note that when player II always pulls up, for every vertex v in the comb there is a finite upper bound $b(v)$ on the expected number of visits to v, which holds regardless of the strategy used by player I. Since in the proof of Lemma 5.1 player II always pulls up, it follows that the value for player I is at most 3 even when f is replaced by $f(\cdot) + q(\cdot)/\big(b(\cdot) + 1\big)$, where $q > 0$ and $\sum_v q(v) = 1$. This modification will certainly not decrease the value for player II. Therefore, in Theorem 1.2, the assumption $\inf f > 0$ cannot be replaced by the assumption $f > 0$, even if $F = 0$ throughout Y.

Note that we may convert the discrete comb graph to a continuous length space by adding line segments corresponding to edges in the comb. It is then easy to define the corresponding continuous f and conclude that in Theorem 1.3 the assumption $\inf |f| > 0$ cannot be relaxed to $f > 0$. The details are left to the reader.

Remark 5.4. The dependence of $u_{\mathrm{II}}(0,0)$ on the growth of $\{\ell_x\}$ is somewhat surprising. If $\{\ell_x\}$ grows slowly enough so that $\sum_{x \geq 0} \ell_x^{-1} = \infty$, then $u_{\mathrm{II}}(0,0) \leq 2$, because (as remarked above) the Borel-Cantelli lemma implies that the strategy of always pulling up is guaranteed to terminate, and thus the proof of Lemma 5.1 applies. We have seen that for some sequences $\{\ell_x\}$ of polynomial growth, $u_{\mathrm{II}}(0,0)$ can be arbitrarily large. However, if $\{\ell_x\}$ grows rapidly enough so that $\ell_{x+1}/\ell_x \to \infty$, then $u_{\mathrm{II}}(0,0) \leq 2$. This is immediate from the following more general statement: for every $k > 0$, we have $u_{\mathrm{II}}(0,0) \leq 2 + 2\ell_k \sum_{j>k} \ell_j^{-1}$. To prove this, suppose player II adopts the strategy of always pulling left when $x_t > k$ and $y_t = 0$ and up otherwise. Let $\psi(t,k)$ denote the payoff accumulated for player I at points in $[0,k] \times \{0\}$ up to time t. Then $\mathbb{E}\psi(t,k) \leq 2$, because

$$\widetilde{M}_t = 2(1 - \frac{y_t}{\ell_{x_t}}) + \psi(t,k)$$

is a positive supermartingale with $\widetilde{M}_0 = 2$. Then we claim that

$$\mathbb{E}[\psi(t) - \psi(t,k)] \leq 2\ell_k \sum_{j>k} \ell_j^{-1}.$$

Fix $j > k$. Starting at $(k,0)$, the expected number of visits to $(j,0)$ before returning to $(k,0)$ is at most 1 (by comparison to simple random walk). The expected number of visits to $(k,0)$ is at most $2\ell_k$, since each time $(x_t, y_t) = (k,0)$,

there is a chance of at least $1/(2\ell_k)$ of terminating at (k, ℓ_k) without returning to $(k, 0)$. Thus the expected accumulated payoff at $(j, 0)$ is at most $2\ell_k \ell_j^{-1}$; summing over $j > k$ proves the claim.

5.2. **Positive Lipschitz function with multiple AM extensions.** Here we show that uniqueness in Theorem 1.4 may fail if F is unbounded; that is, we give an (X, Y, F) (here $f = 0$) for which F is Lipschitz and positive and the continuum value is not the only AM extension of F.

Let T be the rooted ternary tree, where each node has three direct descendants and every node but the root has one parent. Let X be the corresponding length space, where we glue in a line segment of length 1 for every edge in T. For every node v in T, we label the three edges leading to descendants of v by 1, -1 and $*$. (One may interpret T as the set of finite stacks of cards, where each card has one of the three labels 1, -1, and $*$.)

We define a function w on the nodes of T by induction on the distance from the root. For any vertex v let $k(v)$ be the number of edges on the simple path from the root to v whose label is not $*$. Set $w(\text{root}) := 0$. Next, if v is the parent of v' and the edge from v to v' is labeled ± 1, then $w(v') = w(v) \pm 1$, respectively. Finally, if the edge from v to v' is labeled $*$, let $w(v') = w(v) + 1 - 2^{-k(v)}$, say. This defines w on the vertices of T. We define it on X by linear interpolation along the edges.

Let V_0 denote the set of nodes consisting of the root and all vertices v such that the edge from v to its parent is not labeled $*$. For each $v \in V_0$ let $b(v)$ be some large integer, whose value will be later specified, and let $q(v)$ be the vertex at distance $b(v)$ away from v along the (unique) infinite simple path starting at v which contains only edges labeled $*$. Let $Y_1 = \{q(v) : v \in V_0\}$ and let Y_0 be the set of all nodes v such that $w(v) \le 3/2$. Set $Y := Y_0 \cup Y_1$.

We first claim that w is AM on $X \setminus Y$. Indeed, let U be an open subset of $X \setminus Y$. If U does not contain any tree node, then $\operatorname{Lip}_U w = \operatorname{Lip}_{\partial U} w$, because w interpolates linearly inside the edges. Suppose now that $v \in U$ is a node. Let β be the infinite path starting at v always going away from the root which uses only edges labeled -1. For each positive integer m let γ_m be the infinite path starting at v always going away from the root whose first m edges are labeled 1 and the rest are labeled $*$. Observe that γ_m meets Y_1 and β meets Y_0. Consequently, $\gamma_m \cap \partial U \neq \emptyset$ and $\beta \cap \partial U \neq \emptyset$. If $x_1 \in \gamma_m \cap \partial U$ and $x_0 \in \beta \cap \partial U$, then $w(x_1) - w(x_0) \ge (1 - 2^{-m})\, d(x_1, x_0)$ by the construction of w. Thus $\operatorname{Lip}_{\partial U} w \ge 1$. Since $\operatorname{Lip}_X w = 1$, this proves that w is AM on $X \setminus Y$.

Next, we consider the discrete tug-of-war game on the graph T, where $f = 0$ and F is the restriction of w to Y. Let x_0, the starting position of the game, be the third vertex on the infinite simple path from the root whose edges are all labeled 1. We claim that the value u_{I} for player I satisfies $u_{\mathrm{I}}(x_0) < w(x_0)$. Before proving this we explain the idea: at each step, player I has one or two moves that increase the value of w by 1 (either moving away from the root along an edge with label 1 — i.e., "adding a 1 card to the deck" — or moving towards the root along an edge labeled -1 — i.e., "removing a -1 card from the top of the deck"), and similarly, player II has one or two moves that decrease the value of w by 1. Player I can thus make $w(x_j)$ a submartingale by always making moves of this type.

This does not force the game to terminate, however. In order to end the game favorably by reaching a point in Y_1, player I must add a sequence of cards labeled $*$ to the top of the deck. If player II adopts the strategy of always choosing the

edge labeled -1 ("adding a -1 card to the deck"), then (provided that $b(\cdot)$ increases rapidly enough) the expected number of suboptimal moves that player I must make (by moving along an edge with label $*$) to reach any particular element of Y_1 is large enough to significantly decrease the total expected payoff for player I.

We proceed to prove that $u_{\mathrm{I}}(x_0) < w(x_0)$. Let v_0 be a vertex in $V_0 \setminus Y$, which is in the same connected component of $T \setminus Y$ as x_0. Let $(v_0, v_1, \ldots, v_{b(v_0)})$ be the simple path from v_0 to $q(v_0)$. We now abbreviate $n = b(v_0)$, $h = w(v_0)$ and $a = 1 - 2^{-k(v_0)}$.

Let player II use the naive strategy of always trying to move from the current state along the edge labeled -1 going away from the root. For $i \in \{0, 1, \ldots, n\}$ let τ_i be the first time t such that $x_t = v_i$, and if no such t exists let $\tau_i = \infty$. (As before x_t is the game position at time t.) Let $M_t^{(i)} = w(x_t)$ for $t < \tau_i$ and $M_t^{(i)} = w(v_{i-1}) + 1$ for $t \geq \tau_i$. Then $M_t^{(i)}$ is clearly a supermartingale, regardless of the strategy used by player I. Since the game terminates at time t if $w(x_t) \leq 1$, this implies that for $i \in \mathbb{N}_+$

$$\begin{aligned}\Pr\bigl[\tau_i < \infty \bigm| \tau_{i-1} < \infty\bigr] &\leq \frac{w(v_{i-1})}{w(v_{i-1}) + 1} = 1 - \frac{1}{h + (i-1)\,a + 1} \\ &\leq \exp\Bigl(\frac{-1}{h + (i-1)\,a + 1}\Bigr)\,.\end{aligned}$$

Consequently,

$$\begin{aligned}\Pr\bigl[\tau_n < \infty\bigr] \leq \prod_{i=1}^{n} \Pr\bigl[\tau_i < \infty \bigm| \tau_{i-1} < \infty\bigr] &\leq \exp\Bigl(-\sum_{i=1}^{n} \frac{1}{h + (i-1)\,a + 1}\Bigr) \\ &\leq \exp\Bigl(-\int_0^n \frac{ds}{h + 1 + s\,a}\Bigr) = \Bigl(\frac{h+1}{a\,n + h + 1}\Bigr)^{1/a}.\end{aligned}$$

Therefore

$$\Pr\bigl[\tau_n < \infty\bigr] F(v_n) \leq (h + n\,a)\,\Bigl(\frac{h+1}{a\,n + h + 1}\Bigr)^{1/a}.$$

Since $a < 1$, we can make this smaller than any required positive number by choosing $n = b(v_0)$ sufficiently large. We may therefore choose the function $b : V_0 \to \mathbb{N}_+$ so that

$$\sum_{v \in Y_1} \Pr\bigl[\text{game ends at } v\bigr]\, F(v) < 1/2\,.$$

Since $F \leq 3/2$ on Y_0, this implies $u_{\mathrm{I}}(x_0) < 2 < 3 = w(x_0)$.

Let u_* be the linear interpolation of u_{I} to the edges. By Lemma 1.1 u_{I} is discrete ∞-harmonic. It follows that in the II-favored ε-tug-of-war game, for every $\delta > 0$ and $\varepsilon \in (0, 1/2)$, player II can play to make $u_*(x_t) + 2^{-t}\delta$ a supermartingale. Consequently, Lemma 3.3 implies that the value of the continuum game is bounded by u_{I} on the vertices. Since the continuum value is AM on $X \setminus Y$, by Theorem 1.4, this proves our claim that the assumption $\sup F < \infty$ is necessary for uniqueness to follow in the setting of that theorem.

5.3. **Smooth f on a disk with no game value.** The following example is a continuum analog of the triangle example at the beginning of Section 2.2.

Let X be a closed disk in $\mathbb{R}^2$, and let $Y = \partial X$ be its boundary. We now show that for some smooth function $f : X \to \mathbb{R}$, the value $u_{\mathrm{I}}^{\varepsilon}$ for player I and the value $u_{\mathrm{II}}^{\varepsilon}$ for player II differ in ε-tug-of-war with $F = 0$ on Y. Not only are the values

different, but the difference does not shrink to zero as $\varepsilon \searrow 0$. There is a lot of freedom in the choice of the function f, but an essential property, at least for the proof, is that f is anti-symmetric about a line of symmetry of X.

It will be convenient to identify $\mathbb{R}^2$ with $\mathbb{C}$, and use complex numbers to denote points in $\mathbb{R}^2$. Let f be a C^∞ function such that:

(1) $f \geq 1$ inside the disk $|z-1| < 1/2$,
(2) $0 \leq f \leq 2$ in the right half plane $\operatorname{Re} z \geq 0$,
(3) $f = 0$ in $\{z \in \mathbb{C} : \operatorname{Re} z \geq 0,\, |z-1| > 11/20\}$, and
(4) f is anti-symmetric about the imaginary line.

Let $R > 2$ be large, let $X = \{z : |z| \leq R\}$, $Y = \{z : |z| = R\}$ and $F = 0$ on Y. Consider $\varepsilon > 0$ small. We want to show that for an appropriate choice of R, then

$$\liminf_{\varepsilon \searrow 0} \|u_{\mathrm{I}}^\varepsilon - u_{\mathrm{II}}^\varepsilon\|_\infty > 0\,. \tag{5.1}$$

Indeed, suppose that this is not the case. Let $\delta > 0$ be small, and suppose that $\|u_{\mathrm{I}}^\varepsilon - u_{\mathrm{II}}^\varepsilon\|_\infty < \delta$. Symmetry gives $u_{\mathrm{I}}^\varepsilon(x + i\,y) = -u_{\mathrm{II}}^\varepsilon(-x + i\,y)$. The assumption $\|u_{\mathrm{I}}^\varepsilon - u_{\mathrm{II}}^\varepsilon\|_\infty < \delta$ therefore tells us that $|u_{\mathrm{I}}^\varepsilon(x+i\,y) + u_{\mathrm{I}}^\varepsilon(-x+i\,y)| < \delta$. In particular, $|u_{\mathrm{I}}^\varepsilon| < 2\,\delta$ on the imaginary axis. We now abbreviate $w = u_{\mathrm{I}}^\varepsilon$. Clearly $w \geq -\delta$ on $\{z \in X : \operatorname{Re} z \geq 0\}$ and $w \leq \delta$ on $\{z \in X : \operatorname{Re} z \leq 0\}$. By considering strategies which pull towards the imaginary axis and then using whatever strategy gives a value in $[-2\,\delta, 2\,\delta]$, it is easy to see that $|w| = O(1)$ (as usual, $f = O(g)$ means that there is some universal constant C such that $f \leq C\,g$) and that $|w| \leq O(\delta + \varepsilon)$ in $\{z \in X : |\operatorname{Re} z| \leq 2\,\varepsilon\}$, say. (See, e.g., the proof of Lemma 3.5.) It is then easy to see that there is a constant $c_1 > 0$, which does not depend on R and ε (as long as ε is sufficiently small and $R > 2$, say), such that $w \geq c_1$ on the disk $|z-1| < 3/4$ and $w \leq -c_1$ on the disk $|z+1| < 3/4$. Since w is nearly anti-symmetric, it is enough to prove the first claim. Indeed, a strategy for player I which demonstrates this is one in which she pulls towards 1 until she accumulates a payoff of 1. If successful, she then aims towards the boundary Y, but whether successful or not, whenever the game position comes within distance ε of the imaginary axis, she changes strategy and adopts an arbitrary strategy that yields a payoff $O(\delta + \varepsilon)$. We assume that δ and ε are sufficiently small so that the $O(\delta + \varepsilon)$ term is much smaller than c_1.

It now easily follows from Theorem 1.5 that $|w(z)| \leq O(1)\,|z|^{-1/3} + O(\delta + \varepsilon)$. Consequently, there is some constant $r_1 > 2$ such that $|w| < c_1/10$ on $\{z \in X : |z| \geq r_1\}$, provided that $\varepsilon + \delta$ is sufficiently small. Set $L = c_1/(8\,r_1)$, and let $Z_L = Z_L^\varepsilon$ be the set of points $z \in X$ such that

$$\sup\{w(z') : z' \in B_z(\varepsilon)\} - \inf\{w(z') : z' \in B_z(\varepsilon)\} > 2\,L\,\varepsilon\,.$$

Since $|w| \leq c_1/10$ on $|z| = r_1$ and $|w| \geq c_1$ on $S := \{|z-1| < 3/4\} \cup \{|z+1| < 3/4\}$, it follows that every path in the graph (X, E_ε) from S to Y must intersect Z_L. (We are assuming $R > r_1$.) Let $\tilde{Z}_L$ denote the union of Z_L and all connected components of $(X \setminus Z_L, E_\varepsilon)$ that are disjoint from Y.

We now claim that there is some $r_2 > r_1$ (which does not depend on ε) such that $Z_L \subset \{|z| < r_2\}$ and therefore $\tilde{Z}_L \subset \{|z| < r_2\}$. Indeed, if $z_1, z_2 \in X$ satisfy $|z_1 - z_2| < \varepsilon$ and $|z_1| > 3\,r_1$, say, then the strategies for either player of pulling towards z_2 and only giving up when the current position is within distance ε of

$|z| = r_1$ show that

$$|w(z_1) - w(z_2)| \le \frac{\varepsilon}{|z_1| - r_1 - 2\,\varepsilon} \sup\{|w(z)| : |z| \ge r_1\} \le \frac{\varepsilon}{|z_1| - r_1 - 2\,\varepsilon}\, c_1/10\,.$$

This proves the existence of such an r_2.

We now let player II play a strategy very similar to the backtracking strategy she used in the proof of Theorem 2.2, which we presently describe in detail. Let $x_0, x_1, \ldots$ denote the sequence of positions of the game. Assume that $x_0 \in \tilde{Z}_L$. For each $t \in \mathbb{N}$, let $j_t := \max\{j \in \mathbb{N} : j \le t,\, x_j \in \tilde{Z}_L\}$. Note that if $j_t < t$, then $x_{j_t} \in Z_L$. By Lemma 1.1 inside $\{z \in X : |z-1| \ge 3/4,\, |z+1| \ge 3/4\}$ the function w satisfies $\Delta_\infty w = 0$. If $x_t \in \tilde{Z}_L$ and player II gets the turn, she moves to some $z \in B_{x_t}(\varepsilon)$ that satisfies $w(z) < \inf\{w(z') : z \in B_{x_t}(\varepsilon)\} + \delta\, 2^{-t-1}$. While if $x_t \notin \tilde{Z}_L$ and player II gets the turn, she moves to any neighbor z of x_t that is closer (in the graph metric) to x_{j_t} than x_t in the subgraph G_t of (X, E_ε) spanned by the vertices $x_{j_t}, x_{j_t+1}, \ldots, x_t$.

Now consider the game evolving under any strategy for player I and the above strategy for player II. Let d_t be the graph distance from x_t to x_{j_t} in G_t. Set

$$m_t := w(x_{j_t}) + d_t\, \varepsilon\, L + \delta\, 2^{-t} + \sum_{k=0}^{t} f(x_k)\,.$$

As in the proof of Theorem 2.2, it is easy to verify that $w(x_t) \le m_t$ and that m_t is a supermartingale.

Let τ be the time in which the game stops; that is, $\tau := \inf\{t \in \mathbb{N} : x_t \in Y\}$. Assume that $\tau < \infty$ a.s. Suppose that we could apply the optional stopping time theorem. Then we would have

$$w(x_0) \ge \mathbb{E}\big[m_\tau\big] = \mathbb{E}[w(x_{j_\tau}) + \delta\, 2^{-\tau}] + \varepsilon\, L\, \mathbb{E}[d_\tau] + \mathbb{E}\big[\text{payoff}\big].$$

Thus,

$$\mathbb{E}\big[\text{payoff}\big] \le O(1) - \varepsilon\, L\, \mathbb{E}[d_\tau]\,.$$

But $d_\tau \ge (R - r_2)/\varepsilon$, because the Euclidean distance from Y to $\tilde{Z}_L$ is at least $R - r_2$. Since $w(x_0)$ is at most the supremum of expected payoffs under the current strategy for player II and an arbitrary strategy guaranteeing $\tau < \infty$ for player I, we obtain $w(x_0) \le O(1) - R + r_2$. This contradicts our previous conclusion that $\|w\|_\infty = O(1)$, because we may choose r large.

How do we justify the application of the optional stopping time theorem? We slightly modify the strategy for player II. So far, our analysis utilized one advantage to player II in the calculation of $w = u_{\mathrm{I}}^{\varepsilon}$; that is, that it is player I's responsibility to make sure that $\tau < \infty$. Now we have to use the other advantage to player II, which is that player II gets to choose her strategy after knowing what player I's strategy is. Let t_0 be the first time such that with the strategy which player I is using and the above strategy for player II, the game terminates by time t_0 with probability at least $1/2$. The new strategy for player II is to play the above strategy until time t_0, and if the game lasts longer, to use an arbitrary strategy which would guarantee a conditioned expected future payoff of $u_{\mathrm{II}}^{\varepsilon}(x_{t_0}) = O(1)$. We may certainly apply the optional stopping time theorem at time $\tau \wedge t_0$. This does give a contradiction as above, because $\mathbb{E}[d_{\tau \wedge t_0}] \ge (R - r_2)/(2\,\varepsilon)$, and the contradiction proves our claim (5.1).

5.4. **Lipschitz g on unit disk with multiple solutions to $\Delta_\infty u = g$.** Here we construct the counterexample showing that if we omit the assumption $\inf f > 0$, in Corollary 1.9, then it may fail. In the example, X is the unit disk in $\mathbb{R}^2$, Y is its boundary, $F = 0$, and f is Lipschitz and take values of both signs. The example is motivated by and similar to the example of Section 5.3. However, since the assumptions of Corollary 1.9 do not hold (obviously), we need to construct the solutions u to $\Delta_\infty u = g$ not by using tug-of-war, but by other means. In fact, we will use a smoothing of Aronsson's function (4.1).

The following analytic-geometric lemma will replace the use of the backtracking strategy for player II in Section 5.3. The following notation will be needed. For $x \in U$ and $u : R \to \mathbb{R}$ let

$$\operatorname{Lip}_x u := \inf\{\operatorname{Lip}_W u : x \in W \subset U,\, W \text{ open}\}\,.$$

Lemma 5.5. *Let X be a length space, and let F_1, F_2 be bounded Lipschitz real-valued functions on a non-empty subset $Y \subset X$. Let u_1 and u_2 be the corresponding AM extensions to $X \setminus Y$. Fix $L > 0$, and let $Z_L := \{x \in X : \operatorname{Lip}_x u_1 > L\}$. Suppose that $F_1 = F_2$ on $Z_L \cap Y$, that $\operatorname{Lip}_{Y \setminus Z_L} F_2 \le L$ and*

$$\sup\left\{\frac{|F_2(y) - u_1(x)|}{d(y,x)} : y \in Y \setminus Z_L,\, x \in Z_L\right\} \le L\,. \tag{5.2}$$

Then $u_1 = u_2$ in Z_L.

We need a few simple observations before we begin the proof. Suppose that U is an open connected subset of a length space X, that $X \setminus U$ is non-empty, and that $F : X \setminus U \to \mathbb{R}$ is Lipschitz and bounded. Let $\tilde{\partial} U$ denote the set of all points $x \in \partial U$ such that there is a finite length path $\gamma \subset X$ where $\gamma \cap U \neq \emptyset$ and $\gamma \setminus U = \{x\}$. Set $U_* := U \cup \tilde{\partial} U$, and for two points $x, y \in U_*$ let $d_*(x,y) = d_*^U(x,y)$ be the infimum length of paths $\gamma \subset U_*$ joining x and y such that $\gamma \cap \tilde{\partial} U$ is finite. Note that (U_*, d_*) is a length space. Also, since $d_* \ge d$ on $U_* \times U_*$, the restriction of F to $\tilde{\partial} U$ is Lipschitz. Consequently, this restriction of F has an AM extension $u : U_* \to \mathbb{R}$ with respect to d_*. We claim that this is also an AM extension of F with respect to d.

Observe that for any Lipschitz function $w : V \to \mathbb{R}$ where $V \subset X$ is open, we have

$$\operatorname{Lip}_V w \le \Big(\sup_{x \in V} \operatorname{Lip}_x w\Big) \vee \big(\operatorname{Lip}_{\partial V} w\big)\,.$$

(This can be verified by considering for every pair of distinct points $x, y \in V$ the restriction of w to a nearly shortest path joining x and y in X.) Hence, to show that u is AM with respect to d it is enough to prove that

$$\sup_{x \in V} \operatorname{Lip}_x w \le \operatorname{Lip}_{\partial V} w \tag{5.3}$$

holds for arbitrary open $V \subset U$. Observe that for $x \in U$ we have $\operatorname{Lip}_x w = \operatorname{Lip}_x^* w$, where Lip^* refers to d_*. Since w is AM with respect to d_*, we have $\operatorname{Lip}_x^* w \le \operatorname{Lip}_{\partial V}^* w$ for all $x \in V$. Finally, $\operatorname{Lip}_{\partial V}^* w \le \operatorname{Lip}_{\partial V} w$ because $d_* \ge d$ on $U_* \times U_*$. This proves (5.3) and thereby shows that w is AM on U with respect to d. In particular, we conclude that

$$\operatorname{Lip}_{U_*}^* w = \operatorname{Lip}_{\tilde{\partial} U}^* w \tag{5.4}$$

for the AM extension of F to U.

Proof of Lemma 5.5. First, it is clear that Z_L is a closed set. Let

$$F_*(x) = \begin{cases} u_1(x)\,, & x \in Z_L\,, \\ F_2(x)\,, & x \in Y \setminus Z_L\,, \end{cases}$$

and let $w : X \to \mathbb{R}$ be the AM extension of F_* to $X \setminus (Y \cup Z_L)$. We claim that w is also an AM extension of $F_2 : Y \to \mathbb{R}$ to $X \setminus Y$. Since uniqueness of AM extensions holds in this setting, this will imply $w = u_2$ and complete the proof, because $w = u_1$ on Z_L.

Let U be a connected component of $X \setminus Z_L$, and let $d_* = d_*^U$ denote the corresponding metric on $U \cup \tilde{\partial} U$. Note that $\mathrm{Lip}^*_{\tilde{\partial} U} u_1 \le L$, since $\mathrm{Lip}_x u_1 \le L$ for $x \in X \setminus Z_L$. By (5.2) and the assumption $\mathrm{Lip}_{Y \setminus Z_L} F_2 \le L$, it follows that $\mathrm{Lip}^*_{\tilde{\partial} U \cup (Y \cap U)} F_* \le L$. Thus, we get $\mathrm{Lip}_x w = \mathrm{Lip}^*_x w \le L$ for $x \in U$, which implies $\mathrm{Lip}_x w \le L$ for $x \in X \setminus Z_L$.

Now let $V \subset X \setminus Y$ be open. In order to prove that w is AM, we need to establish (5.3). Without loss of generality, we assume that V is connected. If $V \cap Z_L = \emptyset$, then (5.3) certainly holds, since w is AM in $X \setminus (Z_L \cup Y)$.

Suppose now that $V \cap Z_L \neq \emptyset$. Then $L_1 := \mathrm{Lip}_V u_1 > L$, by the definition of Z_L. Since u_1 is AM, by (5.4) with u_1 in place of w, V in place of U and d_*^V in place of d_*^U, there is a sequence of pairs (x_j, y_j) in $\tilde{\partial} V$ and a sequence of paths $\gamma_j \subset V \cup \tilde{\partial} V$ from x_j to y_j, respectively, such that $x_j \neq y_j$,

$$\lim_{j \to \infty} \frac{|u_1(x_j) - u_1(y_j)|}{\mathrm{length}(\gamma_j)} = L_1\,, \tag{5.5}$$

and $\gamma_j \cap \tilde{\partial} V$ is finite. For all sufficiently large j, $|u_1(x_j) - u_1(y_j)| > L\,\mathrm{length}(\gamma_j)$. Hence $\gamma_j \cap Z_L \neq \emptyset$. Let x_j' be the first point on γ_j that is in Z_L, and let y_j' be the last point on γ_j that is in Z_L. Note that $d(x_j', x_j)/\,\mathrm{length}(\gamma_j) \to 0$ as $j \to \infty$, because $L < L_1 = \mathrm{Lip}_V u_1$, (5.5) holds and $|u(x_j') - u(x_j)| \le L\, d(x_j', x_j)$. (If at some significant proportion of the length of γ_j the function u does not change in speed very close to L_1, it will not have enough distance to catch up.) Similarly, $d(y_j', y_j)/\,\mathrm{length}(\gamma_j) \to 0$. Since $\sup_{x \notin Z_L} \mathrm{Lip}_x u_1 \le L$ and $\sup_{x \notin Z_L} \mathrm{Lip}_x w \le L$ and $u_1 = w$ on Z_L, it follows that $|u_1(x_j) - w(x_j)| \le 2\, L\, d(x_j, x_j')$ and similarly $|u_1(y_j) - w(y_j)| \le 2\, L\, d(y_j, y_j')$. Thus, $|w(x_j) - w(y_j)|/\,\mathrm{length}(\gamma_j) \to L_1$, proving that $\mathrm{Lip}_{\partial V} w \ge \mathrm{Lip}_V w$, as needed. □

Our example is based on smoothing two different modifications of Aronsson's ∞-harmonic function G from (4.1). For $L > 0$ let Z_L be the set of points in $\mathbb{R}^2 \setminus \{0\}$ such that $|\nabla G| > L$. Define the inner and outer radii of Z_L: $R_L := \sup\{|z| : z \in Z_L\}$ and $r_L := \inf\{|z| : 0 \neq z \notin Z_L\}$, and note that $\lim_{L \to \infty} R_L = 0$, while $r_L > 0$ for every $L > 0$. Now fix some large L. Let u denote G restricted to the annulus $A_L := \{z \in \mathbb{R}^2 : r_L/2 \le |z| \le 1\}$. Let u_j, $j = 1, 2$, denote the AM function on A_L whose boundary values are equal to u on the inner circle and are j on the outer circle, respectively. Lemma 5.5 implies that $u_1 = u = u_2$ on $A_L \cap Z_L$ (provided that L was chosen sufficiently large). Shortly, we will show that there is a function v defined on the unit disk that agrees with u in A_L and such that $\Delta_\infty v = g$ (in the viscosity sense), where g is a Lipschitz function. Assuming this for the moment, we can then define $v_j = u_j - j$ in A_L and $v_j = v - j$ inside the disk $|z| < r_L/2$. Then v_1 and v_2 are both 0 on the unit circle, and both satisfy $\Delta_\infty v_j = g$, providing the

required example. It thus remains to prove

Lemma 5.6. *For every $r_0 > 0$ there are Lipschitz functions $g, v : \mathbb{R}^2 \to \mathbb{R}$ such that v agrees with G on $|z| > r_0$ and $\Delta_\infty v = g$ holds in the viscosity sense.*

Proof. Let

$$a(\theta) := \left[\frac{\cos\theta\,(1-|\tan(\theta/2)|^{4/3})^2}{1+|\tan(\theta/2)|^{4/3}+|\tan(\theta/2)|^{8/3}}\right]^{1/3}$$

be the angle factor of G in (4.1). First, observe that $a(\theta+\pi) = -a(\theta)$ and $a(-\theta) = a(\theta)$. Next, note that $a(\theta) > 0$ when $\theta \in (-\pi/2, \pi/2)$. We now verify that $a(\theta)$ is C^∞ in a neighborhood of $\pi/2$. Write

$$a(\theta) = \cos\theta\left(1+|\tan(\theta/2)|^{4/3}+|\tan(\theta/2)|^{8/3}\right)^{-1/3}\left(\frac{1-|\tan(\theta/2)|^{4/3}}{\cos\theta}\right)^{2/3}.$$

The first two factors are real-analytic near $\pi/2$. Setting $\varphi = |\tan(\theta/2)|^2/3$, which is real-analytic near $\theta = \pi/2$, allows us to rewrite the last factor as

$$\begin{aligned}\big(\cos(\theta/2)\big)^{-4/3}\left(\frac{1-|\tan(\theta/2)|^{4/3}}{1-\tan^2(\theta/2)}\right)^{2/3} &= \big(\cos(\theta/2)\big)^{-4/3}\left(\frac{1-\varphi^2}{1-\varphi^3}\right)^{2/3}\\ &= \big(\cos(\theta/2)\big)^{-4/3}\left(\frac{1+\varphi}{1+\varphi+\varphi^2}\right)^{2/3},\end{aligned}$$

which is also real-analytic near $\theta = \pi/2$. We conclude that $a(\theta)$ is real-analytic at every $\theta \notin \pi\,\mathbb{Z}$.

It is instructive to see that G is ∞-harmonic in $\mathbb{R}^2 \setminus \{0\}$. Verifying $\Delta_\infty G = 0$ away from the real line can be done by differentiation. At points where $\theta = 0, r > 0$, we have $\nabla G/|\nabla G| = (-1, 0)$. At such points $\partial_x^2 G = \partial_x^2 r^{-1/3} = 4\,r^{-7/3}/9$, and indeed $\Delta_\infty^+ G = 4\,r^{-7/3}/9$ there. However,

$$a(\theta) = 1 - 16^{-1/3}\,\theta^{4/3} - \theta^2/6 + O(\theta^3)\,, \tag{5.6}$$

near $\theta = 0$. This implies that there is no C^∞ function u that satisfies $u(r,0) = G(r,0)$ and $u \le G$ in a neighborhood of $(r,0)$. Thus, $\Delta_\infty^- G = -\infty$.

We return to the proof of the lemma, and first establish that the lemma holds when we only require that g be continuous, instead of Lipschitz. In this case, the function v may be written as

$$v(r,\theta) := \varphi(r)\,\big(\lambda(r)\,a(\theta) + (1-\lambda(r))\,\cos\theta\big) \tag{5.7}$$

for appropriately chosen functions φ and λ. The functions φ and λ will be C^∞, will satisfy $\lambda(r) = \varphi(r) = 0$ in a neighborhood of 0 and $\varphi(r) = r^{-1/3}$, $\lambda(r) = 1$ for all $r \ge r_0/2$. It follows that v is Lipschitz and that $\Delta_\infty v$ is C^∞ away from the x-axis and in a neighborhood of 0. It remains to check the behavior of $\Delta_\infty v$ at points on and close to the x-axis. Based on (5.6), we may estimate $\Delta_\infty v$ for $\theta \neq 0$ close to 0 and $r > 0$:

$$\Delta_\infty v = \varphi'' - \frac{4\,\lambda^3\,\varphi^3}{81\,r^4\,(\varphi')^2} + \frac{P(r,\lambda,\varphi,\varphi',\lambda')}{r^6\,(\varphi')^4}\,\theta^{2/3} + O(\theta)\,, \tag{5.8}$$

where P is some polynomial. Let r_1, r_2, r_3 satisfy $0 < r_3 < r_2 < r_1 < r_0/2$. Choose $\lambda(r)$ as a C^∞ function that is 0 for $r \le r_2$, is 1 for $r \ge r_1$, and is strictly monotone in $[r_2, r_1]$. Choose $\varphi(r)$ as a C^∞ function that is 0 in $[0, r_3]$ and is $r^{-1/3}$ for $r \ge r_2$. When $r < r_2$, $\lambda = 0$, and hence $\Delta_\infty v$ is C^∞ by (5.7). In the range $r \in [r_2, r_1]$, we use (5.8) to conclude that the limit as $\theta \to 0$ of $\Delta_\infty v$ exists.

We now argue that there is a continuous function g such that $\Delta_\infty v = g$ in the viscosity sense. Suppose that η is a C^2 function defined in a small open set U containing the point $z_0 = (r, 0)$ such that the minimum of $\eta - v$ in U occurs at z_0. We will now perturb η. For sufficiently small $s_1 > 0$ the function $\eta_1(z) := \eta + s_1 |z - z_0|^2$ is a C^2 function defined in U and z_0 is the unique minimum of $\eta_1 - v$ in U. If $s_2 > 0$ is chosen much smaller, then the function $\eta_2(z) = \eta_1(z) + s_2\, y$ will still satisfy that the infimum of $\eta_2 - v$ is attained in U. Moreover, that infimum cannot be attained on the x-axis, because $s_2\, y$ is zero on the x-axis, ∇y is not parallel to the x-axis, and ∇v exists. Since s_2 and s_1 are arbitrarily small, we conclude that $\Delta_\infty \eta(z_0) \geq \lim_z \Delta_\infty v(z)$ as z tends to z_0 through a point not on the x-axis. Thus, $\Delta_\infty^+ v(z_0) \geq \lim_z \Delta_\infty v(z)$. Similarly, $\Delta_\infty^- v(z_0) \leq \lim_z \Delta_\infty v(z)$. This proves $\Delta_\infty v = g$, where g is the continuous extension of $\Delta_\infty v$ off the x-axis to the whole plane.

If we want g to be Lipschitz, we need to eliminate the $\theta^{2/3}$ term in (5.8). Define

$$\begin{aligned} a_1(\theta) &:= -16^{1/3}\big(a(\theta) - \cos\theta - (1/12)\, \sin^2 2\theta\big)\,, \\ a_2(\theta) &:= \cos\theta + (1/8)\, \sin^2 2\theta\,, \\ a_3(\theta) &:= (1/4)\, \sin^2 2\theta\,. \end{aligned}$$

Then near $\theta = 0$ we have

$$\begin{aligned} a_1(\theta) &= \theta^{4/3} + O(\theta^3)\,, \\ a_2(\theta) &= 1 + O(\theta^3)\,, \\ a_3(\theta) &= \theta^2 + O(\theta^3)\,. \end{aligned}$$

Set

$$v(r, \theta) := \lambda_1(r)\, a_1(\theta) + \lambda_2(r)\, a_2(\theta) + \lambda_3(r)\, a_3(\theta)\,, \tag{5.9}$$

where λ_1, λ_2 and λ_3 are to be chosen soon. A calculation shows that near $\theta = 0$ we have

$$\Delta_\infty v = \lambda_2'' + \frac{64\, \lambda_1^3}{81\, r^4\, (\lambda_2')^2}$$
$$-16\, \lambda_1 \frac{64\, \lambda_1^4 - 162\, r^4\, \lambda_1' (\lambda_2')^3 + 27\, r^2\, \lambda_1\, (\lambda_2')^2\, \big(3\, r^2\, \lambda_2'' + 3\, r\, \lambda_2' - 10\, \lambda_3\big)}{729\, r^6\, (\lambda_2')^4}\, \theta^{2/3} + O(\theta)\,.$$

Of course, when λ_1, λ_2 and λ_3 are chosen so that $v = G$, all the terms on the right hand side vanish. Our goal is to choose these functions so that the following holds:

(1) for $r \geq r_0$ we have $v = G$,
(2) when r is sufficiently small, we have $v = 0$,
(3) the $\theta^{2/3}$ term vanishes identically,
(4) we don't have a blow up due to $\lambda_2' = 0$, and finally
(5) the functions λ_j are C^∞, say.

This is not hard. Fix r_1, r_2, r_3, r_4 such that $0 < r_4 < r_3 < r_2 < r_1 < r_0/2$. For $r > r_1$, we choose the λ_j so that $v = G$. In particular, $\lambda_2 = r^{-1/3}$ in this range. In the range $r \geq r_3$ we maintain $\lambda_2 = r^{-1/3}$. The function λ_1 is chosen so that throughout $[r_3, r_2]$ we have $\lambda_1 = \lambda_2'$, while λ_1 does not vanish in $[r_3, r_1]$. This is possible, because $\lambda_2' < 0$ for $r \geq r_3$ and $\lambda_1 = -16^{-1/3} < 0$ at $r = r_1$. For every $r \in [r_3, r_1]$, $\lambda_3(r)$ takes the unique value for which the $\theta^{2/3}$ term vanishes. Since λ_1 and λ_2' do not vanish in the interval, it is clear that such a choice for λ_3 is possible

and λ_3 is C^∞ provided that λ_1 and λ_2 are also C^∞. Throughout $r < r_2$ we take $\lambda_1 = \lambda_2'$. This allows us to simplify the expression for $\Delta_\infty v$ in this range, to obtain

$$\Delta_\infty v = \lambda_2'' + \frac{64\,\lambda_2'}{81\,r^4} + 16\,\frac{270\,r^2\,\lambda_3 - (64+81\,r^3)\,\lambda_2' + 81\,r^4\,\lambda_2''}{729\,r^6}\,\theta^{2/3} + O(\theta)\,.$$

Now the denominators cannot vanish before $r = 0$. Thus, λ_3 takes care to make the $\theta^{2/3}$ term vanish while λ_2 evolves to become zero throughout $r \le r_4$. This completes the proof. □

6. Viscosity solutions and quadratic comparison in $\mathbb{R}^n$

In this section we prove Theorem 1.7, which states that in bounded domains in Euclidean space $\mathbb{R}^n$, u is a viscosity solution of $\Delta_\infty u = g$ iff u satisfies g-quadratic comparison. The following lemma will be useful in that proof. Let $D^2\varphi(x)$ denote the Hessian matrix $\left(\partial_i\partial_j\varphi(x)\right)_{j=1,\dots,n}^{i=1,\dots,n}$.

Lemma 6.1. *Let φ be any real-valued function which is C^2 in a neighborhood of $x_0 \in \mathbb{R}^n$ and satisfies $\nabla\varphi(x_0) \neq 0$. Then for every $\delta > 0$, there exist quadratic distance functions φ_1 and φ_2 such that:*

1. *$\varphi_1(x_0) = \varphi(x_0) = \varphi_2(x_0)$; $\nabla\varphi_1(x_0) = \nabla\varphi(x_0) = \nabla\varphi_2(x_0)$.*
2. *$0 < |\Delta_\infty\varphi_i(x_0) - \Delta_\infty\varphi(x_0)| < \delta$ for $i \in \{1,2\}$. Also, as quadratic forms, $D^2\varphi_2(x_0)$ strictly dominates $D^2\varphi(x_0)$, which in turn strictly dominates $D^2\varphi_1(x_0)$.*
3. *Consequently, $\varphi_1 < \varphi < \varphi_2$ at all points $x \neq x_0$ in a neighborhood $\widetilde{U}$ of x_0.*
4. *φ_1 and φ_2 are centered at z_1 and z_2, respectively, with $0 < |z_i - x_0| < \delta$ for $i \in \{1,2\}$.*
5. *In a neighborhood of x_0, the function φ_1 is $\star$-decreasing in the distance from z_1 and φ_2 is $\star$-increasing in the distance from z_2.*

Proof. We begin with preliminary calculations about the quadratic distance function $\varphi_{a,b}(x) = a|x|^2 + b|x|$, centered at the origin. Fix some $x \neq 0$. Let $v = x/|x|$ be the unit vector in the x-direction and v_O be any unit vector orthogonal to v. Write $M = M(x) = D^2\varphi_{a,b}(x)$.

Then direct calculations show:

1. $\nabla\varphi_{a,b}(x) = (2a|x| + b)\,v$.
2. $v^T M v = 2a$.
3. $v_O^T M v_O = 2a + b|x|^{-1}$.
4. $v_O^T M v = 0$.
5. $\Delta_\infty\varphi_{a,b}(x) = 2a$ whenever $\nabla\varphi_{a,b}(x) \neq 0$.

Of course, it is enough to complete this calculation in dimension two when $x = (1,0)$ and $a = 1$ and to deduce the general case by symmetry.

We now construct the φ_2 described in the lemma (the construction of φ_1 is similar). We will take $z_2 = x_0 - \gamma\frac{\nabla\varphi(x_0)}{|\nabla\varphi(x_0)|}$ for some small value of $\gamma > 0$ (specified below) and write $\varphi_2(x) = \varphi_{a,b}(x - z_2) + c$, where we first choose a to be an arbitrary real with the property that $0 < 2a - \Delta_\infty\varphi(x_0) < \delta$, we then choose b so that $\nabla\varphi_2(x_0) = \nabla\varphi(x_0)$ and we then choose c so that $\varphi_2(x_0) = \varphi(x_0)$. We must now show that the φ_2 thus constructed satisfies the requirements of the lemma.

We can compute b explicitly in terms of a, x_0, $\nabla\varphi(x_0)$, γ, and z_2 using the relation $\nabla\varphi_2(x_0) = (2a\gamma + b)\frac{x_0 - z_2}{|x_0 - z_2|} = \nabla\varphi(x_0)$. As γ tends to zero, b tends to $|\nabla\varphi(x_0)|$.

The description of $D^2\varphi_{a,b}(x)$ given above implies that if $v = (x_0 - z_2)/|x_0 - z_2|$ and v_O is a unit vector orthogonal to v, and $M = D^2\varphi_{a,b}(x_0)$, then

(1) $v^T M v = 2a$.
(2) $v_O^T M v_O = 2a + b\gamma^{-1}$.
(3) $v_O^T M v = 0$.

Note that $\chi := v_O^T M v_O = 2a + b\gamma^{-1} = \gamma^{-1}|\nabla\varphi(x_0)|$ tends to ∞ as $\gamma \searrow 0$. Let $M' := D^2\varphi(x_0)$. We claim that by choosing γ sufficiently small we can make sure that M dominates M' as a quadratic form (or any other fixed quadratic form satisfying $v^T M' v < 2a$, for that matter). Let $C := \sup\{w_0^T M' w_1 : w_0, w_1 \in \mathbb{R}^n, |w_0| = |w_1| = 1\}$ and $a' := v^T M' v/2 < a$. If $w \in \mathbb{R}^n$ is non-zero, we may write $w = \alpha\, v + w_O$, where $\alpha \in \mathbb{R}$ and w_O is orthogonal to v. Then

$$w^T M' w \le 2\, a'\, \alpha^2 + 2\, C\, \alpha\, |w_O| + C\, |w_O|^2.$$

On the other hand $w^T M w = 2\, a\, \alpha^2 + \chi\, |w_O|^2$. Since $2\, C\, \alpha\, |w_O| \le \frac{a-a'}{2}\, \alpha^2 + C'\, |w_O|^2$ for some constant $C' = C'(C, a, a')$, the domination follows from $\lim_{\gamma \searrow 0} \chi = \infty$. This constructs φ_2 with the required properties. A similar construction applies to φ_1. □

Proof of Theorem 1.7. Let $x_0 \in U$ and suppose that u satisfies g-quadratic comparison in a neighborhood of x_0. Let φ be a C^2 real-valued function defined in a neighborhood of x_0. Suppose that $\nabla\varphi(x_0) \neq 0$ and $\varphi - u$ has a local minimum at x_0. The above lemma implies that for every $\delta > 0$ there is a quadratic distance function $\varphi_2(x) = a\,|x - z|^2 + b\,|x - z|$ such that $\varphi_2(x) - \varphi(x) > \varphi_2(x_0) - \varphi(x_0)$ for all $x \neq x_0$ in some neighborhood of x_0, $\Delta_\infty\varphi_2 < \Delta_\infty\varphi + \delta$ and φ_2 is $\star$-increasing in a neighborhood of x_0. Since $\varphi_2 - u$ has a strict local minimum at x_0, it follows by g-quadratic comparison that arbitrarily close to x_0 there are points x for which $a > g(x)/2$. By continuity of g, this implies $a \ge g(x_0)/2$. That is, $\Delta_\infty\varphi_2(x_0) \ge g(x_0)$. Since $\delta > 0$ was arbitrary, this also implies $\Delta_\infty\varphi(x_0) \ge g(x_0)$.

Now we remove the assumption that $\nabla\varphi(x_0) \neq 0$ and assume instead that $\varphi(x) = a\,|x - x_0|^2 + o(|x - x_0|^2)$ as $x \to x_0$. If $a \ge 0$, we may take $\varphi_2(x) = (a + \delta)\,|x - x_0|^2$, and the same argument as above gives $\Delta_\infty\varphi(x_0) \ge g(x_0)$. Now suppose $a < 0$ and let $\delta \in (0, |a|)$. In this case, we have to modify the argument, because φ_2 is not $\star$-increasing in a neighborhood of x_0. Recall that $\varphi - u$ has a local minimum at x_0 and $\varphi_2 - \varphi$ has a strict local minimum at x_0. Therefore, $\varphi_2 - u$ has a strict local minimum at x_0. Let $r > 0$ be sufficiently small so that $\varphi_2(x) - u(x) > \varphi_2(x_0) - u(x_0) = -u(x_0)$ on $\{x \in \mathbb{R}^n : 0 < |x - x_0| \le r\}$. By continuity, for every $\eta \in \mathbb{R}^n$ with $|\eta|$ sufficiently small $\varphi_2(x + \eta) - u(x) > -u(x_0)$ for every $x \in \partial B_{x_0}(r)$. Fix such an η satisfying $0 < |\eta| < r$, and let x_* be a point in $\overline{B_{x_0}(r)}$ where $\varphi_2(x + \eta) - u(x)$ attains its minimum. Since

$$\varphi_2(x_0 + \eta) - u(x_0) < -u(x_0) \le \inf\big\{\varphi_2(x + \eta) - u(x) : x \in \partial B_{x_0}(r)\big\},$$

it follows that $x_* \notin \partial B_{x_0}(r)$. On the other hand, $x_* \neq x_0 - \eta$, because

$$\varphi_2\big((x_0 - \eta) + \eta\big) - u(x_0 - \eta) = -u(x_0 - \eta) \ge -\varphi_2(x_0 - \eta) - u(x_0) > -u(x_0)\,.$$

Therefore, $\nabla\varphi_2(x_*) \neq 0$. The first case we have considered therefore gives $g(x_*) \le \Delta_\infty\varphi_2(x_*) = 2\,(a + \delta)$. Since $\delta \in (0, |a|)$ was arbitrary, continuity gives $g(x_0) \le 2\,a$, as required. Thus, we conclude that $\Delta_\infty^+ u \ge g$. By symmetry, $\Delta_\infty^- u \le g$, and hence u is a viscosity solution as claimed.

This completes the first half of the proposition. For the converse, suppose that u is a viscosity solution of $\Delta_\infty u = g$. Let $W \subset \overline{W} \subset U$ be open. Suppose that $\varphi(x) = a\,|x-z|^2 + b\,|x-z| + c$ is a $\star$-increasing quadratic distance function in W, $\varphi > u$ on ∂W and there is some $x_0 \in W$ such that $u(x_0) > \varphi(x_0)$. For all sufficiently small $\delta > 0$ we still have $\varphi_\delta(x) := \varphi(x) - \delta\,|x-z|^2 > u$ on ∂W. Let x_* be a point in $\overline{W}$ where $\varphi_\delta(x) - u(x)$ attains its minimum. Note that $x_* \notin \partial W$, since $\varphi_\delta(x_0) - u(x_0) \le \varphi(x_0) - u(x_0) < 0$. Consequently, $g(x_*) \le \Delta_\infty^+ u(x_*) \le \Delta_\infty \varphi_\delta(x_*) = 2\,(a-\delta)$. Thus, $2\,a > \inf_W g$. Therefore, u satisfies g-quadratic comparison on W, and the proof is complete. □

7. Limiting trajectory

7.1. **A general heuristic.** For every ε, when both players play optimally in ε-tug-of-war, the sequence $\{x_k\}$ of points visited is random. Do the laws of these random sequences, properly normalized, converge in some sense to the law of a random continuous path as ε tends to zero?

We give a complete answer in only a couple of simple cases. However, we can more generally compute the limiting trajectory when u is C^2 in a domain contained in $\mathbb{R}^n$ and the players move to maximize/minimize u instead of u^ε; we suspect but cannot generally prove that the limiting behavior will be the same when the players use u^ε.

Consider a point x_0 in the domain at which $\nabla u(x_0) \neq 0$. If ε is small enough, then $\nabla u \neq 0$ throughout the closed ball $\overline{B}_\varepsilon(x_0)$, so the extrema of u on the closed ball $\overline{B}_\varepsilon(x_0)$ lie on the surface of the ball. Then at any such extremum x, by the Lagrange multipliers theorem, $x - x_0 = \lambda \nabla u(x)$ for some real $\lambda = O(\varepsilon)$. Since u is C^2, we have $\nabla u(x) = \nabla u(x_0) + D^2 u(x_0)(x-x_0) + o(\varepsilon)$, where $D^2 u(x_0) = (\partial_i \partial_j u(x_0))_{j=1,\dots,n}^{i=1,\dots,n}$. We define $\eta = \nabla u(x_0)$, $H = D^2 u(x_0)$, and $c = \eta^T H \eta / |\eta|^2$. Then $x - x_0 = \lambda(\eta + H(x-x_0) + o(\varepsilon))$, where $\lambda = O(\varepsilon)$. If we solve this with small λ, we find

$$x - x_0 = (I - \lambda H)^{-1} \lambda(\eta + o(\varepsilon)) = \lambda\eta + \lambda^2 H\eta + o(\varepsilon^2).$$

Then $\varepsilon^2 = |x-x_0|^2 = |\eta|^2\lambda^2 + 2c|\eta|^2\lambda^3 + o(\varepsilon^3)$, and thus $\pm\varepsilon = |\eta|\lambda + c|\eta|\lambda^2 + o(\varepsilon^2)$, and so $\lambda = \pm\varepsilon/|\eta| - \varepsilon^2 c/|\eta|^2 + o(\varepsilon^2)$. Hence,

$$x - x_0 = \pm\varepsilon|\eta|^{-1}\eta + \varepsilon^2|\eta|^{-2}(H - cI)\eta + o(\varepsilon^2).$$

When u is C^2 and has non-zero gradient in a domain, this suggests the SDE

$$dX_t = r(X_t)\,dB_t + s(X_t)\,dt\,, \tag{7.1}$$

where $r(X_t) = |\nabla u(X_t)|^{-1}\nabla u(X_t)$ and $s(X_t)$ is equal to $|\nabla u(X_t)|^{-2} D^2 u(X_t) \nabla u(X_t)$ minus its projection onto $\nabla u(X_t)$ (so that $r(X_t)$ and $s(X_t)$ are always orthogonal). Now Itô's formula implies that, as expected, $u(X_t) - \frac{1}{2}\int_{s=0}^t \Delta_\infty u(X_s)\,ds$ is a martingale. In particular, when u is infinity harmonic, $u(X_t)$ is a martingale.

All of the above analysis applies only when players make moves to optimize u instead of u^ε, as they would do if they were playing optimally. This difference is what makes the calculation of the limiting trajectory a heuristic, except in a few special cases as described in the next subsection.

7.2. **Special cases.** The above analysis does apply to optimally played games in a couple of simple cases for which $u = u^\varepsilon$. Let $X = \mathbb{R}^2$ and let $Y \subset X$ be the complement of a bounded set. Then the following are infinity harmonic functions u which satisfy both $\Delta_\infty u = 0$ and $\Delta_\infty^\varepsilon u = 0$:

(1) $u(v)$ is the distance from v to a fixed convex set whose ε-neighborhood is contained in Y.
(2) $u(v)$ is the argument of v on a $2\,\varepsilon$-neighborhood of $X \setminus Y$, and defined arbitrarily elsewhere (here we assume that the $2\,\varepsilon$-neighborhood of $X \setminus Y$ does not contain a simple closed curve surrounding 0, so that the argument can be defined there).

(Due to boundary errors the above need not hold when $X \setminus Y$ is a bounded domain and Y is its boundary.) In the first case, X_t is simply a Brownian motion along a (straight) gradient flow line of u. In the second case, X_t is a diffusion with drift in the $-X_t$ direction and diffusion of constant magnitude orthogonal to X_t.

Crandall and Evans [10] have explored the following question in some detail, and it has recently been answered affirmatively by Savin [23] when $n = 2$: is every ∞-harmonic function on a domain in $\mathbb{R}^n$ everywhere differentiable? This question can be rewritten as a question about the amount of variation of the optimal direction of the first move (as a function of the starting point) in ε-tug-of-war. An affirmative answer might be a step towards a more complete analysis of the limiting game trajectories when $f = 0$ and when u is not smooth, since it would at least ensure that $r(X_t)$ is well defined everywhere that the gradient is non-zero.

8. Additional open problems

(1) If $U \subset \mathbb{R}^n$ is open and bounded, $F : \partial U \to \mathbb{R}$ Lipschitz, and $g : \overline{U} \to \mathbb{R}$ Lipschitz, is there a unique viscosity solution for $\Delta_\infty u = g + c$ with boundary values given by F for generic $c \in \mathbb{R}$? Here, generic could mean in the sense of Baire category, or it could mean almost every, or perhaps this could be true except for a countable set of c.
(2) Does Theorem 1.8 continue to hold if F and f are merely continuous instead of uniformly continuous? Can the $\inf |f| > 0$ requirement be replaced with $f \geq 0$ (or $f \leq 0$) when X has finite diameter? When solving $\Delta_\infty u = g$ on bounded domains in $\mathbb{R}^n$, can the condition that g be continuous be replaced with a natural weaker condition (e.g., semicontinuity or piecewise continuity)?
(3) In Section 5.2 we gave a triple (X, Y, F) with F positive and Lipschitz for which the continuum value of tug-of-war is not the unique AM extension of F.

    Is there an example where X is the closure of a connected open subset of $\mathbb{R}^n$ and Y is its boundary? In particular, let X be the set of points in $\mathbb{R}^2$ above the graph of the function $|x|^{1+\delta}$, let Y be the boundary of X in $\mathbb{R}^2$, and let $F(x, y) = y$ on Y. Is y the only AM extension of F to X? What is the value of the corresponding game? (Note added in revision: Changyou Wang and Yifeng Yu (personal communication) have recently shown that y is the only AM extension of F to X. The proof turned out to be rather simple.)

(4) Suppose that $u : U \to \mathbb{R}$ is Lipschitz and $g_1, g_2 : U \to \mathbb{R}$ are continuous on an open set $U \subset \mathbb{R}^n$ and that $\Delta_\infty u = g_1$ as well as $\Delta_\infty u = g_2$, both in the viscosity sense. Does it follow that $g_1 = g_2$? (Note added in revision: Yifeng Yu [25] proved this in the case $n = 2$.)

Acknowledgements

We thank Alan Hammond and Gábor Pete for useful discussions and the referee for corrections to an earlier version.

References

1. Gunnar Aronsson, *Extension of functions satisfying Lipschitz conditions*, Ark. Mat. **6** (1967), 551–561. MR0217665 (36:754)
2. ———, *On the partial differential equation* $u_x^2 u_{xx} + 2u_x u_y u_{xy} + u_y^2 u_{yy} = 0$, Ark. Mat. **7** (1968), 395–425. MR0237962 (38:239)
3. ———, *Construction of singular solutions to the p-harmonic equation and its limit equation for* $p = \infty$, Manuscripta Math. **56** (1986), no. 2, 135–158. MR850366 (87j:35070)
4. Gunnar Aronsson, Michael G. Crandall, and Petri Juutinen, *A tour of the theory of absolutely minimizing functions*, Bull. Amer. Math. Soc. (N.S.) **41** (2004), no. 4, 439–505 (electronic). MR2083637 (2005k:35159)
5. M. Bardi, M. G. Crandall, L. C. Evans, H. M. Soner, and P. E. Souganidis, *Viscosity solutions and applications*, Lecture Notes in Mathematics, vol. 1660, Springer-Verlag, Berlin, 1997, Lectures given at the 2nd C.I.M.E. Session held in Montecatini Terme, June 12–20, 1995, edited by I. Capuzzo Dolcetta and P. L. Lions, Fondazione C.I.M.E. [C.I.M.E. Foundation]. MR1462698 (98b:35005)
6. G. Barles and Jérôme Busca, *Existence and comparison results for fully nonlinear degenerate elliptic equations without zeroth-order term*, Comm. Partial Differential Equations **26** (2001), no. 11-12, 2323–2337. MR1876420 (2002k:35078)
7. T. Bhattacharya, E. DiBenedetto, and J. Manfredi, *Limits as* $p \to \infty$ *of* $\Delta_p u_p = f$ *and related extremal problems*, Rend. Sem. Mat. Univ. Politec. Torino 1989 (1991). Special issue on topics in nonlinear PDEs, 15–68. MR1155453 (93a:35049)
8. Thierry Champion and Luigi De Pascale, *Principles of comparison with distance functions for absolute minimizers*, J. Convex Anal. **14** (2007), no. 3, 515–541. MR2341302
9. M. G. Crandall, L. C. Evans, and R. F. Gariepy, *Optimal Lipschitz extensions and the infinity Laplacian*, Calc. Var. Partial Differential Equations **13** (2001), no. 2, 123–139. MR1861094 (2002h:49048)
10. Michael G. Crandall and L. C. Evans, *A remark on infinity harmonic functions*, Proceedings of the USA-Chile Workshop on Nonlinear Analysis (Viña del Mar-Valparaiso, 2000) (San Marcos, TX), Electron. J. Differ. Equ. Conf., vol. 6, Southwest Texas State Univ., 2001, pp. 123–129 (electronic). MR1804769 (2001j:35076)
11. Michael G. Crandall and Pierre-Louis Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc. **277** (1983), no. 1, 1–42. MR690039 (85g:35029)
12. Lawrence C. Evans and Yifeng Yu, *Various properties of solutions of the infinity-Laplacian equation*, Comm. Partial Differential Equations **30** (2005), no. 7-9, 1401–1428. MR2180310 (2006g:35062)
13. Robert Jensen, *Uniqueness of Lipschitz extensions: minimizing the sup norm of the gradient*, Arch. Rational Mech. Anal. **123** (1993), no. 1, 51–74. MR1218686 (94g:35063)
14. Petri Juutinen, *Absolutely minimizing Lipschitz extensions on a metric space*, Ann. Acad. Sci. Fenn. Math. **27** (2002), no. 1, 57–67. MR1884349 (2002m:54020)
15. Andrew J. Lazarus, Daniel E. Loeb, James G. Propp, Walter R. Stromquist, and Daniel H. Ullman, *Combinatorial games under auction play*, Games Econom. Behav. **27** (1999), no. 2, 229–264. MR1685133 (2001f:91023)
16. Andrew J. Lazarus, Daniel E. Loeb, James G. Propp, and Daniel Ullman, *Richman games*, Games of No Chance (Richard J. Nowakowski, ed.), MSRI Publications, vol. 29, Cambridge Univ. Press, Cambridge, 1996, pp. 439–449. MR1427981 (97j:90101)
17. Donald A. Martin, *The determinacy of Blackwell games*, J. Symbolic Logic **63** (1998), no. 4, 1565–1581. MR1665779 (2000d:03118)

18. E. J. McShane, *Extension of range of functions*, Bull. Amer. Math. Soc. **40** (1934), no. 12, 837–842. MR1562984
19. V. A. Mil′man, *Absolutely minimal extensions of functions on metric spaces*, Mat. Sb. **190** (1999), no. 6, 83–110. MR1719573 (2000j:41041)
20. Abraham Neyman and Sylvain Sorin (eds.), *Stochastic games and applications*, NATO Science Series C: Mathematical and Physical Sciences, vol. 570, Dordrecht, Kluwer Academic Publishers, 2003. MR2032421 (2004h:91004)
21. Adam M. Oberman, *A convergent difference scheme for the infinity Laplacian: construction of absolutely minimizing Lipschitz extensions*, Math. Comp. **74** (2005), no. 251, 1217–1230 (electronic). MR2137000 (2006h:65165)
22. Yuval Peres, Oded Schramm, Scott Sheffield, and David B. Wilson, *Random-turn Hex and other selection games*, Amer. Math. Monthly **114** (2007), no. 5, 373–387. MR2309980 (2008a:91039)
23. Ovidiu Savin, C^1 *regularity for infinity harmonic functions in two dimensions*, Arch. Ration. Mech. Anal. **176** (2005), no. 3, 351–361. MR2185662 (2006i:35108)
24. Hassler Whitney, *Analytic extensions of differentiable functions defined in closed sets*, Trans. Amer. Math. Soc. **36** (1934), no. 1, 63–89. MR1501735
25. Yifeng Yu, *Uniqueness of values of Aronsson operators and applications to "tug-of-war" game theory*, 2007, `http://www.ma.utexas.edu/~yifengyu/uofa2d.pdf`.

Microsoft Research, One Microsoft Way, Redmond, Washington 98052, and Department of Statistics, 367 Evans Hall, University of California, Berkeley, California 94720
URL: `http://research.microsoft.com/~peres/`

Microsoft Research, One Microsoft Way, Redmond, Washington 98052
URL: `http://research.microsoft.com/~schramm/`

Microsoft Research, One Microsoft Way, Redmond, Washington 98052, and Department of Statistics, 367 Evans Hall, University of California, Berkeley, California 94720
Current address: Courant Institute, 251 Mercer Street, New York, New York 10012
URL: `http://www.cims.nyu.edu/~sheff/`

Microsoft Research, One Microsoft Way, Redmond, Washington 98052
URL: `http://dbwilson.com`

ELECTRONIC RESEARCH ANNOUNCEMENTS
IN MATHEMATICAL SCIENCES
Volume 15, Pages 17–23 (January 31, 2008)
S 1935-9179 AIMS (2008)

HYPERFINITE GRAPH LIMITS

ODED SCHRAMM

(Communicated by Krzysztof Burdzy)

ABSTRACT. Gábor Elek introduced the notion of a hyperfinite graph family: a collection of graphs is hypefinite if for every $\epsilon > 0$ there is some finite k such that each graph G in the collection can be broken into connected components of size at most k by removing a set of edges, of size at most $\epsilon |V(G)|$. We presently extend this notion to a certain compactification of finite bounded-degree graphs and show that if a sequence of finite graphs converges to a hyperfinite limit, then the sequence itself is hyperfinite.

1. INTRODUCTION

While studying asymptotic properties of finite graphs, it can be worthwhile to introduce compactifications of various collections of finite graphs. In one such compactification, which is particularly well suited to bounded degree graphs, the notion of convergence is local weak convergence, as studied in [6] and [2]. A different notion of convergence [12] is suitable for dense graphs, in which the number of edges is roughly proportional to $|V(G)|^2$. Here, we will focus on the bounded degree case.

The purpose of this note is to extend the study of the local weak convergence of graphs, whose definition will be recalled below, by noting its properties in relation to the collection of hyperfinite graphs. A collection of finite graphs $\mathcal{G}$ is (k, ϵ)-**hyperfinite** if every $G \in \mathcal{G}$ has a set of edges $S \subset E(G)$ such that $|S| \leq \epsilon |V(G)|$ and every connected component of $G \setminus S$ has at most k vertices. We say that $\mathcal{G}$ is **hyperfinite**, if for every $\epsilon > 0$ there is some finite $k = k(\epsilon)$ such that $\mathcal{G}$ is (k, ϵ)-hyperfinite. Many interesting collections of graphs are hyperfinite. For example, as noted in [13, Theorem 3], it follows from the planar separator theorem that the set of planar graphs with maximal degree at most M is hyperfinite for every $M < \infty$.

Although the notion of hyperfiniteness appears implicitly in the literature (for example, [13]), as far as we know Gábor Elek [8, 9, 10] was the first to give it a name and propose its systematic study.

We now prepare to introduce the notion of local weak convergence. A **rooted graph** is a pair (G, o), where $o \in V(G)$. An **isomorphism** of rooted graphs $\phi : (G, o) \to (G', o')$ is an isomorphism of the underlying graphs which satisfies $\phi(o) = o'$. A graph is **locally-finite** if each vertex is incident to only finitely many

Received by the editors December 27, 2007.
2000 *Mathematics Subject Classification.* Primary 60C05; Secondary 05C80.
Key words and phrases. Graph limits, generalizations of amenability.
We are grateful to Gábor Elek, Christian Borgs and Russell Lyons for beneficial conversations.

©2008 American Institute of Mathematical Sciences

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_19, © Springer Science+Business Media, LLC 2011

edges. Let $\widehat{\mathcal{G}}$ denote the collection of all isomorphism classes of connected, locally-finite rooted graphs. For $M \in \mathbb{N}_+$, let $\widehat{\mathcal{G}}_M \subset \widehat{\mathcal{G}}$ denote the subcollection consisting of rooted graphs with maximal degree at most M. For $(G,o) \in \widehat{\mathcal{G}}$ and $r \geq 0$, let $B_G(o,r)$ denote the subgraph of G spanned by the vertices at distance at most r from o. If $(G,o),(G',o') \in \widehat{\mathcal{G}}$ and r is the largest integer such that $\big(B_G(o,r),o\big)$ is rooted-graph isomorphic to $\big(B_{G'}(o',r),o'\big)$, then we set $\rho\big((G,o),(G',o')\big) = 1/r$. We also set $\rho\big((G,o),(G,o)\big) = 0$. Then ρ, thus defined, is a metric on $\widehat{\mathcal{G}}$. Let $\mathfrak{M}_M$ (respectively $\mathfrak{M}$) denote the space of all probability measures on $\widehat{\mathcal{G}}_M$ (resp. $\widehat{\mathcal{G}}$) that are measurable with respect to the Borel σ-field of ρ. Then $\mathfrak{M}_M$, endowed with the topology of weak convergence, is compact.

Let $\mathcal{G}^0_M$ denote the collection of all isomorphism classes of finite graphs with maximal degree at most M and let $G \in \mathcal{G}^0_M$. Let $o \in V(G)$ be chosen randomly-uniformly and let G_o be the connected component of o. Then the law (of the isomorphism type) of (G_o,o) is an element of $\mathfrak{M}_M$, which will be denoted by $\Psi(G)$. This defines a mapping, $\Psi : \mathcal{G}^0_M \to \mathfrak{M}_M$, whose restriction to the set of connected graphs is injective.

Let $\mathfrak{M}^0_M$ denote the closure of $\Psi(\mathcal{G}^0_M)$ in $\mathfrak{M}_M$. It has been observed in [6] that every $\mu \in \mathfrak{M}^0_M$ satisfies an intrinsic version of the Mass Transport Principle (MTP). The MTP was invented by Olle Häggström in the setting of regular trees [11], and has been used extensively by Benjamini, Lyons, Peres and Schramm in the more general setting of transitive and quasi-transitive graphs [4]. Since then, there have been numerous successful applications of the MTP. A measure, $\mu \in \mathfrak{M}$, satisfies the intrinsic MTP (iMTP) if for every non-negative function $f(G,x,y)$, which takes as arguments a connected graph and two vertices in it, and depends only on the isomorphism type of the triple (G,x,y), we have

$$\int \sum_{x \in V(G)} f(G,x,o)\, d\mu(G,o) = \int \sum_{y \in V(G)} f(G,o,y)\, d\mu(G,o)\,.$$

Let $\mathcal{U}$ denote the set of $\mu \in \mathfrak{M}$ satisfying the iMTP and set $\mathcal{U}_M := \mathcal{U} \cap \mathfrak{M}_M$. The easy proof that $\mathfrak{M}^0_M \subset \mathcal{U}$ proceeds by first noting that $\Psi(\mathcal{G}^0_M) \subset \mathcal{U}$ and then observing that $\mathcal{U}_M$ is closed in $\mathfrak{M}_M$.

Aldous and Lyons [2] coined the term **unimodular** for measures in $\mathcal{U}$, and studied their properties. They also raised the fundametal problem of determining whether $\mathfrak{M}^0_M = \mathcal{U}_M$; that is, whether every $\mu \in \mathcal{U}_M$ can be approximated by elements of $\Psi(\mathcal{G}^0_M)$.

If ν is a measure on triples (G,o,S), where G is a connected graph, $o \in V(G)$, and S is some structure on G (such as a labeling of the edges or vertices, or a subgraph, etc), then ν is **unimodular** if it satisfies the above iMTP, where f is allowed to depend on S.

A measure $\mu \in \mathcal{U}$ is (k,ϵ)**-hyperfinite** if there is a measure, ν, on triples, (G,o,S), such that the projection of ν to (G,o) is μ, $S \subset E(G)$, every connected component of $G \setminus S$ has at most k vertices, ν is unimodular and the ν-expected number of edges in S adjacent to o is at most $2\,\epsilon$.

The reason for choosing $2\,\epsilon$ instead of ϵ is the following. If G_0 is a finite graph which is (k,ϵ)-hyperfinite and $\mu = \Psi(G_0)$, then μ is (k,ϵ)-hyperfinite. This is because if $S_0 \subset E(G_0)$ satsifies $|S_0| \leq \epsilon |V(G_0)|$, then the uniform law on the triples (G_0,o,S_0), where $o \in V(G_0)$, is unimodular and the expected number of edges of

S_0 adjacent to o is $2\,|S_0|/|V(G)| \le 2\,\epsilon$. Conversely, we show in Lemma 2.1 below that if $G \in \mathcal{G}_M^0$ and $\Psi(G)$ is (k,ϵ)-hyperfinite, then G is also (k,ϵ)-hyperfinite.

A measure $\mu \in \mathcal{U}$ is **hyperfinite** if, for every $\epsilon > 0$, there is a finite $k = k(\epsilon)$ such that μ is (k,ϵ)-hyperfinite.

In this paper we will prove the following theorem.

Theorem 1.1. *Let $M \in \mathbb{N}_+$, and let $G_1, G_2, \dots$ be a sequence in $\mathcal{G}_M^0$ with $\Psi(G_j)$ converging weakly to some $\mu \in \mathfrak{M}_M$. Then μ is hyperfinite if and only if the sequence $\{G_1, G_2, \dots\}$ is hyperfinite.*

In the forthcoming paper [7] we give an application of this theorem to planarity testing.

The easy direction of the theorem is to show that μ is hyperfinite if $\{G_1, G_2, \dots\}$ is hyperfinite. The opposite direction is an immediate consequence of the following, more quantitive, version.

Theorem 1.2. *Suppose that $\mu \in \mathcal{U}_M$ is (k,ϵ)-hyperfinite, where $0 < \epsilon < 1$ and $k < \infty$. Then there is some open neighborhood of μ such that every $\mu' \in \mathcal{U}_M$ within that neighborhood is $(k,\tilde{\epsilon})$-hyperfinite, where*

$$\tilde{\epsilon} := 3\,\log(2\,M/\epsilon)\,\epsilon\,.$$

Theorem 1.2 is quantitative in the sense that the dependence of $\tilde{\epsilon}$ on ϵ is explicit. In [7], we present a finitary variant of Theorem 1.2 which bounds the size of the neighborhood. The proof there is also a finitary variant of the proof below.

I. Benjamini (private communication) points out that hyperfiniteness of measures in $\mathcal{U}_M$ is an appropriate analogue for amenability in the setting of transitive graphs. In particular, the Burton-Keane [3] argument can show that percolation on a sample from $\mathcal{U}_M$ can produce at most one infinite cluster a.s. Extending a conjecture from [5] in the transitive setting, Benjamini also conjectures the converse: that if $\mu \in \mathcal{U}_M$ is not hyperfinite, then there is some parameter $p \in [0,1]$ such that Bernoulli(p) percolation on the sample from μ has more than one infinite cluster, with positive probability. See [1] for a related discussion in the context of percolation on finite graphs and, in particular, the discrete hypercube.

2. Proofs

As a warm up, we prove the following.

Lemma 2.1. *Let $\epsilon > 0$, $k \in \mathbb{N}_+$ and $G \in \mathcal{G}_M^0$. Then G is (k,ϵ)-hyperfinite if and only if $\Psi(G)$ is (k,ϵ)-hyperfinite.*

Proof. The "only if" direction was sketched in the introduction, and hence we presently only prove the "if" direction. Assume that $\Psi(G)$ is (k,ϵ)-hyperfinite. Then there is a unimodular probability measure ν on triples (G', S, o) such that G' is ν-a.s. a connected component of G, $S \subset E(G')$, $o \in V(G')$, the ν-distribution of o is uniform in $V(G)$, the connected components of $G' \setminus S$ all have at most k vertices ν-a.s. and the ν-expected number of edges in S that are adjacent to o is at most $2\,\epsilon$.

Let $S(v)$ denote the edges of S incident to a vertex v. Let $G_1, G_2, \dots, G_m$ be the connected components of G. Let $\mathcal{A}_i$ be the event that G' is isomorphic to G_i. Define $g(G', S, x, y)$ to be the number of edges of S that are incident to x and

$f_i(G', S, x, y) = 1_{\mathcal{A}_i}\, g(G', S, x, y)$. Since ν is unimodular, the iMTP may be applied to f_i:

$$\int \sum_{y \in V(G')} f_i(G', S, o, y)\, d\nu = \int \sum_{x \in V(G')} f_i(G', S, x, o)\, d\nu\,. \tag{2.1}$$

The left hand side of (2.1) is equal to

$$\nu\big[|S(o)|\,|V(G')|\,1_{\mathcal{A}_i}\big] = |V(G_i)|\,\nu\big[|S(o)|\,1_{\mathcal{A}_i}\big],$$

while the right hand side is $\nu\big[2\,|S|\,1_{\mathcal{A}_i}\big]$. Hence,

$$|V(G_i)|\,\nu\big[|S(o)|\,1_{\mathcal{A}_i}\big] = \nu\big[2\,|S|\,1_{\mathcal{A}_i}\big].$$

Therefore, there is some $S_i \subset E(G_i)$ such that each connected component of $G_i \setminus S_i$ has at most k vertices and

$$|S_i|\,\nu[\mathcal{A}_i] \le \frac{|V(G_i)|}{2}\,\nu\big[|S(o)|\,1_{\mathcal{A}_i}\big]. \tag{2.2}$$

Define $\tilde{S} := \bigcup_{i=1}^m S_i$. Note that $\nu[\mathcal{A}_i]\,|V(G)|$ is the number of vertices of G that are in connected components of G that are isomorphic to G_i. Therefore $t_i := \nu[\mathcal{A}_i]\,|V(G)|/|V(G_i)|$ is the number of connected components of G that are isomorphic to G_i. Note that

$$\sum_{i=1}^m \frac{1_{\mathcal{A}_i}}{t_i} = 1\,.$$

Dividing (2.2) by $\nu[\mathcal{A}_i]$ and summing over i gives

$$|\tilde{S}| \le \frac{|V(G)|}{2}\,\nu\Big[|S(o)| \sum_i \frac{1_{\mathcal{A}_i}}{t_i}\Big] = \frac{|V(G)|}{2}\,\nu\big[|S(o)|\big] \le 2\,\epsilon\,|V(G)|\,,$$

which completes the proof. □

The idea of the proof of Theorem 1.2 is to replace the random set of edges, S, by a set of edges that still separates G into connected components of size at most k, but has the extra feature that, for any $e \in E(G)$, the event $e \in S$ depends only on the local structure of G near e and some randomness. Thus, it is easy to adapt the law of this new S to every random G' sufficiently close to G.

The proof of the theorem is complicated by a completely uninteresting point of a technical nature. For the sake of simplicity, we choose to present a proof that is not entirely precise, but does convey the essential ideas, and can be adapted to be completely correct. (See, [7].)

Proof of Theorem 1.2. Fix some $\epsilon_0 \in (0, 1/2)$. Let $\mathbf{P}$ be a unimodular measure on triples (G, S, o) such that the marginal law of (G, o) is μ, the $\mathbf{P}$-expected number of edges of S incident to o is at most $2\,\epsilon$, and every connected component of $G \setminus S$ has at most k vertices.

For any $K \subset V(G)$ and any $v \in K$, let $p(K) := \mathbf{P}\big[K(v) = K \;\big|\; (G, o)\big]$. (Here, we are not entirely precise, since the expression $\mathbf{P}\big[K(v) = K \;\big|\; (G, o)\big]$ does not have the standard meaning of the probability of an event conditional on a σ-field. This is because K and v do not, a priori, have a meaning without G. This difficulty is not too hard to overcome, at the expense of obfuscating the proof.) Clearly, $p(K)$ does not depend on the choice of the particular $v \in K$. Moreover, a simple argument using the iMTP shows that $p(K)$ is not effected by a change of the basepoint o; in other words, $p(K)$ is really a function of the isomorphism type of the pair (K, G).

Given (G,o) and $v \in V(G)$, let $\mathfrak{K}_v$ denote the collection of all connected sets of vertices, of cardinality at most k, that contain v. Obviously, $|\mathfrak{K}_v|$ is bounded by a function of k and M. For all $r \in \mathbb{N}_+$, let $\mathcal{B}_r$ denote the (isomorphism type of the) rooted graph $\big(B(o,r),o\big)$. For $r > k$ and $K \in \mathfrak{K}_o$, we set

$$p_r(K) := \mathbf{P}\big[K(o) = K \bigm| \mathcal{B}_r\big].$$

Clearly, $p_r(K)$ is a bounded martingale with respect to r and $\lim_{r\to\infty} p_r(K) = p(K)$ a.s. Consequently, there is some finite R such that

$$\mathbf{E} \sum_{K\in\mathfrak{K}_o} |p_R(K) - p(K)| < \epsilon_0\,.$$

Fix such an R.

For $K \in \mathfrak{K}_o$ let

$$N_R(K) := \bigcup_{v\in K} B(v,R)\,, \qquad \tilde{p}(K) = \tilde{p}(K,G) := \mathbf{P}\big[K = K(o) \bigm| N_R(K)\big].$$

By unimodularity, $\tilde{p}(K)$ does not depend on the choice of basepoint o in K. Therefore, by changing the basepoint, this also defines $\tilde{p}(K)$ for all $\mathfrak{K}_v$, $v \in V(G)$. Note that $\tilde{p}$ is a function whose arguments are a graph and a connected set of size at most k in it and that $\tilde{p}$ depends only on the R-neighborhood of the set. However, $\tilde{p}$ is only defined if its arguments can occur in (G,o). We extend the definition of $\tilde{p}$ to such arguments that have probability 0 of occuring in (G,o) (if such exist) by setting it equal to 1 on such arguments.

Let $\mathfrak{K} := \bigcup_{v\in V(G)} \mathfrak{K}_v$. Given (G,o), let $(X_K : K \in \mathfrak{K})$ be independent random variables such that

$$\mathbf{P}\big[X_K = 1 \bigm| (G,o)\big] = \min\{2\,\log(1/\epsilon_0)\,\tilde{p}(K), 1\}$$

and $X_K \in \{0,1\}$ a.s. Let $F := \bigcup_{K:X_K=1} \partial K$, where ∂K denotes the edge-boundary of K, and let $W := \bigcup\{K \in \mathfrak{K} : X_K = 1\}$. Let $\tilde{F}$ denote the set of all edges incident to some vertex in $V(G) \setminus W$ and set $S' := F \cup \tilde{F}$. It is easy to verify that the law of (G,S',o) is unimodular. Also, it is obvious that every connected component of $G \setminus S'$ has cardinality at most k.

Our present goal is to estimate $\mathbf{E}\big[|E(o) \cap S'|\big]$. We start by estimating $\mathbf{E}\big[|E(o) \cap \tilde{F}|\big]$. Let $\mathcal{A}$ denote the event $\sum_{K\in\mathfrak{K}_o} \tilde{p}(K) < 1/2$.

Lemma 2.2.

$$\mathbf{P}\big[\mathcal{A}\big] < 2\,\epsilon_0\,.$$

Proof. Fix some $K \in \mathfrak{K}_o$. Since the sequence $p_R(K), \tilde{p}(K), p(K)$ is a martingale given $B(o,k)$, we have

$$\mathbf{E}\big[|\tilde{p}(K) - p(K)| \bigm| B(o,k)\big] \le \mathbf{E}\big[|p_R(K) - p(K)| \bigm| B(o,k)\big]\,.$$

Our choice of R therefore gives

$$\mathbf{E}\Big[\sum_{K\in\mathfrak{K}_o} |\tilde{p}(K) - p(K)|\Big] < \epsilon_0\,.$$

Since $\sum_{K\in\mathfrak{K}_o} p(K) = 1$, we get

$$\mathbf{E}\Big[\Big|1 - \sum_{K\in\mathfrak{K}_o} \tilde{p}(K)\Big|\Big] < \epsilon_0$$

and Markov's inequality completes the proof of the lemma. □

Lemma 2.3.

$$\mathbf{P}\big[o \notin W\big] < 3\,\epsilon_0\,.$$

Proof. Given (G,o), the random variables X_K are independent and satisfy $\mathbf{P}\big[X_K = 1 \;\big|\; (G,o)\big] = \min\big\{1, 2\,\log(1/\epsilon_0)\,\tilde{p}(K)\big\}$. Hence,

$$\begin{aligned}\mathbf{P}\big[o \notin W \;\big|\; (G,o)\big] &= \prod_{K\in\mathfrak{K}_o}\Big(1 - \mathbf{P}\big[X_K = 1 \;\big|\; (G,o)\big]\Big)\\ &\le \exp\Big(-\sum_{K\in\mathfrak{K}_o} 2\,\log(1/\epsilon_0)\,\tilde{p}(K)\Big)\\ &\le \exp\big(-2\,\log(1/\epsilon_0)\,(1/2)\big) + 1_{\mathcal{A}} = \epsilon_0 + 1_{\mathcal{A}}\,.\end{aligned}$$

Now Lemma 2.2 completes the proof. □

The lemma implies

$$\mathbf{E}\big[|E(o)\cap\tilde{F}|\big] \le 3\,M\,\epsilon_0\,,$$

where M is the bound on the degrees in G.

Now,

$$\begin{aligned}\mathbf{E}\big[|E(o)\cap F| \;\big|\; (G,o)\big] &\le \sum_{K\in\mathfrak{K}_o} |\partial K\cap E(o)|\,\mathbf{P}\big[X_K = 1 \;\big|\; (G,o)\big]\\ &\le 2\,\log(1/\epsilon_0)\sum_{K\in\mathfrak{K}_o}|\partial K\cap E(o)|\,\tilde{p}(K)\,.\end{aligned}$$

Since,

$$\mathbf{E}\Big[\sum_{K\in\mathfrak{K}_o}|\partial K\cap E(o)|\,\tilde{p}(K)\Big] = \mathbf{E}\big[|E(o)\cap\partial K(o)|\big] = \mathbf{E}\big[E(o)\cap S\big] \le 2\,\epsilon\,,$$

we get

$$\mathbf{E}\big[|E(o)\cap F|\big] \le 4\,\log(1/\epsilon_0)\,\epsilon\,.$$

Thus, we have,

$$\mathbf{E}\big[|E(o)\cap S'|\big] \le 4\,\log(1/\epsilon_0)\,\epsilon + 3\,\epsilon_0\,M\,.$$

We now choose $\epsilon_0 := \epsilon/(2\,M)$.

Observe that the set S' is chosen in a very local way. Namely, given (G,o), you can decide if $e\in S'$ based purely on a fixed radius neighborhood of e and some coin flips that are associated with this neighborhood. Given another unimodular $(\tilde{G},\tilde{o})$, we can choose a $\tilde{S}\subset E(\tilde{G})$ according to the same procedure as S' is obtained for G. If $(\tilde{G},\tilde{o})$ is sufficiently close to (G,o), then the expected size of the set of edges in $\tilde{S}$ adjacent to $\tilde{o}$ will be close to the corresponding quantity in G. This proves the theorem. □

Proof of Theorem 1.1. It is easy to verify that the set of $\nu\in\mathcal{U}_M$ that are (k,ϵ)-hyperfinite is closed. Therefore, Lemma 2.1 implies that μ is hyperfinite if $\{G_1, G_2, \dots\}$ is hyperfinite. The converse follows immediately from Theorem 1.2 and Lemma 2.1. □

References

[1] O. Angel and I. Benjamini, *A Phase Transition for the Metric Distortion of Percolation on the Hypercube*, 2003, arXiv:math/0306355

[2] D. Aldous and R. Lyons, *Processes on Unimodular Random Networks*, 2006, arXiv:math/0603062

[3] R. M. Burton and M. Keane. *Density and uniqueness in percolation*, Comm. Math. Phys., **121** (1989), 501–505. MR 0990777 (90g:60090)

[4] I. Benjamini, R. Lyons, Y. Peres and O. Schramm. *Group-invariant percolation on graphs*, Geom. Funct. Anal., **9** (1999), 29–66. MR 1675890 (99m:60149)

[5] Itai Benjamini and Oded Schramm. *Percolation beyond* $\mathbf{Z}^d$, *many questions and a few answers*, Electron. Comm. Probab., **1** (1996), 71–82 (electronic). MR 1423907 (97j:60179)

[6] Itai Benjamini and Oded Schramm. *Recurrence of distributional limits of finite planar graphs*, Electron. J. Probab., **6** (2001) (electronic). MR 1873300 (2002m:82025)

[7] I. Benjamini, O. Schramm and A. Shapira. *Every minor-closed property of sparse graphs is testable*, 2007. Preprint.

[8] G. Elek, *The combinatorial cost*, 2006, arXiv:math/0608474

[9] G. Elek, L^2*-spectral invariants and convergent sequences of finite graphs*, 2007, arXiv:0709.1261

[10] G. Elek, *A regularity lemma for bounded degree graphs and its applications: Parameter testing and infinite volume limits*, 2007, arXiv:0711.2800.

[11] Olle Häggström, *Infinite clusters in dependent automorphism invariant percolation on trees*, Ann. Probab., **25** (1997), 1423–1436. MR 1457624 (98f:60207)

[12] László Lovász and Balázs Szegedy, *Limits of dense graph sequences*, J. Combin. Theory Ser. B, **96** (2006), 933–957. MR 2274085 (2007m:05132)

[13] Richard J. Lipton and Robert Endre Tarjan, *Applications of a planar separator theorem*, SIAM J. Comput., **9** (1980), 615–627. MR 0584516 (82e:68067)

Microsoft Corporation, One Microsoft Way, Redmond, WA 98074, USA
E-mail address: schramm@microsoft.com

Part IV
Percolation

Percolation beyond $\mathbb{Z}^d$: the contributions of Oded Schramm*

Olle Häggström†

August 27, 2010

Abstract

Oded Schramm (1961–2008) influenced greatly the development of percolation theory beyond the usual $\mathbb{Z}^d$ setting, in particular the case of nonamenable lattices. Here we review some of his work in this field.

Contents

1. Introduction
2. How it began
3. Invariant percolation and mass transport
4. No infinite cluster at criticality
5. Random walks on percolation clusters
6. Cluster indistinguishability
7. In the hyperbolic plane
8. Random spanning forests
9. Postscript

1 Introduction

Oded Schramm was born in 1961 and died in a hiking accident in 2008, in what otherwise seemed to be the middle of an extraordinary mathematical career. Although he made seminal contributions to many areas of mathematics in general and probability in particular, I will here restrict attention to his work on percolation processes taking place on graph structures more exotic than the usual $\mathbb{Z}^d$ setting. The title I've chosen alludes to the short but highly influential paper *Percolation beyond $\mathbb{Z}^d$, many questions and a few answers* from 1996 by Itai Benjamini and Oded Schramm [13].

I need to point out, however, that there are at least two respects in which I will fail to deliver on what my chosen title suggests. Firstly, I will not come anywhere near an exhaustive exposition of Oded's contributions to the field. All I can offer is a personal and highly subjective selection of highlights. Secondly, Oded was a very collaborative mathematician, and I will make no attempt (if it even makes sense) at identifying his individual contributions as opposed to his coauthors'. Suffice it to say that everyone who worked with him knew him as a very generous person and as

*Running head: Percolation beyond $\mathbb{Z}^d$. **Key words and phrases:** percolation, amenability, hyperbolic plane, mass transport, uniform spanning forest. **MSC Subject Classifications:** Primary 60K35. Secondary 82B43, 60B99, 51M10, 60K37.

†Work supported by the Göran Gustafsson Foundation for Research in the Natural Sciences and Medicine.Chalmers University of Technology.

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_20, © Springer Science+Business Media, LLC 2011

someone who would not put his name on a paper unless he had contributed at least his fair share. I'll just quote one recollection from Oded's long-time collaborator and friend Russ Lyons:

> To me, Oded's most distinctive mathematical talent was his extraordinary clarity of thought, which led to dazzling proofs and results. Technical difficulties did not obscure his vision. Indeed, they often melted away under his gaze. At one point when the four of us [Oded, Russ, Itai Benjamini and Yuval Peres] were working on uniform spanning forests, Oded came up with a brilliant new application of the Mass-Transport Principle. We weren't sure it was kosher, and I still recall Yuval asking me if I believed it, saying that it seemed to be "smoke and mirrors". However, when Oded explained it again, the smoke vanished. [59]

Following the spirit of the aformentioned paper [13], I will take "percolation beyond $\mathbb{Z}^d$" to mean percolation process that are not naturally thought of as embedded in d-dimensional Euclidean space. This excludes contributions by Oded not only to the theory of percolation on $\mathbb{Z}^d$ (such as [6]) but also to percolation on the triangular lattice (such as [71] and [73]) and to continuum percolation in $\mathbb{R}^d$ (such as [14]).

Other parts of Oded's work are discussed in the papers by Angel, Garban and Rohde in the present volume. For further reactions to Oded's untimely death, and memories of his life and work, see for instance Lyons [54], Häggström [33] and Werner [79], as well as the blog [59].

2 How it began

It will be assumed throughout that $G = (V, E)$ is an infinite but locally finite connected graph. In **i.i.d. site percolation** on G with retention parameter $p \in [0, 1]$, each vertex $v \in V$ is declared open (retained, value 1) with probability p and closed (deleted, value 0) with the remaining probability $1-p$, and this is done independently for different vertices. Alternatively, one may consider **i.i.d. bond percolation**, which is similar except that it is the edges rather than the vertices that are declared open or closed. Write $\mathbb{P}_{p,site}$ and $\mathbb{P}_{p,bond}$ for the resulting probability measures on $\{0,1\}^V$ and $\{0,1\}^E$, respectively. The choice whether to study bond or site percolation is often (but not always) of little importance and largely a matter of taste. In either case, focus is on the connectivity structure of the resulting random subgraph of G. Of particular interest is the possible occurrence of an infinite connected component – an infinite cluster, for short. The probability under $\mathbb{P}_{p,site}$ or $\mathbb{P}_{p,bond}$ of having an infinite cluster is always 0 or 1, and increasing in p. This motivates defining the site percolation critical value

$$p_{c,site}(G) \;=\; \inf\{p : \mathbb{P}_{p,site}(\exists \text{ an infinite cluster}) > 0\}$$

and the bond percolation critical value $p_{c,bond}(G)$ analogously.

By far the most studied case is where $G = (V, E)$ is the $\mathbb{Z}^d$ lattice with $d \geq 2$, meaning that $V = \mathbb{Z}^d$ and E consists of all pairs of Euclidean nearest neighbors. Some selected landmarks in the history of percolation are the 1960 result of Harris [37] that $p_{c,bond}(\mathbb{Z}^2) \geq \frac{1}{2}$; the 1980 result of Kesten [41] that $p_{c,bond}(\mathbb{Z}^2) = \frac{1}{2}$; the 1987 result of Aizenman, Kesten and Newman [1] establishing uniqueness of the infinite cluster for arbitrary d; and the strikingly short and beautiful alternative proof from 1989

by Burton and Keane [21] of the same result. See Grimmett [26] and Bollobás and Riordan [19] for introductions to percolation theory with emphasis on the $\mathbb{Z}^d$ case.

Benjamini and Schramm [13] were of course not the first to study percolation on more exotic graphs and lattices. The case where G is the $(d+1)$-regular tree $\mathbb{T}_d$ had been well-understood for a long time, essentially because it can be seen as a Galton–Watson process. Lyons [50, 51] had studied percolation on general trees, and Grimmett and Newman [29] had considered percolation on the Cartesian product $\mathbb{T}_d \times \mathbb{Z}$ (the Cartesian product $G = (V, E)$ of two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ has vertex set $V = V_1 \times V_2$, and an edge connecting (x_1, x_2) and (y_1, y_2) iff $x_1 = y_2$ are identical and x_2 and y_2 are neighbors or vice versa). However, it was only with the publication of Benjamini and Schramm [13] that a systematic study of percolation beyond $\mathbb{Z}^d$ began to take off towards anything like escape velocity. Clearly, they managed to find just the right time for launching the kind of informal research programme that their paper proposes. It should be noted, however, that a large part of the meaning of "just the right time" in this context is simply "soon after Oded Schramm had been drawn into probability theory".

When, as in [13], we focus on the occurrence and properties of infinite clusters for i.i.d. site (or bond) percolation on a graph $G = (V, E)$, a first basic issue is of course whether such clusters occur at all for any nontrivial value of p, i.e., whether $p_{c,site}(G) < 1$ (or $p_{c,bond}(G) < 1$). Benjamini and Schramm conjecture the following, where, for $v \in V$, $B(v, n)$ denotes the set of vertices $w \in V$ such that $\mathrm{dist}_G(v, w) \le n$, and dist_G is graph-theoretic distance in G.

Conjecture 2.1 (Conj. 2 in [13]) *If G is a quasi-transitive graph such that for some (hence any) $v \in V$, $|B(v, n)|$ grows faster than linearly, then $p_{c,site}(G) < 1$.*

Here, of course, we need to define quasi-transitivity of a graph (the term used in [13] was *almost transitive*, but the mathematical community quickly decided that quasi-transitive was preferable).

Definition 2.2 *Let $G = (V, E)$ be an infinite locally finite connected graph. A bijective map $f : V \to V$ such that $\langle f(u), f(v)\rangle \in E$ if and only if $\langle u, v\rangle \in E$ is called a* **graph automorphism** *for G. The graph G is said to be* **transitive** *if for any $u, v \in V$ there exists a graph automorphism f such that $f(u) = v$. More generally, G is said to be* **quasi-transitive** *if there is a $k < \infty$ and a partitioning of V into k sets $V_1, \dots, V_k$ such that for $i = 1, \dots, k$ and any $u, v \in V_i$ there exists a graph automorphism f such that $f(u) = v$.*

An important subclass of transitive graphs is the class of graphs arising as the Cayley graph of a finitely generated group. It may be noted that for quasi-transitive graphs (and more generally for bounded degree graphs; cf. [32]) we have $p_{c,bond} < 1$ iff $p_{c,site} < 1$, so Conjecture 2.1 may equivalently be phrased for bond percolation. Benjamini and Schramm found a short and elegant proof of the conjecture for the special case of so-called nonamenable graphs:

Definition 2.3 *The* **isoperimetric constant** *$h(G)$ of a graph $G = (V, E)$ is defined as*

$$h(G) = \inf_S \frac{|\partial S|}{|S|},$$

where the infimum ranges over all finite nonempty subsets of V, and $\partial S = \{u \in V \setminus S : \exists v \in S \text{ such that } \langle u, v\rangle \in E\}$. The graph G is said to be **amenable** *if $h(G) = 0$; otherwise it is said to be* **nonamenable**.

(Sometimes, as in [13], $h(G)$ is also called the Cheeger constant.)

Theorem 2.4 (Thm. 2 in [13]) *Any nonamenable graph G satisfies $p_{c,site}(G) < 1$. In fact,*

$$p_{c,site}(G) \leq \frac{1}{h(G)+1}\,. \tag{1}$$

Proof. Fix p and a vertex $\rho \in V$, and consider the following sequential procedure for searching the open cluster containing ρ. If ρ is open, set $S_1 = \{\rho\}$, otherwise stop. At each integer time n, check the status (open or closed) of some thitherto unchecked vertex v in ∂S_{n-1}; if v is open we set $S_n = S_{n-1} \cup \{v\}$, if v is closed we set $S_n = S_{n-1}$, while if no such v can be found the procedure terminates. Define $X_0 = 0$ and $X_n = |S_n|$ for $n \geq 1$, and note that $\{X_0, X_1, \ldots\}$ is a random walk whose i.i.d. increments take value 0 with probability $1-p$ and 1 with probability p, stopped at some random time. If G has isoperimetric constant $h(G)$, then the random walk can stop only when $\frac{n-X_n}{X_n} \geq h(G)$, i.e., when

$$\frac{X_n}{n} \leq \frac{1}{h(G)+1}\,. \tag{2}$$

But the random walk has drift p, so when $p > \frac{1}{h(G)+1}$ the Strong Law of Large Numbers implies that with positive probability $\frac{X_n}{n}$ never satisfies (2), in which case the walk never stops and ρ belongs to an infinite cluster. □

It may be noted that the result is sharp in the sense that the bound (1) holds with equality for the tree $\mathbb{T}_d$, for which $p_{c,site}(\mathbb{T}_d) = \frac{1}{d}$ and $h(\mathbb{T}_d) = d-1$.

Once $p_{c,site}(G) < 1$, we know that i.i.d. site percolation on G has two distinct phases: for $p < p_{c,site}$ there is no infinite cluster, while for $p > p_{c,site}$ there is. But what happens *at* the critical value? This has long been a central issue in percolation theory. When G is the $\mathbb{Z}^d$ lattice with $d \geq 2$, the consensus belief among percolation theorists is that there is no infinite cluster at criticality; this is known for $d = 2$ (Russo [67]) and $d \geq 19$ (Hara and Slade [36]), but the general case remains open. Benjamini and Schramm suggest that "it might be beneficial to study the problem in other settings":

Conjecture 2.5 (Conj. 4 in [13]) *For any quasi-transitive graph G with $p_{c,site}(G) < 1$, there is a.s. no infinite cluster at criticality.*

This remains open (as of course it must be as long as the $\mathbb{Z}^d$ case with $3 \leq d \leq 18$ stays unsolved), but Benjamini and Schramm were soon to be involved in remarkable progress towards proving it; see Section 4. The quasi-transitivity condition cannot be dropped, as it is easy to construct graphs with a nontrivial critical value which nevertheless percolate at criticality (a tree growing slightly faster than a binary tree will do; see, e.g., [51]).

Another natural next question, once the existence of infinite clusters at nontrivial values of p (i.e. $p_{c,site} < 1$) is established, concerns how many infinite clusters there can be. Benjamini and Schramm [13] noted that the argument of Newman and Schulman [61] for showing that the number of infinite clusters is, for fixed p, an a.s. constant which must equal 0, 1 or ∞ extends to the setting of quasi-transitive graphs. They furthermore saw that the argument of Burton and Keane [21] for ruling out infinitely many infinite clusters extends to amenable quasi-transitive graphs (a similar observation was made earlier in Thm. 1' of Gandolfi, Keane and Newman [25]). In other words:

Theorem 2.6 *For any amenable quasi-transitive graph G and any $p \in [0,1]$, the number of infinite clusters produced by i.i.d. site or bond percolation on G with parameter p is either 0 or 1 a.s.*

In contrast, i.i.d. site percolation the regular tree $\mathbb{T}_d$ with $d \geq 2$ exhibits infinitely many infinite clusters for all $p \in (p_{c,site}, 1)$. Also, the $\mathbb{T}_d \times \mathbb{Z}$ example studied by Grimmett and Newman [29] exhibits the same phenomenon when p is above but sufficiently close to $p_{c,site}$ (Grimmett and Newman showed this for large d, and later Schonmann [69] indicated how to do it for all $d \geq 2$). Benjamini and Schramm [13] conjectured that the Burton–Keane argument is sharp in the sense that whenever G is quasi-transitive and nonamenable, uniqueness of the infinite cluster fails for p above but sufficiently close to $p_{c,site}$. In terms of the so-called uniqueness critical value

$$p_{u,site} = p_{u,site}(G) = \inf\{p \in [0,1] : \text{i.i.d. site percolation on } G \text{ with parameter } p \text{ produces a.s. a unique infinite cluster}\},$$

their conjecture reads:

Conjecture 2.7 (Conj. 6 in [13]) *For any nonamenable quasi-transitive graph G, we have $p_{c,site}(G) < p_{u,site}(G)$.*

In the spirit of the Grimmett–Newman result mentioned above, they proved that for any quasi-transitive graph G, the product graph $\mathbb{T}_d \times G$ satisfies $p_{c,site}(G) < p_{u,site}(G)$ provided d is large enough (this is Cor. 1 in [13]).

Conjecture 2.7 has stimulated further research as well. For instance, Pak and Smirnova-Nagnibeda [63] proved in the case of bond percolation that it holds for the Cayley graph of any nonamenable group provided an appropriate choice of generators. Lalley [43] and a later paper by Benjamini and Schramm [16] proved it for certain classes of nonamenable planar graphs; this will be discussed in more detail in Section 7.

By definition of $p_{u,site}$ and the Newman–Schulman 0-1-∞ law, we have infinitely many infinite clusters a.s. for any $p \in (p_{c,site}, p_{u,site})$. It would be nice to add that uniqueness holds for all $p \in (p_{u,site}, 1)$, but for this we need the monotonicity property that if $p_1 < p_2$ and uniqueness holds a.s. at parameter value p_1, then it holds at p_2 as well; this is part of Question 5 in [13]. The required monotonicity was proved by Häggström and Peres [35] for quasi-transitive graphs under the additional assumption of unimodularity (see Definition 3.5 below), and by Schonmann [68] without this additional assumption.

The other part of Question 5 in [13] concerns, in the case where G is quasi-transitive with $p_{c,site} < p_{u,site} < 1$, the number of infinite cluster at $p = p_{u,site}$: one or infinitely many? Somewhat surprisingly, the answer turned out to depend on the choice of G. For the Grimmett–Newman example, Schonmann [69] showed that there are infinitely many infinite clusters at the uniqueness critical point $p_{u,site}$, a result that Peres [65] extended to more general product graphs. In contrast, Benjamini and Schramm [16] showed that for planar nonamenable graphs with one end, there is a unique infinite cluster at $p_{u,site}$; see Section 7 again.

There is a good deal more to say about the *Percolation beyond $\mathbb{Z}^d$* paper [13], but I must move on to some of Oded Schramm's later contributions. The paper's influence will be evident from the coming sections, but see also Benjamini and Schramm [15], Lyons [53] and Häggström and Jonasson [34] for partially overlapping surveys of what happened in the wake of the paper.

3 Invariant percolation and mass transport

Soon after finishing the *Percolation beyond* $\mathbb{Z}^d$ paper [13], Itai Benjamini and Oded Schramm joined forces with Russ Lyons and Yuval Peres (this quartet of authors will appear frequently in what follows, and will be abbreviated BLPS). In [9] they broadened the scope compared to [13] by considering percolation processes on quasi-transitive graphs in a more general situation than the i.i.d., namely **automorphism invariance**.

Definition 3.1 *Let $G = (V, E)$ be a quasi-transitive graph and let* $\mathrm{Aut}(G)$ *denote the group of graph automorphisms of G. A $\{0,1\}^V$-valued random object X is called a site percolation for G, and it is said to be* **automorphism invariant** *if for any n, any $v_1, \ldots, v_n \in V$, any $b_1, \ldots, b_n \in \{0,1\}$ and any $\gamma \in \mathrm{Aut}(G)$, we have*

$$\mathbb{P}(X(\gamma v_1) = b_1, \ldots, X(\gamma v_n) = b_n) = \mathbb{P}(X(v_1) = b_1, \ldots, X(v_n) = b_n)\,.$$

Automorphism invariance of a bond percolation for G is defined analogously. In fact, much of the work in [9] concerns an even more general setting, namely invariance under certain kinds of subgroups of $\mathrm{Aut}(G)$. Here, for simplicity, I will restrict attention to the case of invariance under the full automorphism group $\mathrm{Aut}(G)$.

There are plenty of automorphism invariant percolation processes beyond i.i.d. that arise naturally. Examples in the site precolation case include certain Gibbs distributions for spin systems such as the Ising model, and certain equilibrium measures for interacting particle systems such as the voter model. In the bond percolation case, they include the random-cluster model [27] as well as the random spanning forest models to be discussed in Section 8.

Amongst the most important contributions of BLPS [9] is the introduction of the so-called **Mass-Transport Principle** in percolation theory, and the beginning of a systematic exploitation of it for understanding the behavior of percolation processes. (This was partly inspired by an application in Häggström [31] of similar ideas in the special case where G is a regular tree.) As a kind of warm-up for readers unfamiliar with the mass-transport technique, let me suggest a very simple toy problem.

Problem 3.2 *Given a transitive graph $G = (V, E)$, does there exist an automorphism invariant bond percolation process which produces, with positive probability, some infinite open cluster consisting of a single self-avoiding path which is infinite in just one direction? (We call such a self-avoiding path* **uni-infinite**.*) In other words, this open cluster should consist of a single vertex of degree* 1 *in the cluster, while all the other (infinitely many) vertices of the cluster should have degree* 2*.*

Call an infinite cluster **slim** if it is of the desired kind. (In a sense, a slim infinite cluster is the smallest infinite cluster there can be.) Also, given an automorphism invariant bond percolation process X taking values in $\{0,1\}^E$, we define random variables $\{Y(v)\}_{v \in V}$ as follows. If v does not belong to a slim infinite cluster in v we set $Y(v) = 0$; otherwise we let $Y(v)$ be one plus the distance in the slim cluster from v to the one endpoint of this cluster. For $k = 0, 1, \ldots$, write $\alpha(v, k) = \mathbb{P}(Y(v) = k)$. Automorphism invariance ensures that this is independent of the choice of v, so we may write $\alpha(k)$ for $\alpha(v, k)$.

When $G = (V, E)$ is the $\mathbb{Z}^d$ lattice, we can argue as follows. Write Λ_n for the box $\{-n, -n+1, \ldots, n\}^d \subset V$. The expected number of vertices $v \in \Lambda_n$ with $Y(v) = 1$ is $(2n+1)^d \alpha(1)$. But for any k, and any vertex v with $Y(v) = 1$, there must be

a corresponding vertex u with $Y(u) = k$ within distance $k-1$ from v. Hence the expected number of vertices $v \in \Lambda_{n+k-1}$ with $Y(v) = k$ is at least $(2n+1)^d \alpha(1)$, so

$$(2(n+k)-1)^d \alpha(k) \;\geq\; (2n+1)^d \alpha(1)\,,$$

and sending $n \to \infty$ yields $\alpha(k) \geq \alpha(1)$. Similarly, $\alpha(1) \geq \alpha(k)$, so that in fact $\alpha(1) = \alpha(k)$, and since k was arbitrary we have $\alpha(1) = \alpha(2) = \alpha(3) = \cdots$. But $\sum_{k=1}^{\infty} \alpha(k) \leq 1$, so $\alpha(k)$ must be 0 for each k, whence slim infinite clusters do not occur in the $G = \mathbb{Z}^d$ case.

The crucial property of the $\mathbb{Z}^d$ lattice that makes the argument work is that $\frac{|\Lambda_{n+k}|}{|\Lambda_n|} \to 1$ as $n \to \infty$. Hence, the argument is easily extended to the more general case where G is transitive and amenable.

But what about the case where G is nonamenable? Now the argument doesn't generalize, and in fact the following example gives us problems. And **end** in a graph $G = (V, E)$ is an equivalence class of uni-infinite self-avoiding paths in X, with two paths equivalent if for all finite $W \subset V$ the paths are eventually in the same connected component of the graph obtained from G by deleting all $v \in W$.

Example 3.3 (Trofimov's graph [77]) *Consider the regular binary tree $\mathbb{T}_2$, and fix an end ξ in this tree. For each vertex v in the tree, there is a unique uni-infinite self-avoiding path from v that belongs to ξ. Call the first vertex after v on this path the* **ξ-parent** *of v, and call the other two neighbors of v its* **ξ-children**. *The* **ξ-grandparent** *of v is defined similarly in the obvious way. Let $G = (V, E)$ be the graph that arises by taking $\mathbb{T}_2$ and adding, for each vertex v, an extra edge connecting v to its ξ-grandparent.*

Clearly, Trofimov's graph G is transitive, and it also inherits the nonamenability property of $\mathbb{T}_2$. It turns out that on G, it is possible to construct an automorphism invariant bond percolation exhibiting slim infinite clusters:

Example 3.4 *Let $G = (V, E)$ be Trofimov's graph, and consider the following automorphism invariant bond percolation on G: Each $v \in V$ will have an open edge to exactly one of its ξ-children, and for each v independently, toss a fair coin to decide which ξ-child to connect to. All grandparent-grandchild edges are closed. (To see that this bond percolation is indeed automorphism invariant, it is necessary, but easy, to check that the end ξ can be identified by just looking at the graph structure of G.) This produces a percolation configuration in which a.s. each v sits in a slim infinite cluster. From v the open path extends downwards (i.e., away from ξ) infinitely, and upwards a geometric($\frac{1}{2}$) number of steps.*

So perhaps slim infinite clusters can arise as soon as G is nonamenable? In fact, no. It turns out that the mass-transport method of BLPS [9] applies to rule out slim infinite clusters also in the nonamenable case, as long as the graphs satisfy the additional assumption of **unimodularity**. Unimodularity holds for all specific examples considered so far except for Trofimov's graph. It holds for Cayley graphs in general, and I daresay it tends to hold for most transitive graphs that are not constructed for the explicit purpose of being nonunimodular.

Definition 3.5 *Let $G = (V, E)$ be a quasi-transitive graph with automorphism group $\mathrm{Aut}(G)$. For $v \in V$, the* **stabilizer** *of v is defined as $\mathrm{Stab}(v) = \{\gamma \in \mathrm{Aut}(G) : \gamma v = v\}$. The graph G is said to be* **unimodular** *if for all $u, v \in V$ in the same orbit of $\mathrm{Aut}(G)$ we have the symmetry*

$$|\mathrm{Stab}(u)v| = |\mathrm{Stab}(v)u|\,.$$

(Note how Trofimov's graph fails to be unimodular: For two vertices u and v such that u is the ξ-parent of v, we get $|\mathrm{Stab}(u)v| = 2$ but $|\mathrm{Stab}(v)u| = 1$. Each vertex has two children but just one parent.)

Proposition 3.6 *In automorphism invariant bond percolation on a quasi-transitive unimodular graph G, there is a.s. no slim infinite cluster.*

This, we will find, is an easy consequence of the Mass-Transport Principle of BLPS [9]. For an automorphism invariant site (or bond) percolation on a quasi-transitive graph $G = (V, E)$, let μ be the corresponding probability measure on $\{0,1\}^V$ (or on $\{0,1\}^E$). Consider a nonnegative function $m(u, v, \omega)$ of three variables: two vertices $u, v \in V$ and the percolation configuration ω taking values in $\Omega = \{0,1\}^V$ (or $\Omega = \{0,1\}^E$). Intuitively, we should think of $m(u, v, \omega)$ as the mass transported from u to v given the configuration ω. We assume that $m(u, v, \omega) = 0$ unless u and v are in the same orbit of $\mathrm{Aut}(G)$, and furthermore that $m(\cdot, \cdot, \cdot)$ is invariant under the diagonal action of $\mathrm{Aut}(G)$, meaning that $m(u, v, \omega) = m(\gamma u, \gamma v, \gamma\omega)$ for all u, v, ω and $\gamma \in \mathrm{Aut}(G)$.

Theorem 3.7 (The Mass-Transport Principle, Sect. 3 in [9]) *Given G, μ and $m(\cdot, \cdot, \cdot)$ as above, let*

$$M(u, v) \;=\; \int_\Omega m(u, v, \omega) d\mu(\omega)$$

for any $u, v \in V$. If G is unimodular, then the expected total mass transported out of any vertex v equals the expected mass transported into v, i.e.,

$$\sum_{u \in V} M(v, u) \;=\; \sum_{u \in V} M(u, v)\,. \tag{3}$$

The Mass-Transport Principle as stated here fails if G is not unimodular. (To see this for Trofimov's graph, we can consider the the mass transport in which each vertex simply sends unit mass to its ξ-parent, regardless of the percolation configuration. Then each vertex sends mass 1 but receives mass 2, thus violating (3).) In fact, BLPS [9] did state a version of the Mass-Transport Principle that holds also in the nonunimodular case; this involves a reweighting of the mass sent from u to v by a factor that depends on $\frac{|\mathrm{Stab}(u)v|}{|\mathrm{Stab}(v)u|}$. But it is in the unimodular case that the Mass-Transport Principle has turned out most useful, and for simplicity we stick to this case.

The proof of the Mass-Transport Principle is particularly simple in the case where G is the Cayley graph of a finitely generated group H, so here I will settle for that case only:

Proof of Theorem 3.7 in the Cayley graph case: For $u, v \in V$, we also have that u and v are group elements of H, and that there is a unique element $h = uv^{-1} \in H$ such that $u = hv$. This gives

$$\begin{aligned} \sum_{u \in V} M(v, u) \;&=\; \sum_{h \in H} M(v, hv) \;=\; \sum_{h \in H} M(h^{-1}v, v) \\ &=\; \sum_{h' \in H} M(h'v, v) \;=\; \sum_{u \in V} M(u, v)\,, \end{aligned}$$

where the second equality follows from automorphism invariance. □

Proof of Proposition 3.6: Consider the mass transport in which each vertex v sitting in a slim infinite cluster sends unit mass to the unique endpoint of this cluster.

Vertices not sitting in a slim infinite cluster send no mass at all. Then the expected mass sent from a vertex is at most 1, while if slim infinite clusters exist with positive probability then some vertices will receive infinite mass with positive probability, so that the expected mass received is infinite, contradicting (3). □

Proposition 3.6 is just an illustrative example, but BLPS [9] proved several other more interesting reuslts using the Mass-Transport Principle, which turns out to be quite a potent tool in nonamenable settings where classical density arguments and ergodic averages are not available in the same way as in the amenable case. The Mass-Transport Principle in itself is not especially deep or difficult. Rather, in the words of BLPS [10], "the creative element in applying the mass-transport method is to make a judicious choice of the transport function $m(u, v, \omega)$". We will see some examples in this section and the next. (For a remarkable recent development of the mass-transport method, see Aldous and Lyons [3] and Schramm [72].)

The following result characterizes amenability of Cayley graphs (and more generally of unimodular transitive graphs) in terms of a certain percolation threshold for invariant percolation. For a transitive graph $G = (V, E)$ and a automorphism invariant site percolation on G with distribution μ on $\{0, 1\}^V$, write $\pi(\mu)$ for the marginal probability that a given vertex is open.

Theorem 3.8 [9] *Let G be a unimodular transitive graph, and define $p_{c,inv}(G)$ is the infimum over all $p \in [0, 1]$ such that any automorphism invariant site percolation μ on G with $\pi(\mu) = p$ is guaranteed to produce at least one infinite cluster with positive probability. Then $p_{c,inv}(G) < 1$ if and only if G is nonamenable.*

Quantitative estimates for $p_{c,inv}(G)$ in the nonamenable case are also provided in [9]. Theorem 3.9 below gives such a bound in the bond percolation case. For site percolation, BLPS [9] show that if $\pi(\mu) \geq \frac{d(G)}{d(G)+h(G)}$, where $d(G)$ is the degree of a vertex in G, and $h(G)$ as before is the isoperimetric constant, then there is at least one infinite cluster with positive probability. Similar bonds are given for the quasi-transitive case as well. For the case $G = \mathbb{T}_d$ the bounds go back to Häggström [31], where they were established using a precursor of the mass-transport method, and also shown to be sharp. The following bound in the bond percolation case is in terms of the **edge-isoperimetric constant** $h_E(G)$, defined by

$$h_E(G) \;=\; \inf_S \frac{|\partial_E S|}{|S|}$$

where as in Definition 2.3 the infimum ranges over all finite $S \in V$, while $\partial_E S = \{\langle u, v\rangle \in E : u \in S, v \in V \setminus S\}$. Clearly $h(G) \leq h_E(G) \leq (d(G) - 1)h(G)$ when G is transitive, so such a G is amenable in the sense of Definition 2.3 if and only if $h_E(G) = 0$.

Theorem 3.9 [9] *Let $G = (V, E)$ be transitive and unimodular, and consider an automorphism invariant bond percolation on G such that for each edge $e \in E$ we have*

$$\mathbb{P}(e \text{ is open}) \;\geq\; \frac{d(G) - h_E(G)}{d(G)}\,. \tag{4}$$

Then the percolation produces an infinite cluster with positive probability.

The proof is worth exhibiting here, but in order to be able to follow the elegant argument from the expository follow-up paper BLPS [10], I will be content with considering the case where (4) holds with strict inequality.

Proof of (almost) Theorem 3.9: For a finite subgraph $G' = (V', E')$ of G define its average internal degree

$$h_E^*(G') \;=\; \frac{2|E'|}{|V'|}$$

and set

$$h_E^*(G) \;=\; \sup_{G'} h_E^*(G') \tag{5}$$

where the supremum is over all finite subgraphs G' of G. For any given such G', we have

$$2|E'| + |\{\langle u, v\rangle \in E : u \in V', v \in V \setminus V'\}| \;\leq\; d(G)|V'|$$

with equality if and only if all $e \in E$ with both endpoints in V' are also in E'. Hence $h_E^*(G) + h_E(G) = d(G)$, so the right-hand side of (4) equals $\frac{h_E^*(G)}{d(G)}$. Now consider a $\{0,1\}^E$-valued automorphism invariant bond percolation X on G such that

$$\mathbb{P}(e \text{ is open}) \;>\; \frac{h_E^*(G)}{d(G)} \tag{6}$$

for each $e \in E$, and assume for contradiction that it a.s. produces no infinite cluster. We may define a mass transport where each vertex sitting in a finite open cluster counts the number of open edges incident to it, sends out exactly this amount of mass, and distributes it equally among all the vertices sitting in its connected component in the percolation process. In other words, we take the transport function to be

$$m(u, v, \omega) \;=\; \begin{cases} \frac{d_\omega(u)}{|K(u)|} & \text{if } u \text{ is in a finite component of } \omega \text{ and } v \in K(u) \\ 0 & \text{otherwise.} \end{cases}$$

where $d_\omega(u)$ is the degree of u in X, and $K(u)$ is the set of vertices having an open path in ω to u. Then (6) and the assumption that X produces no infinite clusters give

$$\begin{aligned} \mathbb{E}\left[\sum_{v \in V} m(u, v, X)\right] \;&\geq\; d(G) \min_{e \in E} \mathbb{P}(e \text{ is open}) \\ &>\; d(G)\frac{h_E^*(G)}{d(G)} \;=\; h_E^*(G) \end{aligned}$$

while the amount $\sum_{v \in V} m(v, u, X)$ received at u is the average internal degree of its connected component, which is bounded by $h_E^*(G)$. Hence

$$\mathbb{E}\left[\sum_{v \in V} m(u, v, X)\right] \;>\; \mathbb{E}\left[\sum_{v \in V} m(v, u, X)\right]$$

contradicting the Mass-Transport Principle. □

The next result from BLPS [9] concerns the expected degree of a vertex given that it belongs to an infinite cluster. Here it seems most natural to consider the bond percolation case. For G transitive and an invariant bond percolation on G with distribution μ that produces at least one infinite cluster with positive probability, define $\beta(G, \mu)$ as the expected degree of a vertex given that it belongs to an infinite cluster, and define $\beta(G) = \inf_\mu \beta(G, \mu)$ where the infimum ranges over all such automorphism invariant bond percolation processes on G.

Theorem 3.10 [9] *For $G = (V, E)$ transitive, we have $\beta(G) = 2$ if G is unimodular, and $\beta(G) < 2$ otherwise.*

An exact expression for $\beta(G)$ in the unimodular case is also given, namely $1 + \inf_{\langle u,v\rangle \in E} \frac{|\mathrm{Stab}(u)v|}{|\mathrm{Stab}(v)u|}$, where transitivity implies that the infimum is in fact a minimum. Note how Theorem 3.10 immediately implies Proposition 3.6 above: a slim infinite cluster would have vertices both of degree 1 and of degree 2, and these are the only degrees appearing, so the average degree would have had to be strictly between 1 and 2. That average degree 2 is needed is quite intuitive, but what is more surprising is that it is possible to go below 2 in the nonunimodular case. In fact, the bond percolation on Trofimov's graph in Example 3.4 has $\beta(G, \mu) = \frac{3}{2}$, and this can be pushed down to $\beta(G, \mu) = \frac{5}{4}$ (which is sharp for Trofimov's graph) by letting the slim infinite clusters live on grandparent-grandchild rather than parent-child edges.

Another striking result in BLPS [9] concerns the number of ends of infinite components: if G is quasi-transitive and unimodular and μ is an automorphism invariant site or bond percolation, then the number of ends of any infinite component must be either 1, 2 or ∞. (In the amenable case the argument of Burton and Keane [21] excludes also the case of infinitely many ends.) Furthermore, for infinite clusters with infinitely many ends, BLPS [9] showed that that such clusters have expected degree strictly greater than 2, and that they have critical values for site or bond percolation that are strictly greater than 1.

4 No infinite cluster at criticality

I have yet to mention what is possibly the most striking result of all from BLPS [9]. Namely, this study of automorphism invariant percolation turned out to have the following implication for i.i.d. percolation, which is a remarkable step in the direction of Conjecture 2.5.

Theorem 4.1 *Let G be a nonamenable unimodular quasi-transitive graph, and consider i.i.d. site percolation on G at the critical value $p = p_{c,site}$. Then there is a.s. no infinite cluster. The analogous statement for i.i.d. bond percolation holds as well.*

Due to the focus in [9] being on the more general setting of automorphism invariant percolation, the proof given there is not the most direct possible. The authors therefore chose to publish a separate expository note, BLPS [10], with a more direct proof which is well worth recalling here. Following [10], I will restrict to the case of bond percolation on a (nonamenable, unimodular) transitive graph; the cases of site percolation and quasi-transitive graphs require only minor modification.

The reason why unimodularity is needed in the proof is that the Mass-Transport Principle is used – in fact, it is used several times (including in the proof of Theorem 3.9 which the proof of Theorem 4.1 falls back on). But we should probably expect the result to be true also in the nonunimodular case (certainly if we trust Conjecture 2.5). See Timár [76] and Peres, Pete and Scolnicov [66] for what are perhaps the best efforts to date towards a better understanding of the nonunimodular case.

Proof of Theorem 4.1 for bond percolation on transitive graphs. Let $G = (V, E)$ be a nonamenable unimodular transitive graph, and consider an i.i.d. bond percolation X on G with $p = p_{c,bond}(G)$. By the Newman-Schulman 0-1-∞ law, the number of infinite clusters in X is an a.s. constant N which equals either 0, 1 or ∞, and we need to rule out the possibilities $N = 1$ and $N = \infty$.

CASE I. RULING OUT $N = 1$. Assume for contradiction that the percolation configuration $X \in \{0,1\}^E$ has a unique infinite cluster a.s. Let $\{Y(e)\}_{e\in E}$ be an i.i.d. collection of random variables, uniformly distributed on the unit interval $[0,1]$ and independent also of X. For each $\varepsilon \in (0,1)$ and $e \in E$, define

$$X_\varepsilon(e) = \begin{cases} 1 & \text{if } X(e) = 1 \text{ and } Y(e) > \varepsilon \\ 0 & \text{otherwise,} \end{cases}$$

and note that

$$X_\varepsilon \in \{0,1\}^E \text{ is an i.i.d. bond percolation on } G \text{ with parameter } (1-\varepsilon)p_{c,bond}\,. \quad (7)$$

For $\varepsilon \in (0,1)$, define yet another bond percolation $Z_\varepsilon \in \{0,1\}^E$, not i.i.d. but automorphism invariant, as follows. As before let dist_G denote graph-theoretic distance in G, and for each $v \in V$ define $U(v)$ as the set of vertices in the infinite cluster of X that minimize $\mathrm{dist}_G(u,v)$. Note that $U(v)$ is finite for all $v \in V$. For an edge $e = \langle v,w\rangle \in E$, set

$$Z_\varepsilon(e) = \begin{cases} 1 & \text{if all vertices in } U(v) \text{ and } U(w) \text{ are} \\ & \text{in the same connected component of } X_\varepsilon \\ 0 & \text{otherwise.} \end{cases}$$

This defines the percolation $Z_\varepsilon \in \{0,1\}^E$. For any $\langle v,w\rangle \in E$, there exists some finite collection $T(v,w) \subset E$ of open edges in X that together connect all the vertices in $U(v)$ and $U(w)$ to each other (this is where the assumption $N = 1$ is used). For definiteness, we take $T(v,w)$ to be the edge set with minimal cardinality having this property, and with minimization of $\sum_{e\in T(v,w)} Y(e)$ acting as tie-breaker. Each edge e' in this collection has $Y(e') > 0$ a.s., and so $\min_{e'\in T(v,w)} Y(e') > 0$ so that $\lim_{\varepsilon\to 0} Z_\varepsilon(e) = 1$ a.s. Hence

$$\lim_{\varepsilon\to 0} \mathbb{P}[Z_\varepsilon(e) = 1] \;=\; 1\,, \quad (8)$$

and Theorem 3.9 ensures that that the percolation process Z_ε contains an infinite cluster with positive probability. But when Z_ε contains an infinite cluster, then so does X_ε. In view of (7), this contradicts the definition of $p_{c,bond}$, so we are done with CASE I.

CASE II. RULING OUT $N = \infty$. This time assume for contradiction that X contains infinitely many infinite clusters a.s. Here we need the concept of **encounter points** introduced by Burton and Keane [21] in their famous short proof of uniqueness of the infinite cluster on $\mathbb{Z}^d$. An encounter point in a percolation process X is a vertex $v \in V$ that has three disjoint open paths to infinity that would fall in different connected components of X if the vertex v were to be removed, but does not have four such paths. Burton and Keane showed that if $N = \infty$, then X contains encounter points a.s.; their proof was formulated for $G = \mathbb{Z}^d$, but goes through unchanged for general (quasi-)transitive graphs.

BLPS [10] begin by noting that

$$\begin{array}{c} \text{if } v \in V \text{ is an encounter point, then a.s. each of the three} \\ \text{infinite clusters } C_1(V), C_2(v), C_3(v) \text{ that the removal} \\ \text{of } v \text{ would produce contains further encounter points.} \end{array} \quad (9)$$

To see this, consider the mass transport in which each vertex u sitting in an infinite cluster with encounter points sends unit mass to the nearest encounter point with respect to dist_X, splitting it equally in case of a tie; here dist_X means graph-theoretic distance in the open subgraph of G defined by X. Failure of (9) would cause a contradiction to the Mass-Transport Principle similarly as in the proof of Proposition 3.6.

Next, we go on to define a random graph $H = (W, F)$, whose vertex set $W \subset V$ is the set of encounter points in X, and whose edge set F, which we are about to specify, will not necessarily be a subset of E (so H is not a subgraph of G). Let $\{Y(v)\}_{v \in W}$ be i.i.d., uniformly distributed on $[0, 1]$ and independent of X. Each $v \in W$ selects three other $u_1, u_2, u_3 \in V$ to form edges to, according to the following rule: one u_i should be chosen in each of the components $C_1(v)$, $C_2(v)$ and $C_3(v)$, and in each such component u_i is chosen to minimize $\mathrm{dist}_X(v, u_i)$, with minimization of $Y(u_i)$ acting as a tie-breaker. An equivalent way to formulate this is that u_i is chosen in C_i to minimize the "distance" $\mathrm{dist}_{X,Y}(v, u_i)$ defined as $\mathrm{dist}_{X,Y}(v, u_i) = \mathrm{dist}_X(v, u_i) + Y(v) + Y(u_i)$.

Each $v \in V$ thus gets H-degree at least 3, but the H-degree may exceed 3 if v is selected by some $w \in W$ which is not among v's preferred triplet. An application of the Mass-Transport Principle shows that the expected number of vertices that choose v is exactly 3, so the expected H-degree of v (conditional on being in W) is somewhere between 3 and 6, and in particular its degree is a.s. finite.

A crucial step of the argument is now to show that

$$\text{the graph } H \text{ has no cycles.} \tag{10}$$

To see this, we first note that if $v \in W$ is in such a cycle, then its two neighbors in this cycle must belong to the same $C_i(v)$, as otherwise we would get a direct contradiction to the definition of an encounter point. Using this, it is not hard to see (or consult [10] for a more detailed argument) that any cycle $v_1 \leftrightarrow v_2 \leftrightarrow \cdots \leftrightarrow v_k \leftrightarrow v_1$ would have to satisfy either

$$\mathrm{dist}_{X,Y}(v_1, v_2) < \mathrm{dist}_{X,Y}(v_2, v_3) < \cdots < \mathrm{dist}_{X,Y}(v_k, v_1) < \mathrm{dist}_{X,Y}(v_1, v_2)$$

or

$$\mathrm{dist}_{X,Y}(v_1, v_2) > \mathrm{dist}_{X,Y}(v_2, v_3) > \cdots > \mathrm{dist}_{X,Y}(v_k, v_1) > \mathrm{dist}_{X,Y}(v_1, v_2)$$

which in either case is of course a contradiction. Hence (10).

Now take an $\varepsilon > 0$ and consider as in Case I the ε-thinned percolation process $X_\varepsilon \in \{0, 1\}^E$. Using X_ε, we define the subgraph $H_\varepsilon = (W, F_\varepsilon)$ obtained from H by deleting each $e = \langle v, w \rangle \in F$ such that v and w fail to be in the same connected component of X_ε. By the definition of $p_{c,bond}$ we have that X_ε has no infinite clusters, so that H_ε has no infinite clusters either.

For $v \in W$, write $K_\varepsilon(v)$ for the set of vertices in W that belong to the same connected component of H_ε as v. Also write $\partial_{int} K_\varepsilon(v)$ (*int* as in "internal boundary") for the set of vertices in $K_\varepsilon(v)$ that have at least one neighbor in H which is not in $K_\varepsilon(v)$.

Now define the following mass transport on $G = (V, E)$. No vertex $v \in V$ sends any mass unless it is an encounter point, i.e. unless $v \in W$. Each encounter point v sends unit mass and divides it equally amongst the vertices in $K_\varepsilon(v)$. This defines the mass transport, and the mass received at v becomes

$$\begin{cases} \frac{|K_\varepsilon(v)|}{|\partial_{int} K_\varepsilon(v)|} & \text{if } v \text{ is an encounter point and belongs to } \partial_{int} K_\varepsilon(v) \\ 0 & \text{otherwise.} \end{cases}$$

Since F is a forest in which each vertex has degree is at least 3, its isoperimetric constant is easily seen to be at least 1 (which holds with equality on the binary tree $\mathbb{T}_2$), and similarly $|K_\varepsilon(v)|/|\partial_{int} K_\varepsilon(v)| \leq 2$. So the expected mass received at v is bounded by $2\mathbb{P}(v \in W, v \in \partial_{int} K_\varepsilon(v))$, while the expected mass sent from $v \in V$ equals $\mathbb{P}(v \in W)$. Hence the Mass-Transport Principle gives

$$\mathbb{P}(v \in W) \;\leq\; 2\mathbb{P}(v \in W, v \in \partial_{int} K_\varepsilon(v))\,. \tag{11}$$

But similarly as in the argument for (8) we get for any $v \in W$ that $v \notin \partial_{int} K_\varepsilon(v)$ for all sufficiently small ε, so that

$$\lim_{\varepsilon \to 0} 2\mathbb{P}(v \in W, v \in \partial_{int} K_\varepsilon(v)) \;=\; 0\,.$$

Since (11) was shown to hold for any $\varepsilon > 0$, we get $\mathbb{P}(v \in W) = 0$, which contradicts $N = \infty$, so Case II is finished and the proof is complete. □

5 Random walks on percolation clusters

One of the most natural probabilistic objects, besides i.i.d. percolation, to define on a graph $G = (V, E)$, is simple random walk (SRW), which is a V-valued random process $\{Z_0, Z_1, \ldots\}$ where one (typically) takes $Z_0 = \rho$ for some pre-specified choice of $\rho \in V$, and then iterates the following: given $Z_0, Z_1, \ldots, Z_{n-1}$, the value of Z_n is chosen uniformly among the neighbors in G of Z_{n-1}.

In contrast to the study of percolation on nonamenable Cayley graphs and related classes of graphs which began to take off only in the 1990's, the literature on SRWs on such graphs goes back much further. An early seminal contribution is the work of Kesten [39, 40] from the late 1950's showing that if G is a Cayley graph for a finitely generated group, then the return probability $\mathbb{P}[Z_n = \rho]$ decays exponentially if and only if G is nonamenable. See, e.g., Woess [81] for an introduction to this field.

Another topic of considerable interest is the study of SRW on percolation clusters. For the $\mathbb{Z}^d$ case, see for instance papers like [23] and [17] on central limit theorems, and the paper [28] which extends Polya's classical $d = 2$ versus $d \geq 3$ recurrence-transience dichotomy for random walk on $\mathbb{Z}^d$ to the case of supercritical percolation on $\mathbb{Z}^d$.

Given these traditions, it was a very natural step for Schramm and his collaborators to go on to consider random walks on percolation clusters on nonamenable graphs. Their work is of two kinds: on one hand, the analysis of SRW on a percolation cluster as a worthwhile object of study in its own right, and, on the other hand, the exploitation of random walk on a percolation cluster as a means towards understanding properties of percolation clusters that do not primarily have anything to do with random walk. A remarkable application of the second kind will be described in Section 6 on so-called cluster indistinguishability, while in the present section I will recall a result of the first kind (Theorem 5.1 below) from a rich paper by Benjamini, Lyons and Schramm, henceforth BLS [12].

A natural question to ask for SRW on an infinite graph G is how fast it escapes from the starting point ρ, i.e. how fast does $\mathrm{dist}_G(\rho, Z_n)$ grow? Define the **speed**

$$S \;=\; \lim_{n \to \infty} \frac{\mathrm{dist}_G(\rho, Z_n)}{n}$$

provided the limit exists. When G is the $\mathbb{Z}^d$ lattice, $\mathrm{dist}_G(\rho, Z_n)$ scales like $\sqrt{n}$, and not surprisingly $S = 0$ a.s. More generally, when G is any transitive graph, the

Subadditive Ergodic Theorem immediately implies that the limit S exists and is an a.s. constant.

If we go on on to consider the speed of SRW on an infinite cluster of, say, i.i.d. bond percolation on G (still with the speed defined with respect to dist_G), then the existence of the speed S is less obvious. However, BLS [12] showed, when G is unimodular, and the percolation process is automorphism invariant, that the speed does exist a.s. and does not depend on the random walk, but only on the percolation configuration. Having come that far, it is easy to see that the speed cannot depend on where in an infinite cluster the random walk starts, so each infinite cluster has a well-defined characteristic SRW speed. For the i.i.d. bond percolation case we can then invoke the cluster indistinguishability result of Lyons and Schramm [56] (Theorem 6.1 below) to deduce that all infinite clusters have the same SRW speed.

Of particular interest is to determine whether the speed is zero or positive. For the nonamenable unimodular case BLS [12] found a general answer:

Theorem 5.1 *The speed S of SRW on an infinite cluster of an i.i.d. bond percolation X on a unimodular nonamenable transitive graph G satisfies $S > 0$ a.s.*

An important step in the proof of this is the following result (also of independent interest) from [12] on the geometry of infinite clusters. That the nonamenability of G should be inherited by the infinite clusters of X is too much to hope for: a sufficient condition for a percolation cluster in X to be amenable is that it contains arbitrarily long "naked" paths, i.e., paths of vertices with X-degree 2, and it can be shown that a.s. all infinite clusters arising from i.i.d. percolation on a transitive graph contain such paths. But the infinite clusters of X do contain nonamenable subgraphs:

Theorem 5.2 *Any infinite cluster in i.i.d. bond percolation X on a unimodular nonamenable transitive graph G contains a nonamenable subgraph a.s.*

The proof in [12] of this result involves yet another application of mass transport. Both Theorem 5.2 and Theorem 5.1 are in fact proved more generally than just for i.i.d. percolation. Automorphism invariance alone does not suffice (each of Ex. 3.1, 3.2 and 3.3 of Häggström [31] shows this, and moreover that the $>$ in condition (d) below cannot be replaced by a $\geq$), but if we add any of the conditions

(a) X is i.i.d.,

(b) X has a unique infinite cluster a.s.,

(c) the infinite clusters of X have at least 3 (hence infinitely many) ends a.s., or

(d) X is ergodic with sufficiently large values of $\mathbb{P}(e \text{ is open})$, more precisely the expected degree of a vertex should strictly exceed the quantity h_E^* defined in (5),

then the conclusions of Theorems 5.2 and 5.1 hold; cf. Thms 3.9 and 4.4 in BLS [12].

6 Cluster indistinguishability

Consider an i.i.d. bond percolation X with parameter $p > p_{c,bond}(C)$ on a graph G, so that X produces one or more infinite clusters. It is then natural to ask questions about properties of these infinite clusters. Properties that we have already discussed

in earlier sections include the number of ends of an infinite cluster, whether it contains encounter points, and the speed of SRW on the cluster.

Yet another natural such property of an infinite cluster C is the value of $p_{c,bond}(C)$, i.e., how much further i.i.d. edge-thinning can the infinite cluster C take before it breaks apart into finite components only? For G transitive, it is known (see Häggström and Peres [35] for the unimodular case, and Schonmann [68] for the full result) that a.s. $p_{c,bond}(C) = p_{c,bond}(G)/p$ for all infinite clusters C of X.

All these properties are examples of **invariant properties**. For $G = (V, E)$ transitive with automorphism group $\mathrm{Aut}(G)$, a property (which may or may not hold for clusters of bond percolation on G) can be identified with a Borel measurable subset of $\{0,1\}^E$, and a property $\mathcal{A} \subset \{0,1\}^E$ is said to be invariant if for all $\omega \in \mathcal{A}$ and all $\gamma \in \mathrm{Aut}(G)$ we have $\gamma\omega \in \mathcal{A}$. Lyons and Schramm [56] proved the following Theorem 6.1, known as **cluster indistinguishability**. Shortly before [56], a weaker result for the case of so-called increasing invariant properties was established in [35].

Theorem 6.1 *Let $G = (V, E)$ be a nonamenable unimodular transitive graph, and consider i.i.d. bond percolation X on G with p in the parameter regime where X produces infinitely many infinite clusters a.s. Then, for any invariant component property $\mathcal{A}$, we have a.s. that either all infinite components of X satisfy $\mathcal{A}$ or all infinite components of X satisfy $\neg\mathcal{A}$.*

Space does not permit me to give the full proof of this beautiful result, but I can explain what the main steps are.

Sketch proof of Theorem 6.1. Let $G = (V, E)$ and the percolation $X \in \{0,1\}^E$ be as in the theorem, and assume for contradiction that $\mathcal{A}$ is an invariant property such that with positive probability, X contains both infinite clusters with property $\mathcal{A}$ and infinite clusters with property $\neg\mathcal{A}$.

Step I. Existence of pivotal edges. For an infinite cluster C of X and an edge $e \in E$ with $X(e) = 0$ that has an endpoint in C, call e **pivotal** for C if either $C \in \mathcal{A}$ and switching on the edge e would create an infinite cluster (containing C) with property $\neg\mathcal{A}$, or vice versa. If there exists an $e \in E$ with $X(e) = 0$ that has one endpoint in an infinite cluster with property $\mathcal{A}$ and the other in an infinite cluster with property $\neg\mathcal{A}$, then clearly e is pivotal for one of the clusters. Otherwise, there exists such a pair of clusters within finite distance from each other, and by sequentially switching on one edge after another on a finite path between them we see that somewhere along the way one infinite cluster of type $\mathcal{A}$ must turn into $\neg\mathcal{A}$ or vice versa. (This is an example of a well-known technique in percolation theory known as **local modification**, pioneered by Newman and Schulman [61] and Burton and Keane [21]; cf. also Coupling 2.5 in [34] for a careful explanation.) Hence pivotal edges exist with positive probability.

Step II. Stationarity of random walk. Consider a SRW $\{Z_0, Z_1, \ldots\}$ on a percolation cluster of X, defined as in Section 5. It would be nice to think that the percolation configuration "as seen from the point of view of the walker", would be stationary. To make this precise, fix for each $v \in V$ a $\gamma_{v,\rho} \in \mathrm{Aut}(G)$ that maps v on ρ, where, as in Section 5, ρ is the starting point of the random walk. The idea of stationarity of the percolation configuration as seen from the random walker is that for any Borel measurable $\mathrm{Stab}(\rho)$-invariant $\mathcal{B} \in \{0,1\}^E$ and any n we should have

$$\mathbb{P}(\gamma_{Z_n,\rho} X \in \mathcal{B}) \;=\; \mathbb{P}(X \in \mathcal{B})\,.$$

This is not true, however, and to see this, think, e.g., about what happens when the starting point ρ happens to be in a finite open cluster C which is simply a path of length 2; in other words, C has three vertices, one of which has degree 2 and two of which have degree 1. Conditioned on ρ being in such a cluster, it has probability $\frac{1}{3}$ of being in the vertex of degree 2 (this follows from a straightforward mass-transport argument). But SRW on such an open cluster will spend half the time (either all even or all odd times) on that vertex, so it cannot possibly be stationary in the desired sense.

All is not lost, however. Stationarity can be recovered by a minor modification of SRW, namely the **delayed simple random walk (DSRW)**, denoted $\{\tilde{Z}_0, \tilde{Z}_1, \ldots\}$. This is again a V-valued random process, and as with SRW we take $\tilde{Z}_0 = \rho$, but the transition mechanism is slightly different: given $\tilde{Z}_0, \ldots, \tilde{Z}_{n-1}$, a vertex w is chosen according to uniform distribution on the set of $\tilde{Z}_{n-1}$'s G-neighbors, and we set

$$\tilde{Z}_n = \begin{cases} w & \text{if } X(\langle \tilde{Z}_{n-1}, w \rangle) = 1 \\ \tilde{Z}_{n-1} & \text{otherwise.} \end{cases}$$

Such a DSRW has the desired stationarity property, i.e.,

$$\mathbb{P}(\gamma_{\tilde{Z}_n,\rho} X \in \mathcal{B}) = \mathbb{P}(X \in \mathcal{B}) .$$

for any $\mathcal{B}$ and any n, and this can be shown via yet another mass-transport argument.

(Stationarity of DSRW is clearly an interesting result in its own right, and the idea was first exploited in [31]. Lyons and Schramm decided to put their most general version of the stationarity result not in [56] but in a separate paper [57]. Without unimodularity, however, stationarity fails, as is easily seen by considering DSRW in Example 3.4: here, $\tilde{Z}_0$ has probability $\frac{1}{2}$ of sitting in the "topmost" (closest to ξ) component of its open cluster, while the probability that $\tilde{Z}_n$ does so tends to 0 as $n \to \infty$.)

Step III. Lots of encounter points. Since X has infinitely many infinite clusters, we can modify X by changing finitely many edges so as to connect three of them to form an encounter point. Hence encounter points exist with positive probability; this is just the local modification argument of Burton and Keane [21]. By (9), we have that every infinite cluster with encounter points must in fact have infinitely many. From this it follows that *every* infinite cluster has infinitely many encounter points, because otherwise we could use the local modification technique to connect an infinite cluster with encounter points to one without and obtain a contradiction to (9).

Step IV. Random walk is transient. Given the prevalence of encounter points, infinite clusters contain, loosely speaking, a large-scale structure similar to the binary tree $\mathbb{T}_2$. This strongly suggests that DSRW on such an infinite cluster should be transient. Lyons and Schramm [56] converts this intuition into a proof by combining the result mentioned in the final paragraph of Section 3 about such infinite clusters C having $p_{c,bond}(C) < 1$, with a basic comparison of random walk and percolation on trees due to Lyons [50]. (Alternatively, we could quote Theorem 5.1 here, but that would be a detour.)

Step V. Local modification applied to a pivotal edge. Let $\mathcal{B}$ be the event that the starting point ρ of the DSRW is in an infinite cluster with property $\mathcal{A}$. Given some small $\varepsilon > 0$, we can find a $k < \infty$ and a $\mathrm{Stab}(\rho)$-invariant event $\mathcal{B}^*$ that depends only on edges within G-distance k from ρ, approximating $\mathcal{B}$ in the sense that

$$\mathbb{P}(\mathcal{B} \triangle \mathcal{B}^*) < \varepsilon . \tag{12}$$

Define, for each n,

$$Y_n = \begin{cases} 1 & \text{if } \gamma_{\tilde{Z}_n,\rho} X \in \mathcal{B}^* \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

Then $\{Y_0, Y_1, \ldots\}$ is a stationary process, so the limit $\bar{Y} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} Y_n$ exists a.s. Furthermore,

$$\text{the limit } \bar{Y} \text{ depends only on the percolation configuration and not on the DSRW.} \tag{14}$$

To see (14), define a second DSRW $\{\tilde{Z}_0', \tilde{Z}_1', \ldots\}$ from ρ which conditionally on X is independent of $\{\tilde{Z}_0, \tilde{Z}_1, \ldots\}$, define $\{Y_0', Y_1', \ldots\}$ analogously as in (13), and set $\bar{Y}' = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} Y_n'$. Then we can define $Y_{-i} = Y_i'$ for each $i \geq 1$, and it turns out (see [56], Lem. 3.13) that the two-sided sequence $\{\ldots, Y_{-1}, Y_0, Y_1, \ldots\}$ becomes stationary. Now, if (14) failed we would with positive probability have $\bar{Y} \neq \bar{Y}'$, which however would contradict stationarity of the two-sided sequence $\{\ldots, Y_{-1}, Y_0, Y_1, \ldots\}$ in view of the ergodic theorem. Hence (14).

Thus we can assign each infinite cluster C a value $\alpha(C)$ as the a.s. value of $\bar{Y}$ that DSRW on that cluster would produce. By (12) we have that infinite clusters with different values of $\alpha(C)$ coexist with positive probability, provided we chose ε small enough to begin with. Call an edge $e \in E$ **α-pivotal** if it is closed ($X(e) = 0$) with endpoints in two different infinite clusters C and C' with

$$\alpha(C) \neq \alpha(C') \,. \tag{15}$$

By the reasoning in Step 1, there exist α-pivotal edges in X with positive probability. In particular, there exists an $e \in E$ which with positive probability is α-pivotal for the infinite cluster containing ρ. Suppose this happens, let C be the infinite cluster containing ρ, and let C' be the other infinite cluster meeting e. Then $\bar{Y} = \alpha(C)$ a.s.

But look what happens if we apply local modification to the edge e. If we turn e on ($X(e) = 1$) and leave the status of all other edges intact, we get a new infinite cluster C'' uniting C and C'. What will then happen with $\bar{Y}$? By transience of DSRW, we get with positive probability that the DSRW escapes to infinity without ever noticing the edge e, and we get a.s. on this event that $\bar{Y} = \alpha(C)$; this uses the fact that Y_n is a function of edges within bounded distance k from $\tilde{Z}_n$ only. On the other hand, we get with positive probability that the DSRW reaches e, crosses it, and then escapes to infinity without ever crossing it back; on this event we get a.s. $\bar{Y} = \alpha(C')$. In view of (15), this contradicts (14) and proves the theorem. □

An inspection of the proof to determine which properties of i.i.d. percolation are actually used reveals that it is enough to assume automorphism invariance plus so-called **insertion tolerance**, which is the term Lyons and Schramm [56] used for a refinement of the finite energy property considered by Newman and Schulman [61], Burton and Keane [21], and others.

Definition 6.2 *A bond percolation X on a graph $G = (V, E)$ is called* **insertion tolerant** *if it admits conditional probabilities such that for every $e \in E$ and every $\xi \in \{0,1\}^{E\setminus\{e\}}$ we have $\mathbb{P}(X(e) = 1 | X(E \setminus \{e\} = \xi) > 0$. If instead $\mathbb{P}(X(e) = 0 | X(E \setminus \{e\} = \xi) > 0$ for all such e and ξ, then X is called* **deletion tolerant**.

(The conjunction of insertion tolerance and deletion tolerance is precisely the finite energy property.)

Theorem 6.3 *Let $G = (V, E)$ be a nonamenable unimodular transitive graph, and consider an insertion tolerant automorphism invariant bond percolation X on G. Then, for any invariant component property $\mathcal{A}$, we have a.s. that either all infinite components of X are in $\mathcal{A}$ or no infinite components of X are in $\mathcal{A}$.*

In fact, the formulation in [56] is even more general than this: here, as in much of the work discussed in Section 3, invariance under the full automorphism group can be weakened to invariance under certain subgroups. Also, the result holds with site percolation in place of bond percolation.

It is worth mentioning some limitations to the scope of cluster indistinguishability. For instance, insertion tolerance cannot be replaced by deletion tolerance in Theorem 6.3, as the following example from [56] shows.

Example 6.4 *Let $G = (V, E)$ be the binary tree $\mathbb{T}_2$, recall that $p_{c,bond}(T_2) = \frac{1}{2}$, and fix p_1 and p_2 such that $\frac{1}{2} < p_1 < p_2 < 1$. Let $X \in \{0,1\}^E$ be i.i.d. bond percolation on G with parameter p_2. Then obtain another bond percolation process $X' \in \{0,1\}^E$ from X as follows. For each infinite cluster C of X independently, toss a fair coin. If the coin comes up heads, delete each edge of C independently with probability $1 - \frac{p_1}{p_2}$; otherwise let all of C's edges be intact. While X' is automorphism invariant and deletion tolerant, some of its infinite clusters C' will have all the characteristics of those produced by i.i.d.(p_2) percolation and thus have $p_{c,bond}(C') = \frac{1}{2p_2}$, while others will look like i.i.d.(p_1) percolation clusters and thus have $p_{c,bond}(C') = \frac{1}{2p_1}$, so cluster indistinguishability fails for X'.*

How about the unimodularity assumption in Theorems 6.1 and 6.3? This cannot be dropped, not even from Theorem 6.1. To see this, let $G = (V, E)$ be Trofimov's graph (Example 3.3), and consider i.i.d. bond percolation with parameter $p \in (p_{c,bond}(G), 1)$. Each infinite cluster C then has a "topmost" vertex $w(C)$ (in the direction of the designated end ξ), and for $i = 0, 1, 2$ we may define the property $\mathcal{A}_i$ by stipulating that $C \in \mathcal{A}_i$ if $w(C)$ is directly linked to exactly i of its ξ-children via open edges. Then $\mathcal{A}_0$, $\mathcal{A}_1$ and $\mathcal{A}_2$ are $\mathrm{Aut}(G)$-invariant properties, and the percolation will a.s. produce infinite clusters of all three kinds, so cluster indistinguishability fails.

7 In the hyperbolic plane

In their *Percolation beyond $\mathbb{Z}^d$* paper [13], Benjamini and Schramm emphasized three properties of graphs that could be expected to be of particular relevance to the behavior of percolation processes: quasi-transitivity, (non-)amenability, and planarity. The first two I have discussed at some length, while the third was only mentioned in passing in Section 2. In this section, I will briefly make up for this.

Planarity plays a crucial role in the classical study of percolation on the $\mathbb{Z}^2$ lattice, such as in the seminal contributions by Harris [37] and Kesten [41]. A key device is the notion of planar duality: For a graph $G = (V, E)$ with a planar embedding in $\mathbb{R}^2$ (or in some other two-dimensional manifold), it is often useful to define its dual graph $G^\dagger = (V^\dagger, E^\dagger)$ by identifying $V^\dagger$ with the faces of the planar embedding of G, and including an edge $e^\dagger \in E^\dagger$ crossing each $e \in E$. If $X \in \{0,1\}^E$ a bond percolation on G, then we can define a bond percolation $X^\dagger \in \{0,1\}^E$ on $G^\dagger$ by declaring, for each $e \in E$, $X^\dagger(e^\dagger) = 1 - X(e)$. If X is i.i.d.(p), then $X^\dagger$ becomes i.i.d.$(1-p)$. Furthermore, if G is the $\mathbb{Z}^2$ lattice, then $G^\dagger$ is isomorphic to G, and for $p = 1 - p = \frac{1}{2}$

the distributions of X and $X^\dagger$ will be the same; this observation is basic to proving the Harris–Kesten theorem that $p_{c,bond}(\mathbb{Z}^2) = \frac{1}{2}$.

Suppose that G is infinite, planar and transitive. It is known (see Babai [5]) that such a graph, equipped with the usual distance dist_G, is quasi-isometric to exactly one of the four spaces $\mathbb{R}$, $\mathbb{R}^2$, $\mathbb{T}_2$ and the hyperbolic plane $\mathbb{H}^2$, the last of which may be defined as the open unit disk $\{z \in \mathbb{C} : |z| < 1\}$ equipped with the metric s given by

$$ds^2 = \frac{dx^2 + dy^2}{(1 - x^2 - y^2)^2}\,.$$

In the paper *Percolation in the hyperbolic plane* [16], Benjamini and Schramm consider the situation where G has one end, in which case it must be quasi-isometric to either $\mathbb{R}^2$ or $\mathbb{H}^2$. Which of these is determined by amenability: if G is amenable, then $\mathbb{R}^2$, while if it is nonamenable, then $\mathbb{H}^2$. The focus of [16] is on the nonamenable case; hence the title of the paper. A happy circumstance here is that a planar nonamenable transitive graph with one end is also unimodular (Prop. 2.1 in [16]), which allows the machinery developed in BLPS [9] to come into play.

Recall the Newman–Schulman 0-1-∞ law about the number of infinite clusters on transitive graphs. This result needs the percolation process to be automorphism invariant and insertion tolerant (cf. Definition 6.2). In general, insertion tolerance cannot be dropped (not even on the $\mathbb{Z}^2$ lattice, see Burton and Keane [22]), although in the present setting, remarkably enough, it can:

Theorem 7.1 (Thm 8.1 in [9]) *If $G = (V, E)$ is a planar nonamenable quasi-transitive graph with one end, then a.s. any automorphism invariant bond percolation $X \in \{0,1\}^E$ has 0, 1, or infinitely many infinite clusters.*

Proof. This proof from [16] differs somewhat from the original one in [9]. There is no loss of generality in assuming that the number of infinite clusters is an a.s. constant k. Assume for contradiction that $k \in \{2, 3, \ldots\}$. By randomly deleting all edges of all infinite clusters but two, chosen uniformly at random, we still preserve automorphism invariance; thus we may assume $k = 2$. Call one of them C_1 and the other C_2, using a fair coin toss to decide which is which. Then turn on each edge e which does not meet C_2 and from which there is a path in G to C_1 that doesn't meet C_2. This preserves automorphism invariance, and expands C_1 to a larger infinite cluster $\hat{C}_1$ that, loosely speaking, sits as close to C_2 as is possible without touching it. Now define a bond percolation $\hat{X}^\dagger$ on the dual graph $G^\dagger$ (which is also planar, nonamenable, quasi-transitive and one-ended) by turning on exactly those edges $e^\dagger \in E^\dagger$ whose corresponding $e \in E$ are pivotal for connecting $\hat{C}_1$ to C_2. Then $\hat{X}^\dagger$ is $\mathrm{Aut}(G^\dagger)$-invariant, and consists (due to one-endedness of G) of a single bi-infinite open path. Hence $\hat{X}^\dagger$ has a unique infinite cluster $C^\dagger$ with $p_{c,bond}(C^\dagger) = 1$. On the other hand, the reasoning in the proof of Theorem 4.1, Case I, shows that any unique infinite cluster arising from automorphism invariant percolation in such a graph has $p_{c,bond} < 1$, and this is the desired contradiction. □

From here, Benjamini and Schramm [16] go on to show a number of interesting results for percolation in the hyperbolic plane. For starters, let G be as in Theorem 7.1, let $G^\dagger$ be its planar dual, let X be an invariant bond percolation process on G, let $X^\dagger$ be its dual, and let k and $k^\dagger$ be their (possibly random) number of infinite clusters. Then Theorem 7.1 leaves nine possibilities for the value of $(k, k^\dagger)$: $(0,0), (0,1), (0,\infty), (1,0), (1,1), (1,\infty), (\infty,0), (\infty,1)$ and (∞,∞), but the following result rules out four of them.

Theorem 7.2 (Thm 3.1 in [16]) *With G, $G^\dagger$, X and $X^\dagger$ as above, we have a.s. that*

$$(k, k^\dagger) \in \{(0,1), (1,0), (1,\infty), (\infty,1), (\infty,\infty)\}\,. \tag{16}$$

All five outcomes in (16) can actually happen. The cases $(1,\infty)$ and $(\infty,1)$ arise in the so-called uniform spanning forest model (see Theorem 8.2 in the next section), but cannot arise in i.i.d. percolation because of a local modification argument (Thm 3.7 in [16]) that can turn $(k, k^\dagger) = (1,\infty)$ into $(2,\infty)$, contradicting Theorem 7.1. Benjamini and Schramm show that, in fact,

$$\begin{gathered}\text{outcomes } (0,1),\ (1,0) \text{ and } (\infty,\infty) \text{ are exactly} \\ \text{those that happen for i.i.d. percolation,}\end{gathered} \tag{17}$$

and they prove the following remarkable result, establishing Conjecture 2.7 for the case of planar hyperbolic graphs:

Theorem 7.3 (Thm 1.1 in [16]) *Let $G = (V, E)$ be a planar nonamenable transitive graph with one end. Then $0 < p_{c,bond}(G) < p_{u,bond}(G) < 1$. The same is true for site percolation.*

This generalizes a result of Lalley [43] who showed $0 < p_{c,site}(G) < p_{u,site}(G) < 1$ for a more restrictive class of graphs. Together with (17), this yields

$$(k, k^\dagger) = \begin{cases} (0,1) & \text{for } p \in (0, p_{c,bond}(G)) \\ (\infty,\infty) & \text{for } p \in (p_{c,bond}(G), p_{u,bond}(G)) \\ (1,0) & \text{for } p \in (p_{u,bond}(G), 1)\,. \end{cases}$$

Concerning the behavior at the critical values, Theorem 4.1 yields $(k, k^\dagger) = (0,1)$ for $p = p_{c,bond}(G)$, and by exchanging the roles of G and $G^\dagger$ we see that $(k, k^\dagger) = (1,0)$ at $p = p_{u,bond}(G)$. As mentioned in Section 2 this uniqueness of the infinite cluster already *at* the uniqueness critical value contrasts with the behavior obtained for certain other nonamenable transitive graphs by Schonmann [69] and Peres [65].

In fact, similar considerations for the hypothetical scenario that Theorem 7.3 fails show how smoothly Theorem 7.3 follows from (17). We would then have $(k, k^\dagger) = (0,1)$ for $p < p_{c,bond}(G)$ and $(k, k^\dagger) = (1,0)$ for $p > p_{c,bond}(G)$. Hence $p_{c,bond}(G^\dagger) = 1 - p_{c,bond}(G)$, and Theorem 4.1 applied to both G and $G^\dagger$ yields $(k, k^\dagger) = (0,0)$ at $p = p_{c,bond}(G)$. This, however, contradicts (17).

Benjamini and Schramm [16] go on to study properties of the infinite clusters and their limit points on the boundary of the hyperbolic disk (see also Lalley [43, 44] for further results in this direction). The final part of their paper [16] concerns i.i.d. site percolation on a certain random lattice in $\mathbb{H}^2$, namely the Delaunay triangulation of a Voronoi tesselation for a Poisson process with intensity $\lambda > 0$ in $\mathbb{H}^2$. The counterpart in $\mathbb{R}^d$ of such a model has also been studied; see Bollobás and Riordan [18, 19] for a recent breakthrough in $\mathbb{R}^2$. However, the phase diagram becomes richer and more interesting in the $\mathbb{H}^2$ case, not just because of the nonuniqueness phase between $p_{c,site}$ and $p_{u,site}$, but also because it becomes a true two-parameter family of models: in $\mathbb{R}^d$, changing λ is just a trivial rescaling of the model, while in $\mathbb{H}^2$ there is no such scale invariance.

More recently, Tykesson [78] considered a different way of doing percolation in $\mathbb{H}^2$ based on a Poisson process, namely to place a ball of a fixed hyperbolic radius R around each point, and consider percolation properties of the region covered by the union of the balls. (This is the so-called Boolean model of continuum percolation,

which has received a fair amount of attention in $\mathbb{R}^d$; see, e.g., Meester and Roy [58].) Equally natural is to consider percolation properties of the vacant region (i.e., the complement of the covered region). To consider percolation properties of both regions simultaneously is analogous to working with the pair $(X, X^\dagger)$ in the discrete lattice setting. The results in [78] turn out mostly analogous to those of Benjamini and Schramm [16] discussed above for the discrete lattice setting. However, in an even more recent paper by Benjamini, Jonasson, Schramm and Tykesson [7], a phenomenon which is particular to the continuum setting is revealed. Namely, are there hyperbolic lines entirely contained in the vacant region, so that someone living in $\mathbb{H}^2$ can actually "see infinity" in some (necessarily random) directions? In $\mathbb{R}^d$ the answer to the analogous question is no (this is related to Olbers' paradox in astronomy; see, e.g., Harrison [38]) while in $\mathbb{H}^2$ the answer turns out to be yes in certain parts of the parameter space. The paper [7] – which, sadly, became one of the last by Oded Schramm – also contains results on various refinements and variants of this question.

8 Random spanning forests

So far, a lot has been said about percolation on nonamenable transitive graphs under the general assumption of automorphism invariance, but hardly anything about particular examples beyond the i.i.d. cases (other than a few examples specifically designed to be counterexamples). But as mentioned in Section 3, there *are* plenty of important examples, and time has come to discuss one of them: the uniform spanning forest. Later in this section, I will go on to discuss its cousin, the *minimal* spanning forest.

A spanning tree for a finite connected graph $G = (V, E)$ is a connected subgraph containing all vertices but no cycles. A **uniform spanning tree** for G is one chosen at random according to uniform distribution on the set of possible spanning trees. Replacing G by, say, the $\mathbb{Z}^d$ lattice, the number of possible spanning trees skyrockets to ∞, and it is no longer obvious how to make sense of the uniform spanning tree. Pemantle [64] managed to make sense of it: For any infinite locally finite connected graph $G = (V, E)$, let $\{G_1 = (V_1, E_1), G_2 = (V_2, E_2), \ldots\}$ be an increasing sequence of finite connected subgraphs of G, which exhausts G in the sense that each $e \in E$ and each $v \in V$ is in all but at most finitely many of the G_i's. Then, it turns out, the uniform spanning tree measures for $G_1, G_2, \ldots$ converge weakly to a probability measure μ_G on $\{0,1\}^E$ which is independent of the exhaustion $\{G_1, G_2, \ldots\}$. In particular, μ_G is $\mathrm{Aut}(G)$-invariant. Furthermore, it is concentrated on the event that there are no open cycles and no finite open clusters, so that in other words what we get is μ_G-a.s. a forest, all of whose trees are infinite. Naively, one might expect to get a single tree, but this is not always the case. Pemantle showed for the $\mathbb{Z}^d$ case that the number of trees is an a.s. constant N satisfying

$$N = \begin{cases} 1 & \text{if } d \leq 4 \\ \infty & \text{otherwise.} \end{cases} \tag{18}$$

This $d \leq 4$ vs $d > 4$ dichotomy is related to the fact that two independent SRW trajectories on the $\mathbb{Z}^d$ lattice intersect a.s. if and only if $d \leq 4$ (though this innocent-looking statement hides the fact that Pemantle [64] had to use deep results by Lawler [45] on so-called loop-erased random walk; the proof was simplified in the paper [11] to be discussed below). The behavior for $d > 4$ suggests **uniform spanning forest** as a better term than uniform spanning tree when G is infinite. The analysis of

the uniform spanning forest μ_G builds to a large extent on the beautiful collection of identities between uniform spanning trees, electrical networks and random walks which has a long and disperse history beginning with the 1847 paper by Kirchhoff [42]. For instance, Rayleigh's Monotonicity Principle for effective resistances in electrical networks (see, e.g., Doyle and Snell [24]) underlies the stochastic monotonicity properties of the sequence $\mu_{G_1}, \mu_{G_2}, \ldots$ that allows us to deduce the existence of the limiting measure μ_G.

Some years after Pemantle's [64] pioneering 1991 paper, and at about the same time that BLPS [9] and BLPS [10] were written, the BLPS quartet started getting seriously interested in uniform spanning forests; see Lyons [54] for a more exact statement about the timing and relation betweens the various BLPS projects. This resulted in the magnificent paper *Uniform spanning forests* by Benjamini, Lyons, Peres and Schramm [11], which provides a unified treatment and a number of simplications of what was known on uniform spanning trees and forests, together with a host of new and important results. The methods involve, in addition to the aforementioned connections to random walks and electrical networks, also mass-transport ideas, Hilbert space projections, and Wilson's [80] substantial improvement on the Aldous–Broder algorithm [2, 20] for generating uniform spanning trees. Here I will mention only a couple of the results from BLPS [11], but see Lyons [52] for a gentle introduction to the same topic (note that despite the inverted publication dates, the 1998 paper [52] surveys much of the original work in the 2001 paper [11]).

It turns out that there is another, equally natural, way to obtain a uniform spanning forest for an infinite graph $G = (V, E)$ via the exhaustion $\{G_1, G_2, \ldots\}$ considered above, namely if for each i we consider the uniform spanning tree not on G_i but on the modified graph where an extra vertex w is introduced, together with edges $\langle v, w \rangle$ for all $v \in V_i \setminus V_{i-1}$. We think of this as a "wired" version of G_i, and the limiting measure on $\{0,1\}^E$, which is denoted $\mathbf{WSF}_G$ and which exists for similar reasons as for μ_G, should be thought of as the **wired uniform spanning forest** for G. For consistency of terminology, we call μ_G the **free uniform spanning forest**, and switch to denoting it $\mathbf{FSF}_G$. (The free and wired uniform spanning forests are analogous to the free and wired limiting measures for the random-cluster model; cf. [27].) The wired measure $\mathbf{WSF}_{\mathbb{Z}^d}$ appeared implicitly in Pemantle [64] together with the result that $\mathbf{WSF}_{\mathbb{Z}^d} = \mathbf{FSF}_{\mathbb{Z}^d}$; this was made explicit in Häggström [30], and BLPS [11] noted that this extends to $\mathbf{WSF}_G = \mathbf{FSF}_G$ whenever G is transitive and amenable. On the other hand, $\mathbf{WSF}_{\mathbb{T}_d} \neq \mathbf{FSF}_{\mathbb{T}_d}$ for the $(d+1)$-regular tree $\mathbb{T}_d$ when $d \geq 2$, so a first guess might be that for transitive graphs, $\mathbf{WSF}_G = \mathbf{FSF}_G$ is equivalent to amenability. This turns out not to be true, however, and BLPS [11] offer instead the following characterization (which does not require G to be transitive or even quasi-transitive). Recall that for a graph $G = (V, E)$ a function $f : V \to \mathbb{R}$ is called **harmonic** if for any $v \in V$ we have that $f(v)$ equals the average of $f(w)$ among all its neighbors w, and that f is called **Dirichlet** if $\sum_{\langle u,v \rangle \in E} (f(u) - f(v))^2 < \infty$.

Theorem 8.1 (Thm 7.3 in [11]) *For any graph G, we have $\mathbf{WSF}_G = \mathbf{FSF}_G$ if and only if G admits no nonconstant harmonic Dirichlet functions.*

As an example of a nonamenable transitive graph for which $\mathbf{WSF}_G = \mathbf{FSF}_G$ holds, we may take the Grimmett–Newman example discussed in Section 2; this follows from Theorem 8.1 in combination with the result of Thomassen [75, 74] that the Cartesian product of any two infinite graphs has no nonconstant harmonic Dirichlet functions.

A nice class of graphs where $\mathbf{WSF}_G \neq \mathbf{FSF}_G$ are the planar hyperbolic lattices considered in the previous section:

Theorem 8.2 (Thms 12.2 and 12.7 in BLPS [11]) *For any planar nonamenable transitive graph $G = (V, E)$ with one end, we have that* $\mathbf{WSF}_G \neq \mathbf{FSF}_G$. *In particular, they differ in the number of infinite clusters:* $\mathbf{FSF}_G$ *produces a unique infinite cluster a.s., while* $\mathbf{WSF}_G$ *produces infinitely many a.s. If $G^\dagger = (V^\dagger, E^\dagger)$ is the planar dual of G, and $X \in \{0,1\}^E$ is given by* $\mathbf{WSF}_G$, *then the dual percolation $X^\dagger \in \{0,1\}^{E^\dagger}$ defined as in Section 7, has distribution* $\mathbf{FSF}_{G^\dagger}$.

There is a lot more to quote from BLPS [11] concerning uniform spanning forests, but I will instead cite a result from a later paper by Benjamini, Kesten, Peres and Schramm [8]. Strictly speaking the result falls outside the scope of the present survey as I defined it in Section 1, because it concerns $\mathbb{Z}^d$, but it is so supremely beautiful that I find this inconsistency of mine to be motivated. For a bond percolation process $X \in \{0,1\}^E$ on a graph $G = (V, E)$, and two vertices $u, v \in V$, define the random variable $D_{closed}(u, v)$ as the minimal number of closed edges that a path in G from u to v needs to traverse, and D^{max}_{closed} as the supremum of $D_{closed}(u, v)$ over all choices of $u, v \in V$. For percolation processes such as $\mathbf{WSF}_G$ and $\mathbf{FSF}_G$ where a.s. all vertices belong to infinite clusters, $D^{max}_{closed} = 0$ means precisely uniqueness of the infinite cluster, while $D^{max}_{closed} = 1$ means that uniqueness fails but that any pair of infinite clusters come within unit distance from each other somewhere in G. The dichotomy by Pemantle quoted in (18) says that for uniform spanning forests on $\mathbb{Z}^d$, we have $D^{max}_{closed} = 0$ a.s. for $d \leq 4$ but $D^{max}_{closed} \geq 1$ a.s. when $d > 4$. Benjamini, Kesten, Peres and Schramm [8] proved the following refinement.

Theorem 8.3 *For the uniform spanning forest measure* $\mathbf{WSF}_{\mathbb{Z}^d}$ *(or equivalently* $\mathbf{FSF}_{\mathbb{Z}^d}$*), we have, a.s.,*

$$D^{max}_{closed} = \begin{cases} 0 & \text{if} \quad d \in \{1,2,3,4\} \\ 1 & \text{if} \quad d \in \{5,6,7,8\} \\ 2 & \text{if} \quad d \in \{9,10,11,12\} \\ 3 & \text{if} \quad d \in \{13,14,15,16\} \\ & \vdots \end{cases}$$

Next, let us return briefly to the case where $G = (V, E)$ is a finite graph. Besides the uniform spanning tree, there is another, much-studied and equally natural, way of picking a random spanning tree for G: attach i.i.d. weights $\{U(e)\}_{e \in E}$ to the edges of G, and pick the spanning tree for G that minimizes the sum of the edge weights; this is the so-called **minimal spanning tree** for G. The marginal distribution of the $U(e)$'s has no effect on the distribution of the tree as long as it is free from atoms, but it turns out to facilitate the analysis to take it to be uniform on $[0, 1]$. It is easy to see that an edge $e \in E$ is included of the minimal spanning tree if and only if $U(e) < \max_{e \in C} U(e')$ for all cycles C containing e.

If instead G is infinite, this characterization can be taken as the definition, and the resulting random subgraph of G is called the **free minimal spanning forest**, the corresponding probability measure on $\{0,1\}^E$ being denoted $\mathbf{FMSF}_G$. An alternative extension to infinite G is to include $e \in E$ if and only if $U(e) < \max_{e \in C} U(e')$ for all *generalized* cycles C containing e, where by a generalized cycle we mean a cycle or a bi-infinite self-avoiding path. This gives rise to the so-called **wired minimal spanning forest**, and a corresponding probability measure $\mathbf{WMSF}_G$ on $\{0,1\}^E$. The study of $\mathbf{FMSF}_G$ and $\mathbf{WMSF}_G$ parallels the study of $\mathbf{FSF}_G$ and $\mathbf{WSF}_G$ in many ways, and I wish to draw the reader's attention to the paper *Minimal spanning*

forests by Lyons, Peres and Schramm [55], where, in the words of the authors, "a key theme [is] to describe striking analogies, and important differences, between uniform and minimal spanning forests".

As for uniform spanning forests, one of the central issues is to determine when $\mathbf{FMSF}_G = \mathbf{WMSF}_G$. And as for uniform spanning forests, equality turns out to hold on $\mathbb{Z}^d$ and more generally on amenable transitive graphs. But moving on to the nonamenable case, the statements $\mathbf{FMSF}_G = \mathbf{WMSF}_G$ and $\mathbf{FSF}_G = \mathbf{WSF}_G$ are no longer equivalent:

Theorem 8.4 (Prop. 3.6 in [55]) *On any G, we have $\mathbf{FMSF}_G = \mathbf{WMSF}_G$ if and only if for Lebesgue-a.e. $p \in [0,1]$ it is the case that i.i.d. bond percolation on G with parameter p produces a.s. at most one infinite cluster.*

Hence we should expect to have $\mathbf{FMSF}_G \neq \mathbf{WMSF}_G$ for nonamenable quasi-transitive G; this is equivalent to Conjecture 2.7. One example where we do know that $\mathbf{FMSF}_G \neq \mathbf{WMSF}_G$ while $\mathbf{FSF}_G = \mathbf{WSF}_G$ is when G is the Grimmett–Newman graph $\mathbb{T}_d \times \mathbb{Z}$; cf. Theorem 8.1 and the comment following it.

Another consequence of having a nonuniqueness phase for i.i.d. bond percolation on G is that uniqueness of the infinite cluster fails for $\mathbf{WMSF}_G$ (this is Cor. 3.7 in [12]). Determining the number of infinite clusters in $\mathbf{WMSF}_G$ (or in $\mathbf{FMSF}_G$) more generally is of course a central issue. On $\mathbb{Z}^d$ for $d \geq 2$ the only case that has been settled is $d = 2$, where Alexander [4] showed that the number of infinite clusters is 1. For higher dimensions, a dichotomy like Pemantle's (18) can be expected; Newman and Stein [62] conjectured that the switch from uniqueness to infinitely many infinite clusters should happen at $d = 8$ or 9.

A general difference – but at the same time a parallel – between the uniform spanning forest and the minimal spanning forest is that while the former has intimate connections to SRW on G, the latter seems equally intimately connected to i.i.d. bond percolation on G, such as in Theorem 8.4. (The key device for exposing the connections between the minimal spanning forest and i.i.d percolation is the coupling in which we use the $U(e)$'s underlying $\mathbf{FMSF}_G$ and $\mathbf{WMSF}_G$ also for generating i.i.d. bond percolation for any p: an edge e is declared open on level p iff $U(e) < p$.) One striking instance of this parallel is the following. Morris [60] showed for any G that a.s. any component C of the wired uniform spanning forest has the property that SRW on it is recurrent, while Lyons, Peres and Schramm [55] showed for any G that a.s. any component C of the wired minimal spanning forest has $p_{c,bond}(C) = 1$. Both results are sharp in the sense that in neither of them can the set of possible C's be narrowed further without making the statement false.

9 Postscript

The reader may have noticed that most of the work surveyed here is from the mid-to-late 1990's, and ask why this is. Is it because Oded got bored with percolation beyond $\mathbb{Z}^d$? No, it isn't, and no, he didn't. The main reason is Oded's discovery in [70] of SLE (Schramm–Loewner evolution, or stochastic Loewner evolution as Oded himself preferred to call it). This led him to the far-reaching and more urgent project, beginning with a now-famous series of papers with Greg Lawler and Wendelin Werner [46, 47, 48, 49], of understanding SLE and how it arises as a scaling limit of various critical models in two dimensions. But that is a different story.

Acknowledgement. I am grateful to Russ Lyons and Johan Tykesson for helpful comments on an earlier draft.

References

[1] Aizenman, M., Kesten, H. and Newman, C.M. (1987) Uniqueness of the infinite cluster and continuity of connectivity functions for short- and long-range percolation, *Comm. Math. Phys.* **128**, 39–55.

[2] Aldous, D. (1990) The random walk construction of uniform spanning trees and uniform labelled trees, *SIAM J. Discrete Math.* **3**, 450–465.

[3] Aldous, D. and Lyons, R. (2007) Processes on unimodular random networks, *Electr. J. Probab.* **12**, 1454–1508.

[4] Alexander, K.S. (1995) Percolation and minimal spanning forests in infinite graphs, *Ann. Probab.* **23**, 87–104.

[5] Babai, L. (1997) The growth rate of vertex-transitive planar graphs, in *Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms (New Orleans, LA, 1997)*, pp 564–573, ACM, New York.

[6] Benjamini, I., Häggström, O. and Schramm, O. (2000) On the effect of adding ε-Bernoulli percolation to everywhere percolating subgraphs of $\mathbb{Z}^d$, *J. Math. Phys.* **41**, 1294–1297.

[7] Benjamini, I., Jonasson, J., Schramm, O. and Tykesson, J. (2009) Visibility to infinity in the hyperbolic plane despite obstacles, *ALEA* **6**, 323–342.

[8] Benjamini, I., Kesten, H., Peres, Y. and Schramm, O. (2004) Geometry of the uniform spanning forest: transitions in dimensions $4, 8, 12, \ldots$, *Ann. Math.* **160**, 465–491.

[9] Benjamini, I., Lyons, R., Peres, Y. and Schramm, O. (1999) Group-invariant percolation on graphs, *Geom. Funct. Anal.* **9**, 29–66.

[10] Benjamini, I., Lyons, R., Peres, Y. and Schramm, O. (1999) Critical percolation on any nonamenable group has no infinite clusters, *Ann. Probab.* **27**, 1347–1356.

[11] Benjamini, I., Lyons, R., Peres, Y. and Schramm, O. (2001) Uniform spanning forests, *Ann. Probab.* **29**, 1–65.

[12] Benjamini, I., Lyons, R. and Schramm, O. (1999) Percolation perturbations in potential theory and random walks, in *Random walks and discrete potential theory* (M. Picardello and W. Woess, eds), pp 56–84, Cambridge University Press.

[13] Benjamini, I. and Schramm, O. (1996) Percolation beyond $\mathbb{Z}^d$, many questions and a few answers, *Electr. Comm. Probab.* **1**, 71–82.

[14] Benjamini, I. and Schramm, O. (1998) Exceptional planes of percolation, *Probab. Theory Related Fields* **111**, 551–564.

[15] Benjamini, I. and Schramm, O. (1999) Recent progress on percolation beyond $\mathbb{Z}^d$, `http://research.microsoft.com/en-us/um/people/schramm/pyondrep/`

[16] Benjamini, I. and Schramm, O. (2001) Percolation in the hyperbolic plane, *J. Amer. Math. Soc.* **14**, 487–507.

[17] Berger, N. and Biskup, M. (2007) Quenched invariance principle for simple random walk on percolation clusters, *Probab. Theory Related Fields* **137**, 83–120.

[18] Bollobás, B. and Riordan, O. (2006) The critical probability for random Voronoi percolation in the plane is 1/2, *Probab. Theory Related Fields* **136**, 417–468.

[19] Bollobás, B. and Riordan, O. (2006) *Percolation*, Cambridge University Press.

[20] Broder, A. (1989) Genertaing random spanning trees, in *30th Annual Symposium on the Foundations of Computer Science*, pp 442–447, IEEE, New York.

[21] Burton, R.M. and Keane, M.S. (1989) Density and uniqueness of percolation, *Comm. Math. Phys.* **121**, 501–505.

[22] Burton, R.M. and Keane, M.S. (1991) Topological and metric properties of infinite clusters in stationary two-dimensional site percolation, *Israel J. Math.* **76**, 299–316.

[23] De Masi, A., Ferrari, P.A., Goldstein, S. and Wick, W.D. (1989) An invariance principle for reversible Markov processes. Applications to random motions in random environments, *J. Statist. Phys.* **55**, 787–855.

[24] Doyle, P.G. and Snell, J.L. (1984) *Random Walks and Electric Networks*, Mathematical Association of America, Washington, D.C.

[25] Gandolfi, A., Keane, M.S. and Newman, C.M. (1992) Uniqueness of the infinite component in a random graph with applications to percolation and spin glasses, *Probab. Theory Related Fields* **92**, 511–527.

[26] Grimmett, G.R. (1999) *Percolation*, 2nd ed, Springer, New York.

[27] Grimmett, G.R. (2006) *The Random-Cluster Model*, Springer, New York.

[28] Grimmett, G.R., Kesten, H. and Zhang, Y. (1993) Random walk on the infinite cluster of the percolation model, *Probab. Theory Related Fields* **96**, 33–44.

[29] Grimmett, G.R. and Newman, C.M. (1990) Percolation in $\infty+1$ dimensions, in *Disorder in Physical Systems* (G.R. Grimmett and D.J.A. Welsh, eds), pp. 167–190, Oxford University Press.

[30] Häggström, O. (1995) Random-cluster measures and uniform spanning trees, *Stochastic Process. Appl.* **59**, 267–275.

[31] Häggström, O. (1997) Infinite clusters in dependent automorphism invariant percolation on trees, *Ann. Probab.* **25**, 1423–1436.

[32] Häggström, O. (2000) Markov random fields and percolation on general graphs, *Adv. Appl. Probab.* **32**, 39-66.

[33] Häggström, O. (2008) Oded Schramm 1961–2008, *Svenska Matematikersamfundets Medlemsutskick*, Oct 15, 27–29. `http://www.math.chalmers.se/~olleh/Oded.pdf`

[34] Häggström, O. and Jonasson, J. (2006) Uniqueness and non-uniqueness in percolation theory, *Probab. Surveys* **3**, 289-344.

[35] Häggström, O. and Peres, Y. (1999) Monotonicity of uniqueness for percolation on Cayley graphs: all infinite clusters are born simultaneously, *Probab. Theory Related Fields* **113**, 273–285.

[36] Hara, T. and Slade, G. (1994) Mean-field behaviour and the lace expansion, in *Probability and Phase Transition* (G.R. Grimmett, ed.), pp 87–122, Kluwer, Dordrecht.

[37] Harris, T.E. (1960) A lower bound for the critical probability in a certain percolation process, *Proc. Cambridge Philos. Soc.* **56**, 13–20.

[38] Harrison, E.R. (1987) *Darkness at Night: A Riddle of the Universe*, Harvard University Press.

[39] Kesten, H. (1959) Symmetric random walks on groups, *Trans. Amer. Math. Soc.* **92**, 336–354.

[40] Kesten, H. (1959) Full Banach mean values on countable groups, *Math. Scand.* **7**, 146–156.

[41] Kesten, H. (1980) The critical probability of bond percolation on the square lattice equals $\frac{1}{2}$, *Comm. Math. Phys.* **74**, 41–59.

[42] Kirchhoff, G. (1847) Ueber die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Vertheilung galvanischer Ströme geführt wird, *Ann. Phys. Chem.* **72**, 497–508.

[43] Lalley, S.P. (1998) Percolation on Fuchsian groups, *Ann. Inst. H. Poincaré Probab. Statist.* **34**, 151–177.

[44] Lalley, S.P. (2001) Percolation clusters in hyperbolic tessellations, *Geom. Funct. Anal.* **11**, 971–1030.

[45] Lawler, G.F. (1991) *Intersections of Random Walks*, Birkhäuser, Boston, MA,

[46] Lawler, G.F., Schramm, O. and Werner, W. (2001) Values of Brownian intersection exponents. I. Half-plane exponents, *Acta Math.* **187**, 237–273.

[47] Lawler, G.F., Schramm, O. and Werner, W. (2001) Values of Brownian intersection exponents. II. Plane exponents, *Acta Math.* **187**, 275–308.

[48] Lawler, G.F., Schramm, O. and Werner, W. (2002) Values of Brownian intersection exponents. III. Two-sided exponents, *Ann. Inst. H. Poincaré Probab. Statist.* **38**, 109–123.

[49] Lawler, G.F., Schramm, O. and Werner, W. (2002) Analyticity of intersection exponents for planar Brownian motion, *Acta Math.* **189**, 179–201.

[50] Lyons, R. (1990) Random walks and percolation on trees, *Ann. Probab.* **18**, 931–958.

[51] Lyons, R. (1992) Random walks, capacity, and percolation on trees, *Ann. Probab.* **20**, 2043–2088.

[52] Lyons, R. (1998) A bird's-eye view of uniform spanning trees and forests, in *Microsurveys in discrete probability (Princeton, NJ, 1997)*, pp 135–162, American Mathematical Society, Providence, RI.

[53] Lyons, R. (2000) Phase transitions on nonamenable graphs, *J. Math. Phys.* **41**, 1099–1126.

[54] Lyons, R. (2008) Oded Schramm 1961–2008, *IMS Bulletin* **37**, no. 9, 10–11.

[55] Lyons, R., Peres, Y. and Schramm, O. (2006) Minimal spanning forests, *Ann. Probab.* **34**, 1665–1692.

[56] Lyons, R. and Schramm, O. (1999) Indistinguishability of percolation clusters, *Ann. Probab.* **27**, 1809–1836.

[57] Lyons, R. and Schramm, O. (1999) Stationary measures for random walks in a random environment with random scenery, *New York J. Math.* **5**, 107-113.

[58] Meester, R. and Roy, R. (1996) *Continuum Percolation*, Cambridge University Press.

[59] *Memories of Oded Schramm*, blog, `http://odedschramm.wordpress.com/`

[60] Morris, B. (2003) The components of the wired spanning forest are recurrent, *Probab. Theory Related Fields* **125**, 259–265.

[61] Newman, C.M. and Schulman, L.S. (1981) Infinite clusters in percolation models, *J. Statist. Phys.* **26**, 613–628.

[62] Newman, C.M. and Stein, D.L. (1996) Ground-state structure in a highly disordered spin-glass model, *J. Statist. Phys.* **82**, 1113–1132.

[63] Pak, I. and Smirnova-Nagnibeda, T. (2000) On non-uniqueness of percolation on nonamenable Cayley graphs, *C. R. Acad. Sci. Paris Sr. I Math.* **330**, 495–500.

[64] Pemantle, R. (1991) Choosing a spanning tree for the integer lattice uniformly, *Ann. Probab.* **19**, 1559–1574.

[65] Peres, Y. (2000) Percolation on nonamenable products at the uniqueness threshold, *Ann. Inst. H. Poincaré Probab. Statist.* **36**, 395–406.

[66] Peres, Y., Pete, G. and Scolnicov, A. (2006) Critical percolation on certain nonunimodular graphs, *New York J. Math.* **12**, 1–18.

[67] Russo, L. (1981) On the critical percolation probabilities, *Z. Wahrsch. Verw. Gebiete* **56**, 229–237.

[68] Schonmann, R.H. (1999) Stability of infinite clusters in supercritical percolation, *Probab. Theory Related Fields* **113**, 287–300.

[69] Schonmann, R.H. (1999) Percolation in $\infty + 1$ dimensions at the uniqueness threshold. in *Perplexing Problems in Probability* (M. Bramson and R. Durrett, eds), pp 53–67, Birkhuser, Boston.

[70] Schramm, O. (2000) Scaling limits of loop-erased random walks and uniform spanning trees, *Israel J. Math.* **118**, 221–288.

[71] Schramm, O. (2001) A percolation formula, *Electr. Comm. Probab.* **6**, 115–120.

[72] Schramm, O. (2008) Hyperfinite graph limits, *Electr. Res. Announc. Math. Sci.* **15**, 17–23.

[73] Schramm, O. and Steif, J. (2007) Quantitative noise sensitivity and exceptional times for percolation, *Ann. Math.*, to appear.

[74] Soardi, P.M. (1994) *Potential Theory on Infinite Networks*, Springer, Berlin.

[75] Thomassen, C. (1989) Transient random walks, harmonic functions, and electrical currents in infinite electrical networks, *Mat-Report 1989-07*, Technical University of Denmark.

[76] Timár, Á. (2006) Percolation on nonunimodular transitive graphs, *Ann. Probab.* **34**, 2344–2364.

[77] Trofimov, V.I. (1985) Groups of automorphisms of graphs as topological groups, *Math. Notes* **38**, 717–720.

[78] Tykesson, J. (2007) The number of unbounded components in the Poisson Boolean model of continuum percolation in hyperbolic space, *Electr. J. Probab.* **12**, 1379–1401.

[79] Werner, W. (2009) Oded Schramm 1961–2008, *EMS Newsletter*, March issue, 9–13.

[80] Wilson, D.B. (1996) Generating random spanning trees more quickly than the cover time, in *Proceedings of the Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996)*, pp 296–303, ACM, New York.

[81] Woess, W. (2000) *Random Walks on Infinite Graphs and Groups*, Cambridge University Press.

Olle Häggström
Mathematical Sciences
Chalmers University of Technology
412 96 Göteborg
Sweden
olleh@chalmers.se
http://www.math.chalmers.se/ olleh/

Elect. Comm. in Probab. **1** (1996) 71–82

ELECTRONIC COMMUNICATIONS in PROBABILITY

PERCOLATION BEYOND $\mathbb{Z}^d$, MANY QUESTIONS AND A FEW ANSWERS

ITAI BENJAMINI & ODED SCHRAMM
The Weizmann Institute, Mathematics Department, Rehovot 76100, ISRAEL
e-mail: itai@wisdom.weizmann.ac.il & schramm@wisdom.weizmann.ac.il

AMS 1991 Subject classification: 60K35, 82B43
Keywords and phrases: Percolation, criticality, planar graph, transitive graph, isoperimeteric inequality.

Abstract

A comprehensive study of percolation in a more general context than the usual $\mathbb{Z}^d$ setting is proposed, with particular focus on Cayley graphs, almost transitive graphs, and planar graphs. Results concerning uniqueness of infinite clusters and inequalities for the critical value p_c are given, and a simple planar example exhibiting uniqueness and non-uniqueness for different $p > p_c$ is analyzed. Numerous varied conjectures and problems are proposed, with the hope of setting goals for future research in percolation theory.

1 Introduction

Percolation has been mostly studied in the lattices $\mathbb{Z}^d$, or in $\mathbb{R}^d$. Recently, several researchers have looked beyond this setting. For instance, Lyons (1996) gives an overview of current knowledge about percolation on trees, while Grimmett and Newman (1990) study percolation on (regular tree) $\times \mathbb{Z}$.

The starting point for a study of percolation on the Euclidean lattices is the fact that the critical probability for percolation, denoted p_c, is smaller than one. (See below for exact definitions and see Grimmett (1989) for background on percolation). The first step in a study of percolation on other graphs, for instance Cayley graphs of finitely generated groups, will be to prove that the critical probability for percolation on these graphs is smaller than one. In this note, we will show that for a large family of graphs, indeed, $p_c < 1$. In particular, this holds for graphs satisfying a strong isoperimetric inequality (positive Cheeger constant).

The second part of the paper discusses non-uniqueness of the infinite open cluster. A criterion is given for non-uniqueness, which is proved by dominating the percolation cluster by a branching random walk. This criterion is useful for proving non-uniqueness in "large" graphs.

Numerous open problems and conjectures are presented, probably of variable difficulty. The main questions are about the relation between geometric or topological properties of the graph, on the one hand, and the value of p_c, uniqueness and structure of the infinite cluster, on the other. It seems that there are many interesting features of percolation on planar graphs. One proposed conjecture is that a planar graph which has infinite clusters for $p = 1/2$ site

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_21, © Springer Science+Business Media, LLC 2011

percolation has infinitely many such clusters. This is proved for graphs that are locally finite in $\mathbb{R}^2$ and are disjoint from the positive x axis.
In a forthcoming paper, we shall discuss a *Voronoi percolation* model. A principle advantage of this model is its generality, which allows to extend various percolation questions beyond $\mathbb{R}^d$, to arbitrary Riemannian manifolds. In this model, the cells of a Voronoi tiling generated by a point Poisson process are taken to be open with probability p, independently. It turns out that this model is advantageous for the study of the conformal invariance conjecture for critical percolation, introduced in Langlands et al (1994). See Benjamini-Schramm (1996). Additionally, we currently study Voronoi percolation in the hyperbolic plane.
We wish to express thanks to Uriel Feige, Rick Kenyon, Ronald Meester, Yuval Peres, and Benjamin Weiss for helpful conversations, and the anonymous referee, for a very careful review.

2 Notations and Definitions

The graphs we shall consider will always be locally finite, that is, each vertex has finitely many neighbors.

Cayley Graphs, Given a finite set of generators $S = \langle g_1^{\pm 1}, \ldots, g_n^{\pm 1} \rangle$ for a group Γ, the *Cayley graph* is the graph $G(\Gamma) = (V, E)$ with $V = \Gamma$ and $\{g, h\} \in E$ iff $g^{-1}h$ is a generator. $G(\Gamma)$ depends on the set of generators. Note that any two Cayley Graphs of the same group are roughly isometric (quasi-isometric). (See Magnus et al (1976) or Ghys et al (1991)).

Almost Transitive Graphs, A Graph G is *transitive* iff for any two vertices u, v in G, there is an automorphism of G mapping u onto v. In particular, Cayley graphs are transitive graphs. $G = (V, E)$ is *almost transitive*, if there is a finite set of vertices $V_0 \subset V$ such that any $v \in V$ is taken into V_0 by some automorphism of G.
For example, the lift to $\mathbb{R}^2$ of any finite graph drawn on the torus $\mathbb{R}^2/\mathbb{Z}^2$ is almost transitive.

Percolation, We assume throughout that the graph G is connected. In Bernoulli site percolation, the vertices are open (respectively closed) with probability p (respectively $1 - p$) independently. The corresponding product measure on the configurations of vertices is denoted by $\mathbb{P}_p$. Let $C(v)$ be the (open) cluster of v. In other words, $C(v)$ is the connected component of the set of open vertices in G containing v, if v is open, and $C(v) = \emptyset$, otherwise. We write

$$\theta^v(p) = \theta_G^v(p) = \mathbb{P}_p\big\{C(v) \text{ is infinite}\big\}.$$

When G is transitive, we may write $\theta(p)$ for $\theta^v(p)$. If $C(v)$ is infinite, for some v, we say that *percolation* occurs. Clearly, if $\theta^v(p) > 0$, then $\theta^u(p) > 0$, for any vertices v, u. Let

$$p_c = \sup\big\{p : \theta^v(p) = 0\big\}$$

be the critical probability for percolation. See Grimmett (1989).
Throughout the paper, site percolation is discussed. Except when dealing with planar graphs, the results and questions remain equally valid, with only minor modifications, for bond percolation.

Uniqueness of the infinite open cluster, In $\mathbb{Z}^d$, when $\theta(p) > 0$, there exists with probability one a unique infinite open cluster. See Grimmett (1989), the charming proof in Burton and Keane (1989), and Meester's (1994) survey article. As was shown by Grimmett and Newman (1990), this is not the case for percolation on (some regular tree) $\times \mathbb{Z}$. For some values of p, uniqueness holds and for others, there are infinitely many infinite disjoint open clusters.

Therefore, it is natural to define

$$p_u = \inf\Big\{p : \mathbb{P}_p\{\text{there is exactly one infinite open cluster}\} = 1\Big\}.$$

Cheeger's constant, Let G be a graph. The Cheeger constant, $h(G)$, of G is

$$h(G) = \inf_S \frac{|\partial S|}{|S|},$$

where S is a finite nonempty set of vertices in G, and ∂S, the *boundary* of S, consists of all vertices in $V \setminus S$ that have a neighbor in S.

3 The percolation critical probability p_c

The starting point of the study of percolation on groups is the following "obvious" conjecture.

Conjecture 1 *If G is the Cayley graph of an infinite (finitely generated) group, which is not a finite extension of $\mathbb{Z}$, then $p_c(G) < 1$.*

Lyons (1994) covers the case of groups with Cayley graphs of exponential volume growth. Babson and Benjamini (1995) covers finitely presented groups with one end. Note that if a group Γ contains as a subgroup another group Γ' for which $p_c(G(\Gamma')) < 1$ then $p_c(G(\Gamma)) < 1$. By results mentioned below, if a group admits a quotient with p_c smaller than 1 for its Cayley graphs, then the same is true for Cayley graphs of the group itself. Thus, the conjecture holds for any group that has a $\mathbb{Z}^2$ quotient.
Suppose that Γ is a group of automorphisms of a graph G. The *quotient graph* G/Γ is the graph whose vertices, $V(G/\Gamma)$, are the equivalence classes $V(G)/\Gamma = \big\{\Gamma v : v \in V(G)\big\}$, or Γ-orbits, and an edge $\{\Gamma u, \Gamma v\}$ appears in G/Γ if there are representatives $u_0 \in \Gamma u, v_0 \in \Gamma v$ that are neighbors in G, $\{u_0, v_0\} \in E(G)$. The map $v \to \Gamma v$ from $V(G)$ to $V(G/\Gamma)$ is called the *quotient map.* If every $\gamma \in \Gamma$, except the identity, has no fixed point in $V(G)$, then G is also called a *covering graph* of G/Γ, and the quotient map is also called a *covering map.*
The following theorem follows from Campanino-Russo (1985). We bring an easy proof here, since it introduces a method which will also be useful below. The proof is reminiscent of the coupling argument of Grimmett and Wierman, which was used by Wierman (1989) in the study of AB-percolation.

Theorem 1 *Assume that G_2 is a quotient graph of G_1, $G_2 = G_1/\Gamma$. Let $v' \in G_1$, and let v be the projection of v' to G_2. Then for any $p \in [0,1]$,*

$$\theta^{v'}_{G_1}(p) \geq \theta^{v}_{G_2}(p),$$

and consequently,

$$p_c(G_1) \leq p_c(G_2).$$

Proof:
We will construct a coupling between percolation on G_2 and on G_1. Consider the following inductive procedure for constructing the percolation cluster of $v \in V(G_2)$. If v is closed, set $C^2_n = \emptyset$ for each n. Otherwise, set $C^2_1 = \{v\}$ and $W^2_1 = \emptyset$. Now let $n \geq 2$. If ∂C^2_{n-1} is

contained in W^2_{n-1}, set $C^2_n = C^2_{n-1}$, $W^2_n = W^2_{n-1}$. Otherwise, choose a vertex $w_n \in V(G_2)$ which is not in $C^2_{n-1} \cup W^2_{n-1}$, but is adjacent to a vertex $z_n \in C^2_{n-1}$. If w_n is open, then let $C^2_n = C^2_{n-1} \cup \{w\}$ and $W^2_n = W^2_{n-1}$, if closed, let $C^2_n = C_{n-1}$ and $W^2_n = W^2_{n-1} \cup \{w_n\}$. $C^2 = \bigcup_n C^2_n$ is $C(v)$ the percolation cluster of v.
We will now describe the coupling with the percolation process in G_1. Let f be the quotient map from G_1 to G_2. Recall that $f(v') = v$. If v is closed, let v' be closed. Otherwise, let v' be open, and let $C^1_1 = \{v'\}$, $W^1_1 = \emptyset$. Assume that $n \geq 2$ and C^1_{n-1}, W^1_{n-1} were defined, and satisfy $f(C^1_{n-1}) = C^2_{n-1}$, $f(W^1_{n-1}) = W^2_{n-1}$. If the construction of $C(v)$ in G_2 stopped at stage n, that is, if $C^2_n = C^2_{n-1}$, $W^2_n = W^2_{n-1}$, then let $C^1_n = C^1_{n-1}$, $W^1_n = W^1_{n-1}$. Otherwise, let z'_n be some vertex in $f^{-1}(z_n) \cap C^1_{n-1}$, and let w'_n be some vertex in $f^{-1}(w_n)$ that neighbors with z'_n. Let w'_n be open iff w_n is open, and define C^1_n and W^1_n accordingly. Then $\bigcup_n C^1_n$ is a connected set of open vertices contained in the percolation cluster $C(v')$ of v'. Hence, $f\big(C(v')\big) \supset C(V)$, and the theorem follows. ■

Question 1 *When does strict inequality hold? We believe that if both G_1 and G_2 are connected almost transitive graphs, G_1 covers but is not isomorphic to G_2 and $p_c(G_2) < 1$, then $p_c(G_1) < p_c(G_2)$.*

Compare with Men'shikov (1987) and Aizenman-Grimmett (1991). Other conjectures regarding p_c for graphs:

Conjecture 2 *Assume that G is an almost transitive graph with ball volume growth faster then linear. Then $p_c(G) < 1$.*

Let G be a graph, define the isoperimetric dimension,

$$\mathrm{Dim}(G) = \sup\left\{ d > 0 : \inf_S \frac{|\partial S|}{|S|^{\frac{d-1}{d}}} > 0 \right\},$$

where S is a finite nonempty set of vertices in G.

Question 2 *Does $Dim(G) > 1$ imply $p_c(G) < 1$?*

Conjecture 3 *Assume that G is a (bounded degree) triangulation of a disc. Any of the following list of progressively weaker assumptions should be sufficient to guarantee $p_c(G) \leq 1/2$:*

(1) $Dim(G) \geq 2$,

(2) $Dim(G) > 1$,

(3) for any finite set A of vertices in G the inequality $|\partial A| \geq f\big(|A|\big) \log |A|$ holds, where f is some function satisfying $\lim_{n\to\infty} f(n) = \infty$.

Moreover, if $h(G) > 0$, then $p_c(G) < 1/2$. The latter might be easier to establish under the assumption of non-positive curvature, that is, minimal degree ≥ 6.

The statement that (1) is sufficient to guarantee $p_c \leq 1/2$ would generalize the fact that $p_c = 1/2$ for the triangular lattice. See Wierman (1989).
So far, we can only show that $p_c < 1$ for graphs with positive Cheeger constant.

Theorem 2

$$p_c(G) \leq \frac{1}{h(G)+1}.$$

Remark 1 *For a degree k regular tree, $p_c(T_k) = (k-1)^{-1} = \big(h(T_k)+1\big)^{-1}$, because $h(T_k) = k-2$. So the theorem is sharp.*

Proof:
Let C_n and W_n be defined as C_n^2, W_n^2 were in the proof of Theorem 1, and let $C = \bigcup_n C_n$. If C is finite (and nonempty), then there is some smallest N such that the boundary of C_N is W_N. By the definition of the Cheeger constant $|W_N| = |\partial C_N| \geq h(G)|C_N|$. That is, we flipped N independent $(p, 1-p)$-coins and $|W_N| \geq Nh(G)/\big(h(G)+1\big)$ turned out closed. But if $p > 1 - h(G)/\big(h(G)+1\big)$, then, with positive probability, a random infinite sequence of independent Bernoulli$(p, 1-p)$ variables does not have an N such that at least $Nh(G)\big(h(G)+1\big)^{-1}$ zeroes appear among the first N elements. In particular, with positive probability we have percolation. ■

Remark 2 *Suppose that $G = (V, E)$ is a finite graph, which is an α-expander; that is, $|\partial A| \geq \alpha|A|$ for any $A \subset V$ with $|A| < |V|/2$. Then the above proof shows that with probability bounded away from zero, the percolation process on G with $p > 1 - \alpha/(\alpha+1)$ will have a cluster with at least half of the vertices of G.*

It has been one of the outstanding challenges of percolation theory to prove that critical percolation in $\mathbb{Z}^d$ $(d > 2)$ dies at p_c, that is, $\theta(p_c) = 0$. This has been proved for $d = 2$ by Kesten (1980) and Russo (1981), and for sufficiently high d by Hara and Slade (1989). It might be beneficial to study the problem in other settings.

Conjecture 4 *Critical percolation dies in every almost transitive graph (assuming $p_c < 1$).*

It is not hard to construct a tree where critical percolation lives, compare Lyons (1996). This shows that the assumption of almost transitivity is essential.

4 Uniqueness and non-uniqueness in almost transitive graphs

In this section, the number and structure of the infinite clusters is discussed, and conditions that guarantee $p_c < p_u$ or $p_u < 1$, are given.

Definition 1 *Let G be a graph. An end of G is a map e that assigns to every finite set of vertices $K \subset V(G)$ a connected component $e(K)$ of $G \setminus K$, and satisfies the consistency condition $e(K) \subset e(K')$ whenever $K' \subset K$.*

It is not hard to see that if $K \subset V(G)$ is finite and F is an infinite component of $G \setminus K$, then there is an end of G satisfying $e(K) = F$.

Definition 2 *The percolation subgraph of a graph G is the graph spanned by all open vertices.*

By Kolmogorov's 0-1 law, the probability of having an infinite component in the percolation subgraph is either 0 or 1. Following is an elementary extension of this 0-1 law in the setting of almost transitive graphs.

Theorem 3 *Let G be an almost transitive infinite graph, and consider a percolation process for some fixed $p \in (0,1)$. Then precisely one of the following situations occurs.*

(1) With probability 1, every component of the percolation graph is finite.

(2) With probability 1, the percolation graph has exactly one infinite component, and has exactly one end.

(3) With probability 1, the percolation graph has infinitely many infinite components, and for every finite n there is some infinite component with more than n ends.

The proof is virtually identical to the proof of a similar theorem for dependent percolation in $\mathbb{Z}^d$, by Newman-Schulman (1981); it is therefore omitted.

Conjecture 5 *Let G be a connected almost transitive graph, and fix a $p \in (0,1)$. Suppose that a.s. there is more than one infinite component in the percolation subgraph. Then, with probability 1, each infinite component in the percolation subgraph has precisely $2^{\aleph_0}$ ends.*

The conjecture is equivalent to saying that in the case of non-uniqueness, there cannot be an infinite component with exactly one end. To demonstrate this, recall that the collection of sets of the form $\{e : e(K) = F\}$ is a sub-basis for a topology on the set of ends of a connected graph, and the set of ends with this topology is a compact (totally disconnected) Hausdorff space. A compact Hausdorff space with no isolated points has cardinality $\geq 2^{\aleph_0}$. Hence, the conjecture follows if one shows that there are no isolated ends in the percolation subgraph. It is easy to see, using the argument of Newman-Schulman (1981), that if there are isolated ends, then there are also infinite clusters with just one end.
Häggström (1996) studies similar questions in the setting of dependent percolation on trees.

It is easy to verify that the proof by Burton and Keane (1989) of the uniqueness of the infinite open cluster for $\mathbb{Z}^d$ works as well for almost transitive graphs with Cheeger constant zero.

Conjecture 6 *Assume that the almost transitive graph G has positive Cheeger constant. Then $p_c(G) < p_u(G)$.*

The conjecture, if true, gives a percolation characterization of amenability. We now present a comparison between the connectivity function and hitting probabilities for a branching random walk on G. This will be useful in showing that $p_c < p_u$, for some graphs. For background on branching random walks on graphs, see Benjamini and Peres (1994).
Let G be a connected graph, and let $p^n(v,u)$ denote the n-step transition probability between v and u, for the simple random walk on G.

$$\rho(G) = \limsup_{n\to\infty} \left(p^n(v,u)\right)^{1/n}$$

is the *spectral radius* of G, and does not depend on v and u. By Dodziuk (1984), for a bounded degree graph, $\rho(G) < 1$ iff the Cheeger constant of G is positive. We have the following

Theorem 4 *Let G be an almost transitive graph, with maximal degree k. Let p be such that percolation occurs at p, that is, $\theta^v(p) > 0$ for some v. If, additionally,*

$$\rho(G)kp < 1,$$

then $\mathbb{P}_p$-almost surely there are infinitely many infinite open clusters. In particular, if $\rho(G)kp_c < 1$, then $p_c < p_u$.

Proof:
Given $p \in [0,1]$, consider the following branching random walk (BRW) on G. Start with a particle at v at time 0. At time 1, at every neighbor of v, a particle is born with probability p, and the particle at v is deleted. Continue inductively: if at time n we have some particles located on G, then at time $n+1$ each one of them gives birth to a particle on each of its neighbors with probability p, independently from the other neighbors, and then dies. We claim that

$$\mathbb{P}_p\{u \in C(v)\} \leq \mathbb{P}\{\text{the BRW starting at } v \text{ hits } u\}.$$

Say that $u \in G$ is in the *support* of the BRW if u is visited by a particle at some time. We will show inductively that the support of the BRW dominates $C(v)$. Consider the following inductive procedure for a coupling of the BRW with the percolation process. Let $C_0 = v$ and $W_0 = \emptyset$. For each $n \geq 1$, choose a vertex w, which is not in $C_{n-1} \cup W_{n-1}$, but is adjacent to a vertex $z \in C_{n-1}$. If, at least once, a particle located at z gave birth to a particle at w, then let $C_n = C_{n-1} \cup w$ and $W_n = W_{n-1}$. Otherwise let $C_n = C_{n-1}$ and $W_n = W_{n-1} \cup \{w\}$. It follows that at each step, the new vertex w is added to C_{n-1} with probability $\geq p$. If at some point there is no vertex w as required, the process stops and we have generated a cluster, which we denote by C. If the process continues indefinitely, we set $C = \bigcup_n C_n$. Note that C is contained in the support of the BRW. Now view the process from a different perspective. Consider C_n to be the set of open vertices and W_n to be the set of closed vertices in the percolation model. Each vertex in $\bigcup_n (W_n \cup C_n)$ is tested only once, and added to C with probability $\geq p$. Thus, the cluster C dominates the open percolation cluster containing v.
The population size of the BRW is dominated by the population size of a Galton-Watson branching process with binomial(p,k) offspring distribution. The mean number of offsprings for that branching process is pk. Hence, by Borel-Cantelli, when $pk < \rho(G)^{-1}$, the BRW is transient, that is, almost surely only finitely many particles will visit v. (Compare Benjamini and Peres (1994)).
Now suppose that p satisfies $\theta^v(p) > 0$ and $\rho(G)pk < 1$. We claim that the transience of the corresponding BRW implies that the probability that two vertices x, y are in the same percolation cluster goes to zero as the distance from x to y tends to infinity. Indeed, suppose that there is an $\epsilon > 0$ and a sequence of vertices x_n, y_n with the distance $d(x_n, y_n)$ tending to infinity, but the probability that they are in the same cluster is greater than ϵ. That would mean that the BRW starting at x_n has probability at least ϵ to reach y_n and the BRW starting at y_n has probability at least ϵ to reach x_n. Consequently, with probability at least ϵ^2, the BRW starting at x_n will reach x_n again at some time after $d(x_n, y_n)$ steps. Since G is almost transitive, by passing to a subsequence and applying an automorphism of G, we may assume that all x_n are the same. This contradicts the transience of the BRW.
Let $r > 0$, and consider m balls in the graph with radius r. If r is large, then with probability arbitrarily close to 1 each of these balls will intersect an infinite open cluster. But the probability that any cluster will intersect two such balls goes to zero as the distances between the balls goes to infinity. Hence, with probability 1, there are more than m infinite open clusters. The theorem follows. ■

Suppose that G is not almost transitive, but has bounded degree. Then the above argument can be modified to show that for p as above, the probability of having at least m infinite open clusters is at least $\left(\inf_v \theta^v(p)\right)^m$.
Grimmett and Newman (1990) showed that $\mathbb{Z} \times$ (some regular tree) satisfies $p_c < p_u < 1$. We now show that $p_c < p_u$ for many products.

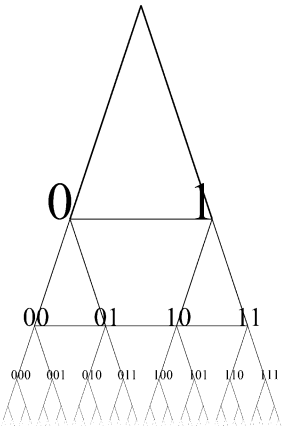


Figure 1: The enhanced binary tree.

Corollary 1 *Let G be an almost transitive graph. Then there is a $k_0 = k_0(G)$ such that the product $G \times T_k$ of G with the k-regular tree satisfies $p_c < p_u$ whenever $k \geq k_0$.*

Proof:
Let m be the maximal degree in G. Note that $p_c(G \times T_k) \leq p_c(T_k) = (k-1)^{-1}$, and the maximal degree in $G \times T_k$ is $m + k$. Observe that $\rho(G \times T_k) \to 0$ as $k \to \infty$. Hence, the corollary follows from the theorem. ■

If one wishes to drop the assumption that G is almost transitive (but still has bounded degree), then the above arguments show that for k sufficiently large, there is a p such that the probability of having more than one infinite open cluster in $G \times T_k$ is positive.

Following is a simple example of a planar graph satisfying $p_c < p_u < 1$. The planarity will make the analysis easy.

Example 1 *Consider the graph obtained by adding to the binary tree edges connecting all vertices of same level along a line (see Figure 1). To be more precise, represent the vertices of the binary tree by sequences of zeros and ones in the usual way, and add to the binary tree an edge between v, w if v and w are at level n and $|0.v - 0.w| = 1/2^n$, where $0.v$ is the number in $[0, 1]$ represented by the sequence corresponding to v. Note that this graph is roughly isometric to a sector in the hyperbolic plane.*

Proposition 1 *For this graph, $p_c < 1 - p_c \leq p_u(G) < 1$, and for p in the range $p \in (p_c, 1 - p_c)$ there are, with probability 1, infinitely many infinite open clusters.*

Proof:
First note that G contains the graph G_0, obtained by adding to the binary tree edges only between vertices that neighbor in G and have the same grandmother. G_0 is a "periodic refinement" of the binary tree. By comparing the number of vertices in G_0 that are the the cluster of the root to a Galton-Watson branching process, it is easy to see that $p_c(G) \leq$

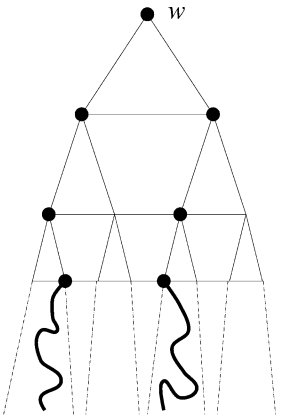


Figure 2: The infinite black clusters separate.

$p_c(G_0) < p_c(T_3) = 1/2$. Let $p \in (p_c, 1-p_c)$. For such p, both the open and the closed clusters percolate, and this is so in any subgraph of G spanned by the binary tree below any fixed vertex, as it is isomorphic to the whole graph. Pick some finite binary word w. Suppose that the vertices $w, w0, w00, w1, w10$ are closed, and each of $w00, w10$ percolates (in closed vertices) in the subgraph below it (see Figure 2). This implies that the open clusters intersecting the subgraph below $w01$ will be disjoint from those below $w11$, which gives non-uniqueness, because each of these subgraphs is sure to contain infinite open clusters. With probability 1, there is some such w. Hence, for $p \in (p_c, 1-p_c)$, there is no uniqueness, and $p_u \geq 1-p_c$. It is easy to see that for $p \in (p_c, 1-p_c)$ there are, with probability 1, infinitely many infinite open clusters. The proof is completed by the following lemma, which gives $p_u < 1$. ■

Lemma 1 *Let G be a bounded degree triangulation of a disk, or, more generally, the 1-skeleton of a (locally finite) tiling of a disk, where the number of vertices surrounding a tile is bounded. Then, for p sufficiently close to 1, there is a.s. at most 1 infinite open cluster.*

Proof:
Suppose that each tile is surrounded by at most k edges. Let G' be the k'th power of G; that is, $V(G') = V(G)$, and an edge appears in G' if the distance in G between its endpoints is at most k. Then G' has bounded degree, and therefore, for some $p^* > 0$ close to zero, there is no percolation in G'. Hence, $\mathbb{P}_{p^*}$ a.s., given any $n > 0$, there is a closed set of vertices in G' that separate a fixed basepoint from 'infinity', and all have distance at least n from the basepoint. If one now thinks of these as vertices in G, they contain the vertices of a loop separating the basepoint from infinity. The distance from this loop to the basepoint is arbitrarily large. Taking $p = 1 - p^*$, then shows that a.s. for $\mathbb{P}_p$ on G there are open loops separating the basepoint from infinity which are arbitrarily far away from the basepoint. Each infinite open cluster must intersect all but finitely many of these loops. Hence there is at most one infinite open cluster. ■

Question 3 *Give general conditions that guarantee $p_u < 1$. For example, is $p_u < 1$ for any transitive graph with one end?*

The proof of the proposition suggests the following conjecture and question.

Conjecture 7 *Suppose G is planar, and the minimal degree in G is at least 7. Then at every p in the range $(p_c, 1-p_c)$, there are infinitely many infinite open clusters. Moreover, we conjecture that $p_c < 1/2$, so the above interval is nonempty.*

Such a graph has positive Cheeger constant, and spectral radius less than 1.

Conjecture 8 *Let G be planar, set $p = 1/2$ and assume that a.s. percolation occurs. Then a.s. there are infinitely many infinite open clusters.*

It is not clear if there is any analogous statement in the setting of bond percolation.
There is a large class of graphs in which we can prove the conjecture.

Theorem 5 *Let G be a planar graph, which is disjoint from the positive x axis, $\{(x,0) : x \geq 0\}$. Suppose that every bounded set in the plane meets finitely many vertices and edges of G. Set $p = 1/2$, and assume that almost surely percolation occurs in G. Then, almost surely, there are infinitely many infinite open clusters.*

Proof:
Let X be the collection of all infinite open or closed clusters. Suppose that X is finite, and let $R > 0$ be sufficiently large so that the disk $x^2 + y^2 < R^2$ intersects each cluster in X. For any $r > R$, and $A \in X$, let $t(A,r)$ be the least $t \in [0, 2\pi]$ such that the point $(r\cos t, r \sin t)$ is on an edge connecting two vertices in A (or is a vertex of A). Let $A, B \in X$ be distinct. Suppose that $R < r_1 < r_2$, $t(A,r_1) < t(B,r_1)$ and $t(A,r_2) < t(B,r_2)$. Take some r in the range $r_1 < r < r_2$. If $t(B,r) < t(A,r)$, then it follows that B is contained in the domain bounded by the arcs $\{(x,0) : x \in [r_1, r_2]\}$, $\{(r_1 \cos t, r_1 \sin t) : t \in [0, t(A,r_1)]\}$, $\{(r_2 \cos t, r_2 \sin t) : t \in [0, t(A,r_2)]\}$ and by A. This is impossible, because B has infinitely many vertices, and therefore, $t(A,r) < t(B,r)$. Consequently, the inequality between $t(A,s)$ and $t(B,s)$ changes at most once in the interval $R < s < \infty$. So either $t(A,s) > t(B,s)$ for all s sufficiently large, or $t(A,s) < t(B,s)$ for all s sufficiently large. In the latter case, we write $A < B$. It is clear that this defines a linear order on X.
Because X is finite, it has a minimal element. Let E be the event that the minimal element in X is an open cluster. By symmetry, $\mathbb{P}\{E\} = 1/2$. But Kolmogorov's 0-1 law implies that $\mathbb{P}\{E\}$ is either 0 or 1. The contradiction implies that X is infinite. Consequently, a.s. there are infinitely many open clusters, or there are infinitely many closed cluster. Consequently, there are infinitely many open clusters, again by Kolmogorov's 0-1 law. ■

Other questions regarding p_u are

Question 4 *Let G, G' be two Cayley graphs of the same group (or, more generally, two roughly isometric almost transitive graphs). Does $p_c(G) < p_u(G)$ imply $p_c(G') < p_u(G')$?*

Question 5 *Assume that G is an almost transitive graph. Is there uniqueness for every $p > p_u$? Is there uniqueness at $p = p_u$?*

References

[1] M. Aizenman and G. R. Grimmett (1991): Strict monotonicity for critical points in percolation and ferromagnetic models. *J. Statist. Phys.* **63**, 817–835.

[2] E. Babson and I. Benjamini (1995): Cut sets in Cayley graphs, (preprint).

[3] I. Benjamini and Y. Peres (1994): Markov chains indexed by trees. *Ann. Prob.* **22**, 219–243.

[4] I. Benjamini and O. Schramm (1996): Conformal invariance of Voronoi percolation. (preprint).

[5] R. M. Burton and M. Keane (1989): Density and uniqueness in percolation. *Comm. Math. Phy.* **121**, 501–505.

[6] M. Campanino, L. Russo (1985): An upper bound on the critical percolation probability for the three-dimensional cubic lattice. *Ann. Probab.* **13**, 478–491.

[7] J. Dodziuk (1984): Difference equations, isoperimetric inequality and transience of certain random walks. *Trans. Amer. Math. Soc.* **284**, 787–794.

[8] E. Ghys, A. Haefliger and A. Verjovsky eds. (1991): Group theory from a geometrical viewpoint, World Scientific.

[9] G. R. Grimmett (1989): Percolation, Springer-Verlag, New York.

[10] G. R. Grimmett and C. M. Newman (1990): Percolation in $\infty+1$ dimensions, in *Disorder in physical systems*, (G. R. Grimmett and D. J. A. Welsh eds.), Clarendon Press, Oxford pp. 219–240.

[11] O. Häggström (1996): Infinite clusters in dependent automorphism invariant percolation on trees, (preprint).

[12] T. Hara and G. Slade (1989): The triangle condition for percolation, *Bull. Amer. Math. Soc.* **21**, 269–273.

[13] H. Kesten (1980): The critical probability of bond percolation on the square lattice equals 1/2, *Comm. Math. Phys.* **74**, 41–59.

[14] R. Langlands, P. Pouliot, and Y. Saint-Aubin (1994): Conformal invariance in two-dimensional percolation, *Bull. Amer. Math. Soc. (N.S.)* **30**, 1–61.

[15] R. Lyons (1995): Random walks and the growth of groups, *C. R. Acad. Sci. Paris,* **320**, 1361–1366.

[16] R. Lyons (1996): Probability and trees, (preprint).

[17] W. Magnus, A. Karrass and D. Solitar (1976): Combinatorial group theory, Dover, New York.

[18] R. Meester (1994): Uniqueness in percolation theory, *Statistica Neerlandica* **48**, 237–252.

[19] M. V. Men'shikov (1987): Quantitative estimates and strong inequalities for the critical points of a graph and its subgraph, (Russian) *Teor. Veroyatnost. i Primenen.* **32**, 599–602. Eng. transl. *Theory Probab. Appl.* **32**, (1987), 544–546.

[20] C. M. Newman and L. S. Schulman (1981): Infinite clusters in percolation models, *J. Stat. Phys.* **26**, 613–628.

[21] L. Russo (1981): On the critical percolation probabilities, *Z. Wahr.* **56**, 229–237.

[22] J. C. Wierman (1989): AB Percolation: a brief survey, in *Combinatorics and graph theory*, **25**, Banach Center Publications, pp. 241–251.

The Annals of Probability
1999, Vol. 27, No. 3, 1347–1356

CRITICAL PERCOLATION ON ANY NONAMENABLE GROUP HAS NO INFINITE CLUSTERS

BY ITAI BENJAMINI, RUSSELL LYONS,[1] YUVAL PERES[2] AND ODED SCHRAMM[3]

Weizmann Institute of Science, Indiana University, Hebrew University and University of California, Berkeley and Weizmann Institute of Science

We show that independent percolation on any Cayley graph of a nonamenable group has no infinite components at the critical parameter. This result was obtained by the present authors earlier as a corollary of a general study of group-invariant percolation. The goal here is to present a simpler self-contained proof that easily extends to quasi-transitive graphs with a unimodular automorphism group. The key tool is a "mass-transport" method, which is a technique of averaging in nonamenable settings.

1. Introduction. The main long-standing open question in percolation theory is to show that critical percolation in $\mathbb{Z}^d$ has no infinite components for all $d \geq 2$. The work of Harris (1960) and Kesten (1980) established the two-dimensional case; Hara and Slade (1994) proved it for $d \geq 19$. Recently, a study of percolation on other graphs, such as Cayley graphs, was initiated. The relevant definitions appear below. Here is our main theorem.

THEOREM 1.1. *Let X be a Cayley graph of a finitely generated nonamenable group, and consider either site or bond percolation on X. Then at the corresponding critical value $p = p_c$, almost surely there is no infinite component.*

In particular, there are no infinite components for critical percolation on any lattice in hyperbolic space, nor on a (k-regular tree) $\times \mathbb{Z}$. The latter was previously known when $k \geq 7$ [Wu (1993)]. Wu's proof goes along the lines of the high-dimensional Euclidean proof and uses the triangle condition.

The key tool in our proof is a technique of "mass transport." This tool is especially valuable in nonamenable settings, where it can sometimes replace the ergodic theorem for approximation of expectations by suitable spatial averages.

Received November 1997; revised August 1998.
[1]Supported in part by the Institute for Advanced Studies, Jerusalem.
[2]Supported in part by NSF Grant DMS-98-03597.
[3]Supported in part by the Sam and Ayala Zacks Professorial Chair.
AMS 1991 *subject classifications*. Primary 60B99; secondary 60D05, 82B43.
Key words and phrases. Percolation, Cayley graphs, amenability.

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_22, © Springer Science+Business Media, LLC 2011

Early forms of the mass-transport method were used by Adams (1990), van den Berg and Meester (1991) and Häggström (1997). The method was developed further in Benjamini, Lyons, Peres and Schramm (1999). The latter paper, denoted BLPS (1999) below, is a study of group-invariant, possibly dependent, percolation models on transitive graphs. After completing the first draft of that paper, we realized that combining two of the results there yields Theorem 1.1.

The resulting proof of Theorem 1.1 was not the most direct possible; in this note, we illustrate the technique of mass transport by presenting the simplest proof we know for Theorem 1.1. Another advantage of the proof given here is that it easily extends to the wider setup of quasi-transitive graphs with a unimodular nonamenable automorphism group. Nevertheless, the present proof still uses invariant percolation models.

In the study of percolation on nonamenable graphs, new phenomena, such as a double phase transition, appear; the first example of this is due to Grimmett and Newman (1990), who considered percolation on the product of a regular tree and $\mathbb{Z}$. Benjamini and Schramm (1996) present an introduction to the subject and a list of problems, some of which have been recently solved; see the final section of the present paper for an update. In that section, we also survey some further results and conjectures concerning percolation on nonamenable graphs.

Terminology. Let Γ be a finitely generated group. Given a finite symmetric set of generators $S = \{g_1^{\pm 1}, \dots, g_n^{\pm 1}\}$ for Γ, the (right) *Cayley graph* of Γ is the graph $X(\Gamma, S) := (\mathsf{V}, \mathsf{E})$ with $\mathsf{V} := \Gamma$ and $[g, h] \in \mathsf{E}$ iff $g^{-1}h \in S$. Thus, the graph $X(\Gamma, S)$ depends on the set of generators. Note that any two Cayley graphs of the same group are roughly isometric (quasi-isometric). More generally, a graph X is called *transitive* if for any vertices x, y, there is a graph automorphism mapping x to y. Clearly, any Cayley graph is transitive by the action of left multiplication.

Let the *edge-isoperimetric constant* of X be

$$\iota_{\mathsf{E}}(X) := \inf\left\{ \frac{|\partial_{\mathsf{E}} K|}{|K|};\ K \subset \mathsf{V} \text{ is finite} \right\},$$

where $\partial_{\mathsf{E}} K$, the edge boundary of K, consists of all edges $[u, v]$ with $u \in K$ and $v \in \mathsf{V} \setminus K$. A finitely generated group is *nonamenable* if for some finite generating set, the edge-isoperimetric constant of its Cayley graph is positive. The isoperimetric constant of a graph is sometimes called the Cheeger constant.

In p-Bernoulli bond percolation on a graph $X = (\mathsf{V}(X), \mathsf{E}(X))$, the edges are *open* with probability p independently. Those edges that are not open are *closed*. The corresponding product measure on the edge configurations is denoted $\mathbf{P}_p$. The *percolation subgraph* is the random graph whose vertices are $\mathsf{V}(X)$ and whose edges are the open edges. Let $K(v)$ be the *cluster* of v,

that is, the connected component of v in the percolation subgraph. We write

$$\theta_v(p) := \mathbf{P}_p[K(v) \text{ is infinite}].$$

On a Cayley graph (or, more generally, on a transitive graph), the value of $\theta_v(p)$ is independent of the choice of v, whence the subscript v is dropped. Let

$$p_c := \inf\{p; \theta(p) > 0\}$$

be the critical probability for percolation. Bernoulli site percolation is defined similarly with vertices replacing edges. See Grimmett (1989) for more information about Bernoulli percolation.

2. Mass transport. Bernoulli percolation is one example of an invariant percolation model. If a group Γ acts by automorphisms on a graph X, we call a probability measure on the subgraphs of X a Γ-*invariant percolation model* if it is invariant under the action of Γ. An invariant percolation model $\mathbf{P}$ is a *bond percolation model* if $\mathbf{P}[\mathsf{V}(\omega) = \mathsf{V}(X)] = 1$, where ω denotes a subgraph of X.

We now present the mass-transport principle. Let $m(x, y; \omega)$ be a nonnegative function of three variables: two vertices x, y and a subgraph ω of X. Intuitively, $m(x, y; \omega)$ represents the mass transported from x to y when the percolation subgraph is ω. We suppose that $m(\cdot, \cdot; \cdot)$ is invariant under the diagonal action of Γ, that is, $m(x, y; \omega) = m(gx, gy; g\omega)$ for all x, y, g and ω. The *mass-transport principle* asserts that for any invariant percolation, the expected total mass transported out of any vertex x equals the expected total mass transported into x, that is,

$$\forall\, x \in \mathsf{V} \quad \sum_{y \in \mathsf{V}} M(x, y) = \sum_{y \in \mathsf{V}} M(y, x), \tag{2.1}$$

where $M(x, y) := \mathbf{E}m(x, y; \omega)$ is also invariant under the diagonal action of Γ.

On a Cayley graph, it is straightforward to verify (2.1):

$$\sum_{y \in \mathsf{V}} M(x, y) = \sum_{g \in \Gamma} M(x, gx) = \sum_{g \in \Gamma} M(g^{-1}x, x) = \sum_{y \in \mathsf{V}} M(y, x).$$

The creative element in applying the mass-transport method is to make a judicious choice of the transport function $m(x, y; \omega)$. This depends on the particular application; three such applications are given in this note.

The identity (2.1) is valid on many transitive graphs that are not Cayley graphs, provided their automorphism group is unimodular; see BLPS (1999). The proof of Theorem 1.1 below uses only transitivity of X and (2.1), so it applies to these graphs as well. With minor modifications, the proof also applies to unimodular *quasi-transitive* graphs, that is, graphs $X = (\mathsf{V}, \mathsf{E})$ for which there is some unimodular group Γ of automorphisms with finitely many orbits in V.

Let S be a finite set of generators for a nonamenable group Γ and let $X := X(\Gamma, S)$ be the corresponding Cayley graph. For a finite subgraph $K \subset X$,

set

$$\alpha_K := \frac{1}{|\mathsf{V}(K)|} \sum_{x \in \mathsf{V}(K)} \deg_K(x),$$

where $\deg_K(x)$ refers to the degree of x as a vertex in the graph K. Let

$$\alpha(X) := \sup\{\alpha_K;\ K \subset X \text{ is finite}\}.$$

Since every vertex of X has the same degree, it follows that

$$\alpha(X) + \iota_{\mathsf{E}}(X) = \deg_X(o), \tag{2.2}$$

where $o \in \mathsf{V}(X)$ is the identity of Γ.

THEOREM 2.1. *Let X be a Cayley graph of a nonamenable group Γ with respect to some finite set of generators. Let $o \in \mathsf{V}(X)$ denote the identity of Γ. Suppose that ξ is a random subgraph of X and the distribution of ξ is an invariant bond percolation model on X. If $\mathbf{E}\deg_\xi(o) > \alpha(X)$, then ξ has an infinite component with positive probability.*

REMARK 2.2. It suffices that $\mathbf{E}\deg_\xi(o) \geq \alpha(X)$. See BLPS (1999) for details.

COROLLARY 2.3. *If $\mathbf{P}[e \notin \xi] < \iota_{\mathsf{E}}(X)/\deg_X(o)$ for all $e \in \mathsf{E}(X)$, then ξ has an infinite component with positive probability.*

The proof follows immediately from Theorem 2.1 and (2.2).

PROOF OF THEOREM 2.1. For each $v \in \mathsf{V}(X)$, let $K(v)$ denote the component of ξ containing v. Set

$$m(v,u;\xi) := \begin{cases} \deg_\xi(v)/|K(v)|, & \text{if } K(v) \text{ is finite and } u \in K(v), \\ 0, & \text{otherwise.} \end{cases}$$

Thus,

$$\sum_{u \in \mathsf{V}(X)} m(u,o;\xi) = \alpha_{K(o)} \leq \alpha(X)$$

whenever $K(o)$ is finite, and $\sum_{u \in \mathsf{V}(X)} m(u,o;\xi) = 0$ otherwise. On the other hand,

$$\sum_{u \in \mathsf{V}(X)} m(o,u;\xi) = \begin{cases} \deg_\xi(o), & \text{when } K(o) \text{ is finite,} \\ 0, & \text{otherwise.} \end{cases}$$

Since $m(\cdot,\cdot;\cdot)$ is diagonally invariant, the mass-transport principle gives

$$\begin{aligned} &\mathbf{E}\big[\deg_\xi(o)\,|\,|K(o)| < \infty\big]\mathbf{P}\big[|K(o)| < \infty\big] \\ &\quad = \mathbf{E} \sum_{u \in \mathsf{V}(X)} m(o,u;\xi) = \mathbf{E} \sum_{u \in \mathsf{V}(X)} m(u,o;\xi) \leq \alpha(X)\mathbf{P}\big[|K(o)| < \infty\big]. \end{aligned}$$

When $\mathbf{P}[|K(o)| < \infty] = 1$, we get $\mathbf{E}\deg_\xi(o) \leq \alpha(X)$, as required. □

3. Ruling out a unique infinite cluster.

PROOF OF THEOREM 1.1. For simplicity, we restrict our attention to critical bond percolation; the proof for critical site percolation requires only minor modifications.

The number of infinite clusters is invariant under the action of Γ. By ergodicity, the number of infinite clusters is constant a.s. It follows that this constant must be 0, 1 or ∞; see Newman and Schulman (1981). In this section, we rule out a unique infinite cluster at $p = p_c$, and in the next section, we rule out infinitely many infinite clusters.

Denote the graphical distance in X by d_X, and let ω be critical Bernoulli bond percolation on X. Suppose that almost surely there is a unique infinite cluster U of ω. Let γ_ε be an ε-Bernoulli bond percolation on X that is independent of ω. We use γ_ε to define a bond percolation model ξ_ε on X as follows. For $v \in \mathsf{V}$, let $U(v) \subset U$ be the set of vertices $u \in U$ such that $d_X(v,u) = d_X(v,U)$; that is, $U(v)$ is the set of vertices of U that are closest to v. Given two neighbors $v, w \in \mathsf{V}$, put the edge $[v,w]$ in ξ_ε iff $d_X(v,U) < 1/\varepsilon$, $d_X(w,U) < 1/\varepsilon$, and $U(v) \cup U(w)$ is contained in a connected component of $\omega \setminus \gamma_\varepsilon$. If we fix ω, then

$$\lim_{\varepsilon \to 0} \mathbf{P}\big[[v,w] \notin \xi_\varepsilon | \omega\big] = 0,$$

because every pair of vertices in U is joined by some finite path in U that is unlikely to intersect γ_ε when ε is small. Consequently, the bounded convergence theorem gives

$$\lim_{\varepsilon \to 0} \mathbf{P}\big[[v,w] \notin \xi_\varepsilon\big] = 0.$$

By Corollary 2.3, it follows that for $\varepsilon > 0$ sufficiently small, with positive probability, there will be an infinite component of ξ_ε. However, when there is an infinite ξ_ε-component, there must be an infinite component in $\omega \setminus \gamma_\varepsilon$. As $\omega \setminus \gamma_\varepsilon$ is a $(1-\varepsilon)p_c$-Bernoulli bond percolation, this contradicts the definition of p_c and proves that ω cannot have a unique infinite cluster.

4. Ruling out multiple infinite clusters. In this section, we continue with the proof of Theorem 1.1, dealing with the case that critical percolation produces more than one infinite cluster.

DEFINITION 4.1. Let G be an infinite graph. A vertex $v \in \mathsf{V}(G)$ is an *encounter point* if there are three or more neighbors of v that belong to distinct infinite connected components of $G \setminus \{v\}$.

Suppose that with positive probability, ω has more than one infinite cluster. It is then well known [Burton and Keane (1989)] that the set Y of encounter points of ω is nonempty almost surely.

We say that C has an *ω-neighbor* of v if there is some $x \in C$ with $[x, v] \in \omega$.

LEMMA 4.2. *Consider Bernoulli bond percolation ω with $p \geq p_c$ on a Cayley graph. With $\mathbf{P}_p$-probability* 1, *for every encounter point v of ω and every infinite component C of $\omega \setminus \{v\}$ having an ω-neighbor of v, there is some encounter point of ω in C.*

PROOF. Let $u, v \in \mathsf{V}$. Set $m(u, v; \omega) := 1$ if u and v are in the same cluster of ω and v is the unique encounter point of ω with minimal distance to u (where distances are measured in ω). Otherwise, set $m(u, v; \omega) := 0$. Then $\sum_{u \in \mathsf{V}} m(v, u; \omega) \leq 1$. By the mass-transport principle, it follows that almost surely $\sum_{u \in \mathsf{V}} m(u, v; \omega) < \infty$ for all $v \in \mathsf{V}$. This implies the lemma. □

We now construct an invariant forest ξ whose vertices are Y, the encounter points of ω. The edges of ξ will not necessarily be in E. For each $v \in \mathsf{V}$, let $x(v)$ be a uniformly distributed random variable in $[0, 1]$ such that ω and all $x(v)$ are mutually independent. Suppose that $v \in Y$. For every infinite connected component Z of $\omega \setminus \{v\}$ having an ω-neighbor of v, let $Y(Z, v) \subset Y \cap Z$ be the set of encounter points of ω in Z that are closest (in ω) to v, and let $u(Z, v)$ be the vertex $u \in Y(Z, v)$ that minimizes $x(u)$. Let ξ be the set of all edges $[v, u(Z, v)]$, where $v \in Y$ and Z is an infinite connected component of $\omega \setminus \{v\}$ having an ω-neighbor of v. It is clear that ξ is an invariant graph on V. For every $v \in Y$, the degree of v in ξ is at least 3 by Lemma 4.2.

We now verify that ξ is almost surely a forest; that is, it contains no cycles. Assume otherwise, and suppose that C is a cycle in ξ of length greater than or equal to 3. When $u = u(Z, v)$, consider edges in ξ of the form $[v, u]$ to be directed from v to u. Let v be a vertex in C and note that $C \setminus \{v\}$ is contained in some connected component Z of $\omega \setminus \{v\}$, because v appears only once in C. This shows that the number of edges in C that are directed away from v is at most 1. Since the number of edges in C equals the number of vertices in C, every vertex must have one edge directed toward it and one edge directed away from it. If $[a, b]$ and $[b, c]$ are directed edges in C, then the distance in ω from b to c cannot be larger than the distance from a to b. As C is a cycle, it follows that these distances must be the same. Let z be the vertex in C that maximizes $x(z)$. Then there must be vertices a, b in C such that $[a, b]$ and $[b, z]$ are directed edges of C and $a \neq z$. But a and z are at the same distance from b and $x(a) < x(z)$. This contradicts the existence of the directed edge $[b, z]$ in C, and this contradiction proves that ξ is a forest.

Let $\varepsilon > 0$ and let γ_ε be an ε-Bernoulli bond percolation process. Since ω is critical, almost surely $\omega \setminus \gamma_\varepsilon$ has only finite components. Now define an invariant bond percolation model $\xi_\varepsilon \subset \xi$ as follows. For any edge $[v, u] \in \xi$, let $[v, u] \in \xi_\varepsilon$ iff v and u are in the same component of $\omega \setminus \gamma_\varepsilon$. Then almost surely ξ_ε has only finite components.

We now perform mass transport on V. For any $v \in \mathsf{V}$, let $K(v)$ denote its ξ_ε-connected component (which is almost surely finite) and let $K_\partial(v)$ denote

the set of vertices $u \in K(v)$ that have some ξ-neighbor outside of $K(v)$. Set

$$m(v,u;\xi) := \begin{cases} 1/|K_\partial(v)|, & \text{if } v \in Y \text{ and } u \in K_\partial(v), \\ 0, & \text{otherwise.} \end{cases}$$

By the mass-transport principle, we know that

$$\sum_{u \in \mathsf{V}} \mathbf{E}[m(v,u;\xi)] = \sum_{u \in \mathsf{V}} \mathbf{E}[m(u,v;\xi)]. \tag{4.1}$$

Note that

$$\sum_{u \in \mathsf{V}} m(u,v;\xi) = \begin{cases} |K(v)|/|K_\partial(v)|, & \text{when } v \in Y \cap K_\partial(v), \\ 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

Because ξ is a forest and each vertex in Y has ξ-degree at least 3, it easily follows that $2|K_\partial(v)| \geq |K(v)|$. Consequently, (4.2) shows that $\sum_{u \in \mathsf{V}} m(u,v;\xi) \leq 2$ almost surely. On the other hand, $\sum_{u \in \mathsf{V}} m(v,u;\xi)$ is 1 if $v \in Y$ and 0 otherwise. Therefore, (4.1) and (4.2) give

$$\mathbf{P}[v \in Y] \leq 2\mathbf{P}[v \in Y,\ v \in K_\partial(v)]. \tag{4.3}$$

If we fix ω, ξ, and $e \in \xi$, then the probability that $e \notin \xi_\varepsilon$ tends to 0 as $\varepsilon \to 0$: there is some finite path in ω joining the endpoints of e, and when ε is small, this path is unlikely to meet γ_ε. From the bounded convergence theorem, it therefore follows that for any $v \in \mathsf{V}$, the probability that v has some edge in $\xi \setminus \xi_\varepsilon$ tends to 0 as $\varepsilon \to 0$. In other words,

$$\lim_{\varepsilon \to 0} \mathbf{P}[v \in Y,\ v \in K_\partial(v)] = 0.$$

This contradicts (4.3) since $\mathbf{P}[v \in Y]$ is positive and does not depend on ε. Hence the proof of Theorem 1.1 is complete. □

5. Further results and problems on percolation in nonamenable groups. Recall that for percolation on $\mathbb{Z}^d$, if $\theta(p) > 0$, then there exists with probability 1 a unique infinite open cluster. [See Aizenman, Kesten and Newman (1987) and the elegant proof in Burton and Keane (1989).] As was shown by Grimmett and Newman (1990), for percolation on a (k-regular tree) $\times \mathbb{Z}$, this is not the case if k is sufficiently large: for some values of p, uniqueness holds, while for others, there are infinitely many infinite clusters. Benjamini and Schramm (1996) defined

$$p_u := \inf\{p; \mathbf{P}_p[\text{there is exactly one infinite cluster}] = 1\}$$

and conjectured that $p_c < p_u$ for percolation on nonamenable Cayley graphs. This was established by Lalley (1996) for certain planar Cayley graphs, then by Benjamini and Schramm (1998a) for all nonamenable planar Cayley graphs. Häggström and Peres (1998), in response to another question from Benjamini and Schramm (1996), showed that for any $p > p_u$, uniqueness holds almost surely. Benjamini and Schramm (1996) also asked whether there is uniqueness at p_u in transitive nonamenable graphs. This is true when the graphs are planar [Benjamini and Schramm (1998a)], but Schonmann (1998) has shown it to be false for (k-regular tree) $\times \mathbb{Z}$ when $k \geq 3$. We

note that for Cayley graphs of finitely presented groups with one end, Babson and Benjamini (1999) showed that $p_u < 1$. A brief description of some recent developments concerning percolation on general graphs is presented by Benjamini and Schramm (1998b).

REMARK 5.1. Lemma 4.2 has the following consequence. Consider Bernoulli percolation ω with $p > p_c$ on a Cayley graph, and suppose that ω has more than one infinite cluster a.s. Then with probability 1, any infinite cluster C of ω has infinitely many encounter points. Consequently, the space of ends of C is topologically a Cantor set (in particular, it is uncountable). The same proof works for unimodular quasi-transitive graphs. The result for all quasi-transitive graphs was conjectured by Benjamini and Schramm (1996) and first established by Häggström and Peres (1998) in the unimodular case, using a more complicated argument. Very recently, this result on ends was extended to the nonunimodular case by Häggström, Peres and Schonmann (1998). Their argument is based on different principles; in particular, it is not known in the nonunimodular case whether for $p \in (p_c, p_u)$, every infinite cluster must contain infinitely many encounter points.

Theorem 1.1 extends easily to some other percolation models; for example, if different (positive) retention probabilities are assigned to different generators in a nonamenable Cayley graph X, then for parameter values on the corresponding critical submanifold, there are no infinite clusters in X.

To analyze the Ising model on nonamenable groups, it would be useful to know whether Theorem 1.1 also extends to the FK random cluster model for $q \leq 2$; see, for example, Grimmett (1995) for definitions and background. For $q > 2$, the (wired) random cluster model on a regular tree has infinite clusters at criticality; see Häggström (1996).

Combining Theorem 1.1 with a result of Häggström and Peres (1998) gives that for nonamenable Cayley graphs, the function $\theta(p) = \mathbf{P}_p[K(o) \text{ is infinite}]$ is continuous for all $p \in [0, 1]$. One may wish for quantitative versions of this result, for example as in the following.

PROBLEM 5.1. Bound the modulus of continuity of $\theta(\cdot)$ for nonamenable Cayley graphs. Is $\sup_{p > p_c} \theta'(p) < \infty$?

Since $\theta(p_c) = 0$, the component $K(o)$ satisfies $\lim_{R \to \infty} \mathbf{P}_{p_c}[\text{diam } K(o) > R] = 0$.

PROBLEM 5.2. For a nonamenable Cayley graph X, find a bound for $\mathbf{P}_{p_c}[\text{diam } K(o) > R]$ that tends to zero as $R \to \infty$.

It seems reasonable that such a bound would involve only the isoperimetric constant of X and perhaps the degree of the vertices.

Let $\tau_p(o, v)$ denote the connectivity function, that is, the $\mathbf{P}_p$-probability that o and v are in the same cluster.

PROBLEM 5.3. For a nonamenable Cayley graph X, find a bound for $\tau_{p_c}(o, v)$ that goes to zero as $d_X(o, v) \to \infty$.

We expect that $\tau_{p_c}(o, v) \leq \alpha^{d_X(o,v)}$ for some $\alpha < 1$. Lyons and Schramm (1998) have recently shown that on Cayley graphs, $\inf_v \tau_p(o, v) = 0$ when there is no uniqueness at p. Schramm (unpublished) has also shown that for every nonamenable Cayley graph G, there is an $\alpha < 1$ with $\tau_{p_c}(o, v) < \alpha^{d_X(o,v)}$ for infinitely many v.

REFERENCES

ADAMS, S. (1990). Trees and amenable equivalence relations. *Ergodic Theory Dynamic Systems* **10** 1–14.

AIZENMAN, M., KESTEN, H. and NEWMAN, C. M. (1987). Uniqueness of the infinite cluster and continuity of connectivity functions for short and long range percolation. *Comm. Math. Phys.* **111** 505–531.

BABSON, E. and BENJAMINI, I. (1999). Cut sets and normed cohomology with applications to percolation. *Proc. Amer. Math. Soc.* **127** 589–597.

BENJAMINI, I., LYONS, R., PERES, Y. and SCHRAMM, O. (1999). Group-invariant percolation on graphs. *Geom. Funct. Anal.* **9** 29–66.

BENJAMINI, I. and SCHRAMM, O. (1996). Percolation beyond $\mathbb{Z}^d$, many questions and a few answers. *Electronic Comm. Probab.* **1** 71–82.

BENJAMINI, I. and SCHRAMM, O. (1998a). Percolation in the hyperbolic plane. Unpublished manuscript.

BENJAMINI, I. and SCHRAMM, O. (1998b). Recent progress on percolation beyond $\mathbb{Z}^d$. http://www.wisdom.weizmann.ac.il/~schramm/papers/pyond-rep/.

BURTON, R. M. and KEANE, M. (1989). Density and uniqueness in percolation. *Comm. Math. Phys.* **121** 501–505.

GRIMMETT, G. R. (1989). *Percolation*. Springer, New York.

GRIMMETT, G. R. (1995). The stochastic random-cluster process and the uniqueness of random-cluster measures. *Ann. Probab.* **23** 1461–1510.

GRIMMETT, G. R. and NEWMAN, C. M. (1990). Percolation in $\infty + 1$ dimensions. In *Disorder in Physical Systems* (G. R. Grimmett and D. J. A. Welsh, eds.) 219–240. Clarendon Press, Oxford.

HÄGGSTRÖM, O. (1996). The random-cluster model on a homogeneous tree. *Probab. Theory Related Fields* **104** 231–253.

HÄGGSTRÖM, O. (1997). Infinite clusters in dependent automorphism invariant percolation on trees. *Ann. Probab.* **25** 1423–1436.

HÄGGSTRÖM, O. and PERES, Y. (1998). Monotonicity of uniqueness for percolation on transitive graphs: all infinite clusters are born simultaneously. *Probab. Theory Related Fields* **113** 273–285.

HÄGGSTRÖM, O., PERES, Y. and SCHONMANN, R. (1998). Percolation on transitive graphs as a coalescent process: relentless merging followed by simultaneous uniqueness. In *Perplexing Probability Problems: Papers in Honor of H. Kesten* (M. Bramson and R. Durrett, eds.). Birkhäuser, Boston. To appear.

HARA, T. and SLADE, G. (1994). Mean-field behaviour and the lace expansion. In *Probability and Phase Transition* (G. R. Grimmett, ed.) 87–122. Kluwer, Dordrecht.

HARRIS, T. E. (1960). A lower bound for the critical probability in a certain percolation process. *Proc. Cambridge Philos. Soc.* **56** 13–20.

KESTEN, H. (1980). The critical probability of bond percolation on the square lattice equals $\frac{1}{2}$. *Comm. Math. Phys.* **74** 41–59.

LALLEY, S. P. (1996). Percolation on Fuchsian groups. *Ann. Inst. H. Poincaré Probab. Statist.* **34** 151–178.

LYONS, R. and SCHRAMM, O. (1998). Indistinguishability of percolation clusters. *Ann. Probab.* To appear.

NEWMAN, C. M. and SCHULMAN, L. S. (1981). Infinite clusters in percolation models. *J. Statist. Phys.* **26** 613–628.

SCHONMANN, R. H. (1998). Percolation in $\infty + 1$ dimensions at the uniqueness threshold. In *Perplexing Probability Problems: Papers in Honor of H. Kesten* (M. Bramson and R. Durrett, eds.). Birkhäuser, Boston. To appear.

VAN DEN BERG, J. and MEESTER, R. W. J. (1991). Stability properties of a flow process in graphs. *Random Structures Algorithms* **2** 335–341.

WU, C. C. (1993). Critical behavior of percolation and Markov fields on branching planes. *J. Appl. Probab.* **30** 538–547.

I. BENJAMINI
MATHEMATICS DEPARTMENT
WEIZMANN INSTITUTE OF SCIENCE
REHOVOT 76100
ISRAEL
E-MAIL: itai@wisdom.weizmann.ac.il

R. LYONS
DEPARTMENT OF MATHEMATICS
INDIANA UNIVERSITY
BLOOMINGTON, INDIANA 47405-5701
E-MAIL: rdlyons@indiana.edu

Y. PERES
INSTITUTE OF MATHEMATICS
HEBREW UNIVERSITY
GIVAT RAM, JERUSALEM 91904
ISRAEL
AND
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720-3860
E-MAIL: peres@stat.berkeley.edu

O. SCHRAMM
MATHEMATICS DEPARTMENT
WEIZMANN INSTITUTE OF SCIENCE
REHOVOT 76100
ISRAEL
E-MAIL: schramm@wisdom.weizmann.ac.il

The Annals of Probability
1999, Vol. 27, No. 4, 1809–1836

INDISTINGUISHABILITY OF PERCOLATION CLUSTERS

BY RUSSELL LYONS[1] AND ODED SCHRAMM[2]

Indiana University and Weizmann Institute of Science

We show that when percolation produces infinitely many infinite clusters on a Cayley graph, one cannot distinguish the clusters from each other by any invariantly defined property. This implies that uniqueness of the infinite cluster is equivalent to nondecay of connectivity (a.k.a. long-range order). We then derive applications concerning uniqueness in Kazhdan groups and in wreath products and inequalities for p_u.

1. Introduction. Grimmett and Newman (1990) showed that if T is a regular tree of sufficiently high degree, then there are $p \in (0,1)$ such that Bernoulli(p) percolation on $T \times \mathbb{Z}$ has infinitely many infinite components a.s. Benjamini and Schramm (1996) conjectured that the same is true for any Cayley graph of any finitely generated nonamenable group. (A finitely generated group Γ is nonamenable iff its Cayley graph satisfies $\inf_K |\partial K|/|K| > 0$, where K runs over the finite nonempty vertex sets. See Section 2 for all other definitions.) This conjecture has been verified for planar Cayley graphs of high genus by Lalley (1998) and for all planar lattices in the hyperbolic plane by Benjamini and Schramm (1998).

The present paper is concerned with the percolation phase where there are multiple infinite clusters. We show that under quite general assumptions, the infinite clusters are indistinguishable from each other by any invariantly defined property.

Let G be a Cayley graph of a finitely generated group (more generally, G can be a transitive unimodular graph).

THEOREM 1.1. *Consider Bernoulli bond percolation on G with some survival parameter $p \in (0,1)$, and let $\mathscr{A}$ be a Borel measurable set of subgraphs of G. Assume that $\mathscr{A}$ is invariant under the automorphism group of G. Then either a.s. all infinite percolation components are in $\mathscr{A}$, or a.s. they are all outside of $\mathscr{A}$.*

For example, $\mathscr{A}$ might be the collection of all transient subgraphs of G, or the collection of all subgraphs that have a given asymptotic rate of growth or the collection of all subgraphs that have no vertex of degree 5.

Received December 1998.
[1]Supported in part by Varon Visiting Professorship at the Weizmann Institute of Science and NSF Grant DMS-98-02663.
[2]Supported in part by the Sam and Ayala Zacks Professorial Chair.
AMS 1991 *subject classifications.* Primary 82B43, 60B99; secondary 60K35, 60D05.
Key words and phrases. Finite energy, Cayley graph, group, Kazhdan, wreath product, uniqueness, connectivity, transitive, nonamenable.

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_23, © Springer Science+Business Media, LLC 2011

If $\mathscr{A}$ is the collection of all transient subgraphs of G, then this shows that almost surely, either all infinite clusters of ω are transient (meaning that simple random walk on them is transient), or all clusters are recurrent. In fact, as shown in Proposition 3.11, if G is nonamenable, then a.s. all infinite clusters are transient if Bernoulli percolation produces more than one infinite component. In Benjamini, Lyons and Schramm (1999), it is shown that the same is true if Bernoulli percolation produces a single infinite component.

Theorem 1.1 is a particular case of Theorem 3.3 below, which applies also to some non-Bernoulli percolation processes and to more general $\mathscr{A}$.

A collection of subgraphs $\mathscr{A}$ in G is *increasing* if whenever $H \in \mathscr{A}$ and $H \subset H' \subset G$, we have $H' \in \mathscr{A}$. Häggström and Peres (1999) have proved that for increasing $\mathscr{A}$, Theorem 1.1 holds, except possibly for a single value of $p \in (0, 1)$. They have also proved the following theorem.

THEOREM 1.2 (Uniqueness monotonicity). *Let $p_1 < p_2$ and $\mathbf{P}_i$ $(i = 1, 2)$ be the corresponding Bernoulli(p_i) bond percolation processes on G. If there is a unique infinite cluster $\mathbf{P}_1$-a.s., then there is a unique infinite cluster $\mathbf{P}_2$-a.s. Furthermore, in the standard coupling of Bernoulli percolation processes, if there exists an infinite cluster $\mathbf{P}_1$-a.s., then a.s. every infinite p_2-cluster contains an infinite p_1-cluster.*

Here, we refer to the standard coupling of Bernoulli(p) percolation for all p, where each edge $e \in \mathsf{E}$ is assigned an independent uniform $[0, 1]$ random variable $U(e)$ and the edges where $U(e) \le p$ are retained for Bernoulli(p) percolation.

The first part of this theorem partially answers a question of Benjamini and Schramm (1996); the full answer, removing the assumption of unimodularity, has been provided by Schonmann (1999a). It follows from Theorem 1.2 that the set of p such that Bernoulli(p) bond percolation on G has more than one infinite cluster a.s. is an interval, called the *nonuniqueness phase*. It is well known that in the nonuniqueness phase, the number of infinite clusters is a.s. infinite [see Newman and Schulman (1981)]. The interval of p such that Bernoulli(p) percolation on G produces a single infinite cluster a.s. is called the *uniqueness phase*. The infimum of the p in the uniqueness phase is denoted $p_u = p_u(G)$.

To illustrate the usefulness of Theorem 1.1, we now show that it implies Theorem 1.2.

PROOF OF THEOREM 1.2. Suppose that there exists an infinite cluster $\mathbf{P}_1$-a.s. Let ω be the open subgraph of the $\mathbf{P}_2$ process and let η be an independent Bernoulli(p_1/p_2) percolation process. Thus, $\omega \cap \eta$ has the law of $\mathbf{P}_1$ and, in fact, $(\omega \cap \eta, \omega)$ has the same law as the standard coupling of $\mathbf{P}_1$ and $\mathbf{P}_2$. By assumption, $\omega \cap \eta$ has an infinite cluster a.s. Thus, for some cluster C of ω, we have $C \cap \eta$ is infinite with positive probability, hence, by Kolmogorov's 0–1 law, with probability 1. By Theorem 1.1, this holds for every infinite cluster C of ω. □

Our main application of Theorem 1.1 is to prove in Section 4 a sufficient condition for having a unique infinite cluster. Define $\tau(x, y)$ to be the probability that x and y are in the same cluster. The process is said to exhibit *connectivity decay* if

$$\inf\{\tau(x, y)\colon x, y \in \mathsf{V}\} = 0.$$

When a process does not have connectivity decay, it is said to exhibit *long-range order*. We show that for Bernoulli percolation (and some other percolation processes), nonuniqueness of the infinite cluster is equivalent to connectivity decay. One direction is easy; namely, if a.s. there is a unique infinite percolation cluster, then there is long-range order. This is an immediate consequence of Harris's inequality (that increasing events are positively correlated). Although the other direction seems intuitively obvious as well, its proof is surprisingly difficult (and it is still open for nonunimodular transitive graphs).

We note that connectivity decay in the nonuniqueness phase easily implies Theorem 3.2 of Schonmann (1999a) for the case of unimodular transitive graphs. (Our result also deals with percolation at p_u and extends Schonmann's theorem to more general percolation processes, but his result extends to Bernoulli percolation on nonunimodular transitive graphs.)

REMARK 1.3. Fix a base vertex $o \in \mathsf{V}$ in G, and consider Bernoulli percolation in the nonuniqueness phase. Although $\inf\{\tau(o, v)\colon v \in \mathsf{V}\} = 0$, it may happen that

$$\lim_{d \to \infty} \sup\{\tau(o, v)\colon \operatorname{dist}(o, v) > d\} > 0.$$

For example, take the Cayley graph of the free product $\mathbb{Z}^2 * \mathbb{Z}_2$ with the usual generating set, that is, the Cayley graph corresponding to the presentation $\langle a, b, c \mid ab = ba,\ c = c^{-1}\rangle$. We have $p_u = 1$, because the removal of any edge of the Cayley graph corresponding to the generator c disconnects the graph. On the other hand, for $p > p_c(\mathbb{Z}^2)$, we have that all pairs of vertices in the same $\mathbb{Z}^2$ fiber (i.e., pairs g, h such that $g^{-1}h$ is in the group generated by a and b) are in the same cluster with probability bounded away from 0.

Benjamini and Schramm (1996) asked for which Cayley graphs G one has $p_u(G) < 1$. It is easy to show that $p_u(G) = 1$ when G has more than one end. Babson and Benjamini (1999) showed that $p_u(G) < 1$ when G is a finitely presented group with one end. Häggström, Peres and Schonmann (1999) have shown that $p_u(G) < 1$ if G is a Cartesian product of infinite transitive graphs. Here, we show that this is also true of certain other classes of groups with one end, namely infinite Kazhdan groups (Corollary 6.6) and wreath products (Corollary 6.8). Presumably, all quasi-transitive (infinite) graphs with one end have $p_u < 1$.

As another consequence of indistinguishability, we prove in Theorem 6.12 the inequalities $p_u(G \times H') \geq p_u(G \times H)$ and, in particular, $p_u(G) \geq p_u(G \times H)$, for Cayley graphs (or, more generally, unimodular transitive graphs) G, H', H such that G is infinite and $H' \subset H$.

Several other uses of cluster indistinguishability appear in Benjamini, Lyons, and Schramm (1999).

Crucial techniques for our proofs are the mass-transport principle and stationarity of delayed simple random walk, both explained below. These techniques were introduced in the study of percolation by Häggström (1997). In Section 5, we prove an ergodicity property for delayed random walk.

2. Background. We begin with some graph-theoretic terminology. Let $G = (\mathsf{V}(G), \mathsf{E}(G))$ be a graph with vertex set $\mathsf{V}(G)$ and (symmetric) edge set $\mathsf{E}(G) \subseteq \mathsf{V}(G) \times \mathsf{V}(G)$. When there is an edge in G joining vertices u, v, we say that u and v are *adjacent* and write $u \sim v$. We always assume that the number of vertices adjacent to any given vertex is finite. The *degree* $\deg v = \deg_G v$ of a vertex $v \in \mathsf{V}(G)$ is the number of edges incident with it. A *tree* is a connected graph with no cycles. A *forest* is a graph whose connected components are trees. The *distance* between two vertices $u, v \in \mathsf{V}(G)$ is denoted by $\mathrm{dist}(v, u) = \mathrm{dist}_G(v, u)$ and is the least number of edges of a path in G connecting v and u. Given a set of vertices $K \subset \mathsf{V}(G)$, we let ∂K denote its edge boundary, that is, the set of edges in $\mathsf{E}(G)$ having one vertex in K and one outside of K. A transitive graph G is *amenable* if $\inf |\partial K|/|K| = 0$, where K runs over all finite nonempty vertex sets $K \subset \mathsf{V}(G)$.

An infinite set of vertices $V_0 \subset \mathsf{V}(G)$ is *end convergent* if for every finite $K \subset \mathsf{V}(G)$, there is a component of $G - K$ that contains all but finitely many vertices of V_0. Two end-convergent sets V_0, V_1 are *equivalent* if $V_0 \cup V_1$ is end convergent. An *end* of G is an equivalence class of end-convergent sets. Let ξ be an end of G. A *neighborhood* of ξ is a set of vertices in G that intersects every end-convergent set in ξ. In particular, when $K \subset \mathsf{V}(G)$ is finite, there is a component of $G - K$ that is a neighborhood of ξ.

Let Γ be a finitely generated group and S a finite symmetric generating set for Γ. Then the (right) *Cayley* graph $G = G(\Gamma, S)$ is the graph with vertices $\mathsf{V}(G) := \Gamma$ and edges $\mathsf{E}(G) := \{[v, vs]: v \in \Gamma,\ s \in S\}$.

Now suppose that $G = (\mathsf{V}(G), \mathsf{E}(G))$ is any graph, not necessarily a Cayley graph. An *automorphism* of G is a bijection of $\mathsf{V}(G)$ with itself that preserves adjacency; $\mathrm{Aut}(G)$ denotes the group of all automorphisms of G with the topology of pointwise convergence. If $\Gamma \subset \mathrm{Aut}(G)$, we say that Γ is (vertex) *transitive* if for every $u, v \in \mathsf{V}(G)$, there is a $\gamma \in \Gamma$ with $\gamma u = v$. The graph G is *transitive* if $\mathrm{Aut}(G)$ is transitive. The graph G is *quasi-transitive* if $\mathsf{V}(G)/\mathrm{Aut}(G)$ is finite; that is, there is a finite set of vertices V_0 such that $\mathsf{V}(G) = \{\gamma v: \gamma \in \mathrm{Aut}(G),\ v \in V_0\}$. Note that any finitely generated group acts transitively on any of its Cayley graphs by the automorphisms $\gamma: v \mapsto \gamma v$.

It may seem that the most natural class of graphs on which percolation should be studied is the class of transitive (or quasi-transitive) locally finite graphs. At first sight, one might suspect that any theorem about percolation on Cayley graphs should hold for transitive graphs. However, somewhat surprisingly, this impression is not correct. It turns out that theorems about percolation on Cayley graphs "always" generalize to unimodular transitive

graphs (to be defined shortly), but nonunimodular transitive graphs are quite different.

Recall that on every closed subgroup $\Gamma \subset \mathrm{Aut}(G)$, there is a unique (up to a constant scaling factor) Borel measure that, for every $\gamma \in \Gamma$, is invariant under left multiplication by γ; this measure is called (left) *Haar measure*. The group Γ is *unimodular* if Haar measure is also invariant under right multiplication. For example, when Γ is finitely generated and G is the (right) Cayley graph of Γ, on which Γ acts by left multiplication, then $\Gamma \subset \mathrm{Aut}(G)$ is (obviously) closed, unimodular and transitive. The Haar measure in this case is (a constant times) counting measure. A [quasi-]transitive graph G is said to be *unimodular* if $\mathrm{Aut}(G)$ is unimodular. By Benjamini, Lyons, Peres and Schramm (1999) [hereinafter referred to as BLPS (1999)], a transitive graph is unimodular iff there is some unimodular transitive closed subgroup of $\mathrm{Aut}(G)$. It is not hard to show that a transitive closed subgroup $\Gamma \subset \mathrm{Aut}(G)$ is unimodular iff for all $x, y \in \mathsf{V}(G)$, we have

$$\begin{aligned}&\left|\{z \in \mathsf{V}(G)\colon \exists\ \gamma \in \Gamma\ \gamma x = x \text{ and } \gamma y = z\}\right| \\ &\quad = \left|\{z \in \mathsf{V}(G)\colon \exists\ \gamma \in \Gamma\ \gamma y = y \text{ and } \gamma x = z\}\right|\end{aligned}$$

[see Trofimov (1985)].

Here, unimodularity will be used only in the following form.

THEOREM 2.1 (Unimodular mass-transport principle). *Let G be a graph with a transitive unimodular closed automorphism group $\Gamma \subset \mathrm{Aut}(G)$. Let $o \in \mathsf{V}(G)$ be an arbitrary base point. Suppose that $\phi\colon \mathsf{V}(G) \times \mathsf{V}(G) \to [0, \infty]$ is invariant under the diagonal action of Γ. Then*

$$\sum_{v \in \mathsf{V}(G)} \phi(o, v) = \sum_{v \in \mathsf{V}(G)} \phi(v, o). \tag{2.1}$$

See BLPS (1999) for a discussion of this principle and for a proof. In fact, Γ is unimodular iff (2.1) holds for every such ϕ. Also see BLPS (1999) for further discussion of the relevance of unimodularity to percolation.

Given a set A, let 2^A be the collection of all subsets $\eta \subset A$, equipped with the σ-field generated by the events $\{a \in \eta\}$, where $a \in A$. A *bond percolation process* is a pair $(\mathbf{P}, \omega)$, where ω is a random element in 2^{E} and $\mathbf{P}$ denotes the distribution (law) of ω. Sometimes, for brevity, we shall just say that ω is a (bond) percolation. A *site percolation process* $(\mathbf{P}, \omega)$ is given by a probability measure $\mathbf{P}$ on $2^{\mathsf{V}(G)}$, while a (mixed) *percolation* is given by a probability measure on $2^{\mathsf{V}(G) \cup \mathsf{E}(G)}$ that is supported on subgraphs of G. If ω is a bond percolation process, then $\hat{\omega} := \mathsf{V}(G) \cup \omega$ is the associated mixed percolation. In this case, we shall often not distinguish between ω and $\hat{\omega}$, and think of ω as a subgraph of G. Similarly, if ω is a site percolation, there is an associated mixed percolation $\hat{\omega} := \omega \cup (\mathsf{E}(G) \cap (\omega \times \omega))$, and we shall often not bother to distinguish between ω and $\hat{\omega}$.

If $v \in \mathsf{V}(G)$ and ω is a percolation on G, the *cluster* (or *component*) $C(v)$ of v in ω is the set of vertices in $\mathsf{V}(G)$ that can be connected to v by paths contained in ω. We shall often not distinguish between the cluster $C(v)$ and the graph $(C(v), (C(v) \times C(v)) \cap \omega)$ whose vertices are $C(v)$ and whose edges are the edges in ω with endpoints in $C(v)$.

Let $p \in [0,1]$. Then *Bernoulli*(p) *bond percolation* $(\mathbf{P}_p, \omega)$ on G is the product measure on $2^{\mathsf{E}(G)}$ that satisfies $\mathbf{P}_p[e \in \omega] = p$ for all $e \in \mathsf{E}(G)$. Similarly, one defines *Bernoulli*(p) *site percolation* on $2^{\mathsf{V}(G)}$. The *critical probability* $p_c(G)$ is the infimum over all $p \in [0,1]$ such that there is $\mathbf{P}_p$-positive probability for the existence of an infinite connected component in ω.

Aizenman, Kesten and Newman (1987) showed that Bernoulli(p) percolation in $\mathbb{Z}^d$ has a.s. at most one infinite cluster. Burton and Keane (1989) gave a much simpler argument that generalizes from $\mathbb{Z}^d$ to any amenable Cayley graph, though this generalization was not mentioned explicitly until Gandolfi, Keane and Newman (1992). It follows that $p_u(G) = p_c(G)$ when G is an amenable Cayley graph. For background on percolation, especially in $\mathbb{Z}^d$, see Grimmett (1989).

Suppose that Γ is an automorphism group of a graph G. A percolation process $(\mathbf{P}, \omega)$ on G is Γ-*invariant* if $\mathbf{P}$ is invariant under each $\gamma \in \Gamma$. This is, of course, the case for Bernoulli percolation.

3. Cluster indistinguishability.

DEFINITION 3.1. Let G be graph and Γ a closed (vertex-) transitive subgroup of $\mathrm{Aut}(G)$. Let $(\mathbf{P}, \omega)$ be a Γ-invariant bond percolation process on G. We say that $\mathbf{P}$ has *indistinguishable infinite clusters* if for every measurable $\mathscr{A} \subset 2^{\mathsf{V}(G)} \times 2^{\mathsf{E}(G)}$ that is invariant under the diagonal action of Γ, almost surely, for all infinite clusters C of ω, we have $(C, \omega) \in \mathscr{A}$, or for all infinite clusters C, we have $(C, \omega) \notin \mathscr{A}$.

DEFINITION 3.2. Let $G = (\mathsf{V}(G), \mathsf{E}(G))$ be a graph. Given a set $A \in 2^{\mathsf{E}(G)}$ and an edge $e \in \mathsf{E}(G)$, denote $\Pi_e A := A \cup \{e\}$. For $\mathscr{A} \subset 2^{\mathsf{E}(G)}$, we write $\Pi_e \mathscr{A} := \{\Pi_e A \colon A \in \mathscr{A}\}$. A bond percolation process $(\mathbf{P}, \omega)$ on G is *insertion tolerant* if $\mathbf{P}[\Pi_e \mathscr{A}] > 0$ for every $e \in \mathsf{E}(G)$ and every measurable $\mathscr{A} \subset 2^{\mathsf{E}(G)}$ satisfying $\mathbf{P}[\mathscr{A}] > 0$.

For example, Bernoulli(p) bond percolation is insertion tolerant when $p \in (0,1]$.

A percolation ω is *deletion tolerant* if $\mathbf{P}[\Pi_{\neg e} \mathscr{A}] > 0$ whenever $e \in \mathsf{E}(G)$ and $\mathbf{P}[\mathscr{A}] > 0$, where $\Pi_{\neg e} \omega := \omega - \{e\}$. It turns out that deletion tolerance and insertion tolerance have very different implications. For indistinguishability of infinite clusters, we shall need insertion tolerance; deletion tolerance does not imply indistinguishability of infinite clusters (see Example 3.15 below). A percolation that is both insertion and deletion tolerant is usually said to have "finite energy." Gandolfi, Keane and Newman (1992) use the words "positive finite energy" in place of "insertion tolerance."

THEOREM 3.3 (Cluster indistinguishability). *Let G be a graph with a transitive unimodular closed automorphism group $\Gamma \subset \mathrm{Aut}(G)$. Every Γ-invariant, insertion-tolerant, bond percolation process on G has indistinguishable infinite clusters.*

Similar statements hold for site and mixed percolations and the proofs go along the same lines. Likewise, the proof extends to quasi-transitive unimodular automorphism groups.

Initially, we could only establish the theorem under the assumption of *strong insertion-tolerance*; that is, $\mathbf{P}[\Pi_e \mathscr{A}] \geq \delta \mathbf{P}[\mathscr{A}]$ for some constant $\delta > 0$. We are grateful to Olle Häggström for pointing out how to deal with the general case.

REMARK 3.4 (Scenery). For some purposes, the following more general form of this theorem is useful. Let G be a graph and Γ a transitive group acting on G. Suppose that X is either V, E, or $\mathsf{V} \cup \mathsf{E}$. Let Q be a measurable space and $\Omega := 2^{\mathsf{E}} \times Q^X$. A probability measure $\mathbf{P}$ on Ω will be called a *bond percolation with scenery* on G. The projection onto 2^{E} is the underlying percolation and the projection onto Q^X is the scenery. If $(\omega, q) \in \Omega$, we set $\Pi_e(\omega, q) := (\Pi_e \omega, q)$. We say that the percolation with scenery $\mathbf{P}$ is *insertion-tolerant* if, as before, $\mathbf{P}[\Pi_e \mathscr{B}] > 0$ for every measurable $\mathscr{B} \subset \Omega$ of positive measure. We say that $\mathbf{P}$ has *indistinguishable infinite clusters* if for every $\mathscr{A} \subset 2^{\mathsf{V}} \times 2^{\mathsf{E}} \times Q^X$ that is invariant under the diagonal action of Γ, for $\mathbf{P}$-a.e. (ω, q), either all infinite clusters C of ω satisfy $(C, \omega, q) \in \mathscr{A}$ or they all satisfy $(C, \omega, q) \notin \mathscr{A}$. Theorem 3.3 holds also for percolation with scenery (with the same proof as given below).

A percolation with scenery that comes up naturally is as follows. Fix $p_1 < p_2$ in $(0, 1)$. Let $z_e (e \in \mathsf{E})$ be i.i.d. with each z_e distributed according to uniform measure on $[0, 1]$. Let ω_j be the set of $e \in \mathsf{E}$ with $z_e < p_j$, $j = 1, 2$. Then (ω_2, ω_1) is an insertion-tolerant percolation with scenery, where $\omega_1 \in 2^{\mathsf{E}}$ is considered the scenery. See Häggström and Peres (1999), Häggström, Peres and Schonmann (1999), Schonmann (1999a) and Alexander (1995) for examples where this process is studied.

Say that an infinite cluster C is of *type* $\mathscr{A}$ if $(C, \omega) \in \mathscr{A}$; otherwise, say that it is of *type* $\neg \mathscr{A}$. Suppose that there is an infinite cluster C of ω and an edge $e \in \mathsf{E}$ with $e \notin \omega$ such that the connected component C' of $\omega \cup \{e\}$ that contains C has a type different from the type of C. Then e is called *pivotal* for (C, ω).

We begin with an outline of the proof of Theorem 3.3. Assume that the theorem fails. First, we shall show that with positive probability, given that a vertex belongs to an infinite cluster, there are pivotal edges at some distance r of the vertex. By Proposition 3.11, a.s. when there is more than one infinite cluster of ω, each such cluster is transient for the simple random walk, and hence for the so-called delayed simple random walk (DSRW). The DSRW on a cluster of ω is stationary (in the sense of Lemma 3.13). Fix a base point $o \in \mathsf{V}$, and let W be the DSRW on the ω-cluster of o with $W(0) = o$. Let n be

large, and let e be a uniform random edge at distance r from $W(n)$. When DSRW is transient, with probability bounded away from zero, e is pivotal for $(C(o), \omega)$ and $W(j)$ is not an endpoint of e for any time $j < n$. On this event, set $\omega' := \Pi_e \omega$. Then ω' is, up to a controllable factor, as likely as ω by insertion tolerance. By transience, e is far from o with high probability. Since e is pivotal, the type of $C(o)$ is different in ω and in ω'. Since ω and ω' are the same in a large neighborhood of o, this shows that the type of $C(o)$ cannot be determined with arbitrary accuracy by looking at ω in a large neighborhood of o. This contradicts the measurability of $\mathscr{A}$ and establishes the theorem.

One can say that the proof is based on the contradictory prevalence of pivotal edges. To put the situation in the correct perspective, we point out that there are events depending on i.i.d. zero–one variables that have infinitely many "pivotals" with positive probability. For example, let $m_1, m_2, \dots$ be a sequence of positive integers such that $\sum_k 2^{-m_k} < \infty$ but $\sum_k m_k 2^{-m_k} = \infty$, and let $\langle x_{k,j} : j \in \{1, \dots, m_k\} \rangle$ be i.i.d. random variables with $\mathbf{P}[x_{k,j} = 0] = \mathbf{P}[x_{k,j} = 1] = 1/2$. Let $\mathscr{X}$ be the event that there is a k such that $x_{k,j} = 1$ for $j = 1, 2, \dots, m_k$. Then with positive probability $\mathscr{X}$ has infinitely many pivotals; that is, there are infinitely many (k, j) such that changing $x_{k,j}$ from 0 to 1 will change from $\neg \mathscr{X}$ to $\mathscr{X}$. [This is a minor variation on an example described by Häggström and Peres (private communication).]

We now prove the lemmas necessary for Theorem 3.3.

LEMMA 3.5 (Pivotals). *Suppose that* $(\mathbf{P}, \omega)$ *is an insertion-tolerant percolation process on a graph* G, *and let* $\mathscr{A} \subset 2^{\mathsf{V}(G)} \times 2^{\mathsf{E}(G)}$ *be measurable. Assume that there is positive probability for coexistence of infinite clusters in* $\mathscr{A}$ *and* $\neg \mathscr{A}$. *Then with positive probability, there is an infinite cluster* C *of the percolation that has a pivotal edge.*

PROOF. Let k be the least integer such that there is positive probability that there are infinite clusters of different types with distance between them equal to k. Clearly, $k > 0$. Suppose that γ is a path of length k such that with positive probability, γ connects infinite clusters of different types; let $\mathscr{G}$ be the event that γ connects infinite clusters of different types. Given $\mathscr{G}$, there are exactly two infinite clusters that intersect γ, by the minimality of k. Let e be the first edge in γ. When $\omega \in \mathscr{G}$, let C_1 and C_2 be the two infinite clusters that γ connects and let C_1' and C_2' be the infinite clusters of $\Pi_e \omega$ that contain C_1 and C_2, respectively. Note that, conditioned on $\mathscr{G}$, the distance between C_1' and C_2' is less than k. (The possibility that $C_1' = C_2'$ is allowed.) Since $\Pi_e \mathscr{G}$ has positive probability, the definition of k ensures that the type of $(C_1', \Pi_e \omega)$ equals the type of $(C_2', \Pi_e \omega)$ a.s. Hence, e is pivotal for (C_1, ω) or for (C_2, ω) whenever $\omega \in \mathscr{G}$. This proves the lemma. □

LEMMA 3.6. *Let* Γ *be a transitive closed subgroup of the automorphism group of a graph* G. *If* $(\mathbf{P}, \omega)$ *is a* Γ*-invariant insertion-tolerant percolation process on an infinite graph* G, *then almost every ergodic component of* ω *is insertion tolerant.*

REMARK 3.7. A stronger and more general statement is Lemma 1 of Gandolfi, Keane and Newman (1992).

PROOF OF LEMMA 3.6. This is the same as saying that for every Γ-invariant event $\mathscr{G}$ of positive probability, the probability measure $\mathbf{P}[\,\cdot \mid \mathscr{G}]$ is insertion tolerant. If $\mathscr{G}$ is a tail event, then $\Pi_e \mathscr{G} = \mathscr{G}$, and insertion tolerance of $\mathbf{P}[\,\cdot \mid \mathscr{G}]$ follows from the calculation

$$\mathbf{P}[\,\Pi_{e^+}\mathscr{A} \mid \mathscr{G}] = \frac{\mathbf{P}[\,\Pi_{e^+}\mathscr{A} \cap \mathscr{G}]}{\mathbf{P}[\,\mathscr{G}]} = \frac{\mathbf{P}[\,\Pi_{e^+}\mathscr{A} \cap \Pi_e \mathscr{G}]}{\mathbf{P}[\,\mathscr{G}]} = \frac{\mathbf{P}[\,\Pi_e(\mathscr{A} \cap \mathscr{G})]}{\mathbf{P}[\,\mathscr{G}]}.$$

But every Γ-invariant event is a tail event (mod 0): Write E as an increasing union of finite subsets E_n. Let $\mathscr{G}_n$ be events that do not depend on edges in E_n and such that $\sum_n \mathbf{P}[\mathscr{G}_n \triangle \mathscr{G}] < \infty$; such events exist because $\mathbf{P}$ and $\mathscr{G}$ are Γ-invariant. (Here $\mathscr{H} \triangle \mathscr{K} := (\mathscr{H} \setminus \mathscr{K}) \cup (\mathscr{K} \setminus \mathscr{H})$ is the exclusive or.) Then $\limsup_n \mathscr{G}_n$ is a tail event and equals $\mathscr{G}$(mod 0). □

COROLLARY 3.8. *Let Γ be a transitive closed subgroup of the automorphism group of a graph G. Consider a Γ-invariant insertion-tolerant percolation process $(\mathbf{P}, \omega)$ on G. Then almost surely, the number of infinite clusters of ω is $0, 1$ or ∞.*

The proof is standard for the ergodic components; compare Newman and Schulman (1981).

Benjamini and Schramm (1996) conjectured that for Bernoulli percolation on any quasi-transitive graph, if there are infinitely many infinite clusters, then a.s. every infinite cluster has continuum many ends. This was proved by Häggström and Peres (1999) in the unimodular case and then by Häggström, Peres and Schonmann (1999) in general. For the unimodular case, we give a simpler proof that extends to insertion-tolerant percolation processes. We begin with the following proposition.

PROPOSITION 3.9. *Let G be a graph with a transitive unimodular closed automorphism group $\Gamma \subset \mathrm{Aut}(G)$. Consider some Γ-invariant percolation process $(\mathbf{P}, \omega)$ on G. Then a.s. each infinite cluster that has at least three ends has no isolated ends.*

PROOF. For each $n = 1, 2, \ldots,$ let $\hat{A}_n$ be the union of all vertex sets A that are contained in some percolation cluster $K(A)$, have diameter at most n in the metric of the percolation cluster and such that $K(A) - A$ has at least three infinite components. Note that if ξ is an isolated end of a percolation cluster K, then for each finite n, some neighborhood of ξ in K is disjoint from $\hat{A}_n$. Also observe that if K is a cluster with at least three ends, then K intersects $\hat{A}_n$ for some n.

Fix some $n \geq 1$. Consider the mass transport that sends one unit of mass from each vertex v in a percolation cluster that intersects $\hat{A}_n$ and distributes it equally among the vertices in $\hat{A}_n$ that are closest to v in the metric of the

percolation cluster of v. In other words, let $K(v)$ be the set of vertices in $C(v) \cap \hat{A}_n$, that are closest to v in the metric of ω, and set $F(v,w;\omega) := |K(v)|^{-1}$ if $w \in K(v)$ and otherwise $F(v,w;\omega) := 0$. Then $F(v,w;\omega)$, and hence the expected mass $f(v,w) := \mathbf{E}F(v,w;\omega)$ transported from v to w, is invariant under the diagonal Γ action. If ξ is an isolated end of an infinite cluster K that intersects $\hat{A}_n$, then there is a finite set of vertices B that gets the mass from all the vertices in a neighborhood of ξ. But the mass-transport principle tells us that the expected mass transported to a vertex is finite. Hence, a.s. clusters that intersect $\hat{A}_n$ do not have isolated ends. Since this holds for all n, we gather that a.s. infinite clusters with isolated ends do not intersect $\bigcup_n \hat{A}_n$, whence they have at most two ends. □

PROPOSITION 3.10. *Let G be a graph with a transitive unimodular closed automorphism group $\Gamma \subset \mathrm{Aut}(G)$. If $(\mathbf{P}, \omega)$ is a Γ-invariant insertion-tolerant percolation process on G with infinitely many infinite clusters a.s., then a.s. every infinite cluster has continuum many ends and no isolated end.*

PROOF. As Benjamini and Schramm (1996) noted, it suffices to prove that there are no isolated ends of clusters. To prove this in turn, observe that if some cluster has an isolated end, then because of insertion tolerance, with positive probability, some cluster will have at least three ends with one of them being isolated. Hence Proposition 3.10 follows from Proposition 3.9. □

PROPOSITION 3.11 (Transience). *Let G be a graph with a transitive unimodular closed automorphism group $\Gamma \subset \mathrm{Aut}(G)$. Suppose that $(\mathbf{P}, \omega)$ is a Γ-invariant insertion-tolerant percolation process on G that has almost surely infinitely many infinite clusters. Then a.s. each infinite cluster is transient.*

PROOF. By Proposition 3.10, every infinite cluster of ω has infinitely many ends. Consequently, there is a random forest $\mathcal{F} \subset \omega$ whose distribution is Γ-invariant such that a.s. each infinite cluster C of ω contains a tree of $\mathcal{F}$ with more than two ends [Lemma 7.4 of BLPS (1999)]. From Remark 7.3 of BLPS (1999), we know that any such tree has $p_c < 1$. By Lyons (1990), it follows that such a tree is transient. The Rayleigh monotonicity principle [e.g., Lyons and Peres (1998)] then implies that C is transient. □

REMARK 3.12. Examples show that Proposition 3.11 does not hold when Γ is not unimodular. However, we believe that when Γ is not unimodular and $\mathbf{P}$ is Bernoulli percolation, it does still hold.

Let $(\mathbf{P}, \omega)$ be a bond percolation process on G, and let $\omega \in 2^{\mathsf{E}}$. Let $x \in \mathsf{V}(G)$ be some base point. It will be useful to consider *delayed simple random walk* on ω starting at x, $W = W_x^{\omega}$, defined as follows. Set $W(0) := x$. If $n \geq 0$, conditioned on $\langle W(0), \ldots, W(n) \rangle$ and ω, let $W'(n+1)$ be chosen from the neighbors of $W(n)$ in G with equal probability. Set $W(n+1) :=$

$W'(n+1)$ if the edge $[W(n), W'(n+1)]$ belongs to ω; otherwise, let $W(n+1) := W(n)$.

Given ω, let W and W^* be two independent delayed simple random walks starting at x. Set $w(n) := W(n)$ for $n \geq 0$ and $w(n) := W^*(-n)$ for $n < 0$. Then w is called *two-sided* delayed simple random walk. Let $\hat{\mathbf{P}}_x$ denote the law of the pair (w, ω); it is a probability measure on $\mathsf{V}^{\mathbb{Z}} \times 2^{\mathsf{E}}$. Define $\mathscr{S}$: $\mathsf{V}^{\mathbb{Z}} \to \mathsf{V}^{\mathbb{Z}}$ by

$$(\mathscr{S}w)(n) := w(n+1),$$

and let

$$\mathscr{S}(w, \omega) := (\mathscr{S}w, \omega) \qquad \forall (w, \omega) \in \mathsf{V}^{\mathbb{Z}} \times 2^{\mathsf{E}}.$$

For $\gamma \in \Gamma$, we set

$$\gamma(w, \omega) := (\gamma w, \gamma\omega),$$

where $(\gamma w)(n) := \gamma(w(n))$.

The following lemma generalizes similar lemmas in Häggström (1997), in Häggström and Peres (1999) and in Lyons and Peres (1998).

LEMMA 3.13 (Stationarity of delayed random walk). *Let G be a graph with a transitive unimodular closed automorphism group $\Gamma \subset \mathrm{Aut}(G)$. Let $o \in \mathsf{V}(G)$ be some base point. Let $(\mathbf{P}, \omega)$ be a Γ-invariant bond percolation process on G. Let $\hat{\mathbf{P}}_o$ be the joint law of ω and two-sided delayed simple random walk on ω, as defined above. Then $\hat{\mathbf{P}}_o[\mathscr{A}] = \hat{\mathbf{P}}_o[\mathscr{S}\mathscr{A}]$ for every Γ-invariant $\mathscr{A} \subset \mathsf{V}^{\mathbb{Z}} \times 2^{\mathsf{E}}$. In other words, the restriction of $\hat{\mathbf{P}}_o$ to the Γ-invariant σ-field is $\mathscr{S}$-stationary.*

The lemma will follow from two identities. The first is based on the fact that the transition operator for delayed simple random walk on ω is symmetric, and the second is based on the mass-transport principle, that is, on unimodularity.

PROOF OF LEMMA 3.13. For $j \in \mathbb{Z}$ and $x \in \mathsf{V}$, set

$$\mathscr{W}_x^j := \{(w, \omega) \in \mathsf{V}^{\mathbb{Z}} \times 2^{\mathsf{E}} : w(j) = x\}$$

and

$$\mu := \sum_{x \in \mathsf{V}} \hat{\mathbf{P}}_x.$$

Let $\mathscr{B} \subset \mathsf{V}^{\mathbb{Z}} \times 2^{\mathsf{E}}$ be measurable. Observe that for all $\omega \in 2^{\mathsf{E}}$,

$$\mu[\mathscr{B} \mid \omega] = \mu[\mathscr{S}\mathscr{B} \mid \omega], \tag{3.1}$$

where $\mu[\mathscr{B} \mid \omega]$ means $\sum_{x \in \mathsf{V}} \hat{\mathbf{P}}_x[\mathscr{B} \mid \omega]$. This follows from the fact that for any $j, k \in \mathbb{Z}$ with $j < 0 < k$ and any $v_j, v_{j+1}, \ldots, v_k \in \mathsf{V}$, we have

$$\mu\Big[\mathscr{W}_{v_j}^j \cap \mathscr{W}_{v_{j+1}}^{j+1} \cap \cdots \cap \mathscr{W}_{v_k}^k \mid \omega\Big] = \prod_{i=j}^{k-1} a_i,$$

where $a_i := \deg_G(v_i)^{-1} = \deg_G(o)^{-1}$ if $[v_i, v_{i+1}] \in \omega$, $a_i := (\deg_G(o) - \deg_\omega(v_i))/\deg_G(o)$ if $v_i = v_{i+1}$ and $a_i := 0$ otherwise. By integrating (3.1) over

ω, we obtain

$$\mu[\mathscr{B}] = \mu[\mathscr{S}\mathscr{B}], \tag{3.2}$$

which is our first identity. Observe also that μ is Γ-invariant.

Now let $\mathscr{A} \subset \mathsf{V}^{\mathbb{Z}} \times 2^{\mathsf{E}}$ be Γ-invariant and measurable. For $x, y \in \mathsf{V}$, define

$$\phi(x, y) := \mu\bigl[\mathscr{A} \cap \mathscr{W}_x^{-1} \cap \mathscr{W}_y^0\bigr].$$

Then ϕ is invariant under the diagonal action of Γ on $\mathsf{V} \times \mathsf{V}$, because μ and $\mathscr{A}$ are Γ-invariant. Consequently, the mass-transport principle gives $\sum_{x \in \mathsf{V}} \phi(x, o) = \sum_{x \in \mathsf{V}} \phi(o, x)$, which translates to our second identity,

$$\mu\bigl[\mathscr{A} \cap \mathscr{W}_o^0\bigr] = \mu\bigl[\mathscr{A} \cap \mathscr{W}_o^{-1}\bigr]. \tag{3.3}$$

Observe that $\mu[\mathscr{C} \cap \mathscr{W}_o^0] = \hat{\mathbf{P}}_o[\mathscr{C}]$ for all measurable $\mathscr{C} \subset \mathsf{V}^{\mathbb{Z}} \times 2^{\mathsf{E}}$. By using (3.2) with $\mathscr{B} := \mathscr{A} \cap \mathscr{W}_o^0$ and then using (3.3) with $\mathscr{S}\mathscr{A}$ in place of $\mathscr{A}$, we obtain finally

$$\begin{aligned}\hat{\mathbf{P}}_o[\mathscr{A}] &= \mu\bigl[\mathscr{A} \cap \mathscr{W}_o^0\bigr] = \mu\bigl[\mathscr{S}\bigl(\mathscr{A} \cap \mathscr{W}_o^0\bigr)\bigr] \\ &= \mu\bigl[\mathscr{S}\mathscr{A} \cap \mathscr{W}_o^{-1}\bigr] = \mu\bigl[\mathscr{S}\mathscr{A} \cap \mathscr{W}_o^0\bigr] = \hat{\mathbf{P}}_o[\mathscr{S}\mathscr{A}]. \qquad \square\end{aligned}$$

In Theorem 5.1 below, we show that in an appropriate sense, if ω is ergodic and has indistinguishable components, then the delayed simple random walk on the infinite components of ω is ergodic.

REMARK 3.14 (A generalization of Lemma 3.13). Let V be a countable set acted on by a transitive unimodular group Γ. Let Q be a measurable space, let $\Xi := Q^{\mathsf{V} \times \mathsf{V}}$, and let $\mathbf{P}$ be a Γ-invariant probability measure on Ξ. Suppose that $z\colon Q \to [0, 1]$ is measurable, and that $z(\xi(x, y)) = z(\xi(y, x))$ and $\sum_v z(\xi(x, v)) = 1$ for all $x, y \in \mathsf{V}$ and for $\mathbf{P}$-a.e. ξ. Given $o \in \mathsf{V}$ and a.e. $\xi \in \Xi$, there is an associated random walk starting at o with transition probabilities $p_\xi(x, y) = z(\xi(x, y))$. Let $\hat{\mathbf{P}}_o$ denote the joint distribution of ξ and this random walk. Then the above proof shows that the restriction of $\hat{\mathbf{P}}_o$ to the Γ-invariant σ-field is $\mathscr{S}$-invariant.

See Lyons and Schramm (1999) for a still greater generalization.

Let $(\mathbf{P}, \omega)$ be a bond percolation process on an infinite graph G. For every $e \in \mathsf{E}(G)$, let $\mathscr{F}_{\neg e}$ be the σ-field generated by the events $\{e' \in \omega\}$ with $e' \neq e$. Set $Z(e) := \mathbf{P}[e \in \omega \mid \mathscr{F}_{\neg e}]$, and call this the *conditional marginal* of e. Note that insertion tolerance is equivalent to $Z(e) > 0$ a.s. for every $e \in \mathsf{E}(G)$.

PROOF OF THEOREM 3.3. Let $(\mathbf{P}, \omega)$ be insertion tolerant and Γ-invariant. Let $o \in \mathsf{V}$ be some fixed base point of G. Assume that the theorem is false. Then by Corollary 3.8, there is positive probability that ω has infinitely many infinite clusters. If we condition on this event, then ω is still insertion tolerant, as shown in Lemma 3.6. Consequently, we henceforth assume, with no loss of generality, that a.s. ω has infinitely many infinite clusters.

Fix $\varepsilon > 0$. From Lemma 3.5, we know that there is a positive probability for pivotal edges of clusters of type $\mathscr{A}$, or there is a positive probability for

pivotal edges of clusters of type $\neg\mathscr{A}$. Since we may replace $\mathscr{A}$ by its complement, assume, with no loss of generality, that there is a positive probability for pivotal edges of clusters of type $\mathscr{A}$. Fix some $r > 0$ and $\delta > 0$ such that with positive probability, $C(o)$ is infinite of type $\mathscr{A}$ and there is an edge e at distance r from o that is pivotal for $(C(o), \omega)$ and satisfies $Z(e) \geq \delta$, where $Z(e)$ is the conditional marginal of e, as described above.

Let $\mathscr{A}_o$ be the event that $C(o)$ is infinite and $(C(o), \omega) \in \mathscr{A}$, and let $\mathscr{A}_o'$ be an event that depends on only finitely many edges such that $\mathbf{P}[\mathscr{A}_o \triangle \mathscr{A}_o'] < \varepsilon$. Let R be large enough that $\mathscr{A}_o'$ depends only on edges in the ball $B(o, R)$.

Let $W\colon \mathbb{Z} \to \mathsf{V}$ be two-sided delayed simple random walk on ω, with $W(0) = o$. For $n \in \mathbb{Z}$, let $e_n \in \mathsf{E}$ be an edge chosen uniformly among the edges at distance r from $W(n)$. Write $\hat{\mathbf{P}}$ for the probability measure where we choose ω according to $\mathbf{P}$, choose W, and choose $\langle e_n\colon n \in \mathbb{Z}\rangle$ as indicated.

Given any $e \in \mathsf{E}$, let $\mathscr{P}_e$ be the event that $\omega \in \mathscr{A}_o$, that e is pivotal for $C(o)$, and that $Z(e) \geq \delta$. Let $\mathscr{E}_e^n$ be the event that $e_n = e$ and $W(j)$ is not an endpoint of e whenever $-\infty < j < n$. Note that for all $\zeta \in 2^{\mathsf{E}}$, $n \in \mathbb{Z}$, and $e \in \mathsf{E}$,

$$\hat{\mathbf{P}}\big[\mathscr{E}_e^n \mid \omega = \zeta\big] = \hat{\mathbf{P}}\big[\mathscr{E}_e^n \mid \omega = \Pi_e \zeta\big].$$

Thus, for all measurable $\mathscr{B} \subset 2^{\mathsf{E}}$ with $\mathbf{P}[\mathscr{B}] > 0$ and $\mathbf{P}[Z(e) \geq \delta \mid \mathscr{B}] = 1$, we have

$$\begin{aligned}\hat{\mathbf{P}}\big[\mathscr{E}_e^n \cap \Pi_e\mathscr{B}\big] &= \hat{\mathbf{P}}\big[\mathscr{E}_e^n \mid \Pi_e\mathscr{B}\big]\mathbf{P}\big[\Pi_e\mathscr{B}\big] = \hat{\mathbf{P}}\big[\mathscr{E}_e^n \mid \mathscr{B}\big]\mathbf{P}\big[\Pi_e\mathscr{B}\big] \\ &= \frac{\mathbf{P}\big[\Pi_e\mathscr{B}\big]}{\mathbf{P}\big[\mathscr{B}\big]}\hat{\mathbf{P}}\big[\mathscr{E}_e^n \cap \mathscr{B}\big] \geq \delta\hat{\mathbf{P}}\big[\mathscr{E}_e^n \cap \mathscr{B}\big].\end{aligned}$$

In particular,

$$\hat{\mathbf{P}}\big[\mathscr{E}_e^n \cap \Pi_e\mathscr{A}_o' \cap \Pi_e\mathscr{P}_e\big] \geq \delta\hat{\mathbf{P}}\big[\mathscr{E}_e^n \cap \mathscr{A}_o' \cap \mathscr{P}_e\big] = \delta\hat{\mathbf{P}}\Big[\mathscr{E}_e^n \cap \mathscr{A}_o' \cap \mathscr{P}_{e_n}\Big]. \tag{3.4}$$

Let

$$\begin{aligned}\mathscr{E}^n &:= \bigcup_{e \in \mathsf{E}} \mathscr{E}_e^n, \\ \mathscr{E}_R^n &:= \bigcup_{e \in \mathsf{E} - B(o,\,R)} \mathscr{E}_e^n,\end{aligned}$$

and note that these are disjoint unions. Since $\Pi_e\mathscr{P}_e \subset \neg\mathscr{A}_o$ and since $\Pi_e\mathscr{A}_o' \subset \mathscr{A}_o'$ when $e \notin B(o, R)$, we may sum (3.4) over all $e \notin B(o, R)$ to obtain that

$$\begin{aligned}\mathbf{P}\big[\mathscr{A}_o' \cap \neg\mathscr{A}_o\big] &\geq \hat{\mathbf{P}}\big[\mathscr{E}_R^n \cap A_o' \cap \neg\mathscr{A}_o\big] \geq \delta\hat{\mathbf{P}}\Big[\mathscr{E}_R^n \cap \mathscr{A}_o' \cap \mathscr{P}_{e_n}\Big] \\ &\geq \delta\hat{\mathbf{P}}\Big[\mathscr{E}_R^n \cap \mathscr{A}_o \cap \mathscr{P}_{e_n}\Big] - \delta\varepsilon.\end{aligned} \tag{3.5}$$

Fix n to be sufficiently large that the probability that $C(o)$ is infinite and $e_n \in B(o, R)$ is smaller than ε; this can be done by Proposition 3.11. Then $\hat{\mathbf{P}}[\mathscr{A}_o \cap \mathscr{E}^n - \mathscr{E}_R^n] \leq \varepsilon$. Hence, we have from (3.5) that

$$\varepsilon \geq \mathbf{P}\big[\mathscr{A}_o' \triangle \mathscr{A}_o\big] \geq \mathbf{P}\big[\mathscr{A}_o' \cap \neg\mathscr{A}_o\big] \geq \delta\hat{\mathbf{P}}\Big[\mathscr{E}^n \cap \mathscr{A}_o \cap \mathscr{P}_{e_n}\Big] - 2\varepsilon\delta. \tag{3.6}$$

Recall that $\hat{\mathbf{P}}[\mathscr{A}_o \cap \mathscr{P}_{e_0}] > 0$. Moreover, conditioned on $\mathscr{A}_o \cap \mathscr{P}_{e_0}$, transience guarantees a.s. a least $m \in \mathbb{Z}$ such that $W(m)$ is at distance r to e_0. Consequently, for some $m \le 0$,

$$\hat{\mathbf{P}}\big[\mathscr{E}^m \cap \mathscr{A}_o \cap \mathscr{P}_{e_m}\big] > 0. \tag{3.7}$$

Let $\mathscr{B}_m$ be the event that $Z(e_m) \ge \delta$, that e_m is not an endpoint of $W(j)$ for $j < m$, that $C(W(m))$ is infinite and of type $\mathscr{A}$, and that e_m is pivotal for $C(W(m))$. Then $\mathscr{B}_m = \mathscr{E}^m \cap \mathscr{A}_o \cap \mathscr{P}_{e_m}$. But $\mathscr{B}_m$ is Γ-invariant. Therefore, Lemma 3.13 shows that the left-hand side of (3.7) does not depend on m, and it certainly does not depend on ε. Hence, when we take ε to be a sufficiently small positive number, (3.6) gives a contradiction. This completes the proof of the theorem. □

EXAMPLE 3.15. A deletion-tolerant process that does not have indistinguishable clusters is obtained as follows. Let X be a 3-regular tree and $p \in (1/2, 1)$. Let $p' \in (1/(2p), 1)$. Begin with Bernoulli(p) percolation on X. Independently for each cluster C, with probability $1/2$, intersect it with an independent Bernoulli(p') percolation. The resulting percolation process is clearly deletion tolerant, yet some infinite clusters ω have $p_c(\omega) = 1/(2p)$, while others have $p_c(\omega) = 1/(2pp')$.

REMARK 3.16. Let T be the 3-regular tree, let Γ be the subgroup of automorphisms of T that fixes an end ξ of T and let $\mathbf{P}$ be Bernoulli(p) bond percolation on T, where $p \in (1/2, 1)$. Then a.s. each infinite cluster C has a unique vertex v_C "closest" to ξ. The degree of v_C in the percolation configuration distinguishes among the infinite clusters. Hence, Theorem 3.3 does not hold without the assumption that Γ is unimodular. By a simple modification, a similar example can be constructed where Γ is the full automorphism group of a graph on which the percolation is performed [compare BLPS (1999)].

However, Häggström, Peres and Schonmann (1999) have recently shown that even without the unimodularity assumption, when $p > p_c$ so-called "robust" properties do not distinguish between the infinite clusters of Bernoulli(p) percolation.

QUESTION 3.17. In the case that Γ is nonunimodular, write μ_x for the Haar measure of the stabilizer of $x \in \mathsf{V}$. The infinite clusters C divide into two types: the *heavy* clusters for which $\sum_{x \in C} \mu_x = \infty$ and the others, the *light* clusters. It can be that the light clusters are distinguishable: for example, consider a 3-regular tree T with a fixed end ξ. Let Γ be the group of automorphisms of T that fix ξ. Then Bernoulli(2/3) percolation has infinitely many light clusters, which can be distinguished by the degree of the vertex they contain that is closest to ξ. But is it the case that heavy clusters are indistinguishable for every insertion-tolerant percolation process that is invariant with respect to a transitive automorphism group?

4. Uniqueness and connectivity. Our goal is to prove the following theorem.

THEOREM 4.1 (Uniqueness and connectivity). *Let G be an infinite graph with a transitive unimodular closed automorphism group $\Gamma \subset \mathrm{Aut}(G)$. Let $\mathbf{P}$ be a Γ-invariant and ergodic insertion-tolerant percolation process on G. If $\mathbf{P}$ has more than one infinite component a.s., then connectivity decays,*

$$\inf\{\tau(x, y)\colon x, y \in \mathsf{V}\} = 0.$$

The intuitive idea behind our proof is that if $\inf \tau(x, y) > 0$, then each infinite cluster has a positive "density." Since the densities are the same by cluster indistinguishability, there are only finitely many infinite clusters. By Corollary 3.8, there is only one. To make the idea of "density" precise, we use simple random walk X on the whole of G with the percolation subgraph as the scenery, counting how many times we visit each cluster.

For a set $C \subset \mathsf{V}$, write

$$\alpha(C) := \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\{X(k)\in C\}},$$

where the limit exists, for the frequency of visits to C by the simple random walk on G.

LEMMA 4.2 (Cluster frequencies). *Let G be a graph with a transitive unimodular closed automorphism group $\Gamma \subset \mathrm{Aut}(G)$. There is a Γ-invariant measurable function* $\mathrm{freq}\colon 2^{\mathsf{V}} \to [0,1]$ *with the following property. Suppose that $(\mathbf{P}, \omega)$ is a Γ-invariant bond percolation process on G, and let $\hat{\mathbf{P}} := \mathbf{P} \times \mathbf{P}_o$, where $\mathbf{P}_o$ is the law of simple random walk on G starting at the base point o. Then $\hat{\mathbf{P}}$-a.s. $\alpha(C) = \mathrm{freq}(C)$ for every cluster C.*

PROOF. Given a set $C \subset \mathsf{V}$ and $m, n \in \mathbb{Z}$, $m < n$, let

$$\alpha_m^n(C) := \frac{1}{n-m} \sum_{k=m}^{n-1} \mathbf{1}_{\{X(k)\in C\}}.$$

For every $\alpha \in [0,1]$, let

$$\mathscr{Z}_\alpha := \left\{C \subset \mathsf{V}\colon \lim_{n\to\infty} \alpha_0^n(C) = \alpha \ \mathbf{P}_o\text{-a.s.}\right\},$$

and set $\mathscr{Z} := \bigcup_{\alpha\in[0,1]} \mathscr{Z}_\alpha$. Define $\mathrm{freq}(C) := \alpha$ when $C \in \mathscr{Z}_\alpha$. If $C \notin \mathscr{Z}$, put $\mathrm{freq}(C) := 0$, say. It is easy to verify that freq is measurable. Let $\gamma \in \Gamma$. To prove that a.s. $\mathrm{freq}(C) = \mathrm{freq}(\gamma C)$, note that there is an $m \in \mathbb{N}$ such that with positive probability $X(m) = \gamma o$. Hence for every measurable $A \subset [0,1]$ such that $\alpha(C) \in A$ with positive probability, we have $\alpha(\gamma C) \in A$ with positive probability. This implies that freq is Γ-invariant.

It remains to prove that $\mathbf{P}$-a.s., every component of ω is in $\mathscr{Z}$. First observe that the restriction of $\hat{\mathbf{P}}$ to the Γ-invariant σ-field is $\mathscr{S}$-invariant. This can be

verified directly, but is also a special case of Remark 3.14: take $Q = \{-1, 0, 1\}$, take the value in Q associated to a pair $(x, y) \in \mathsf{V} \times \mathsf{V}$ to be 1 if $[x, y] \in \omega$, 0 if $[x, y] \in \mathsf{E} - \omega$ and -1 if $[x, y] \notin \mathsf{E}$ and take $z(0) = z(1) = 1/\deg_G(o)$ and $z(-1) = 0$.

The following argument is modeled on the proof of Theorem 1 of Burton and Keane (1989). Let $F^n(j)$ be the number of times that the j most frequently visited clusters in $[0, n-1]$ are visited in $[0, n-1]$. That is,

$$F^n(j) := n \max\bigl\{\alpha_0^n(C_1) + \cdots + \alpha_0^n(C_j)\colon C_1, \ldots, C_j \text{ are distinct clusters}\bigr\}.$$

For each fixed j, it is easy to see that $F^n(j)$ is a subadditive sequence, that is, $F^{n+k}(j) \le F^n(j) + \mathscr{S}^n F^k(j)$. Note that the random variables $F^n(j)$ are invariant with respect to the diagonal action of Γ on $\mathsf{V}^{\mathbb{Z}} \times 2^{\mathsf{E}}$. By the subadditive ergodic theorem, $\lim_n F^n(j)/n$ exists a.s. Set $\alpha(j) := \lim_n F^n(j)/n - \lim_n F^n(j-1)/n$ for $j \ge 1$. We claim that a.s.,

$$\lim_{n\to\infty} \max\bigl\{|\alpha_0^m(C) - \alpha_0^k(C)|\colon k, m \in \mathbb{Z}, k, m \ge n, C \text{ is a cluster}\bigr\} = 0 \tag{4.1}$$

the Cauchy property uniform in C. Indeed, let $\varepsilon > 0$. Observe that $\sum_j \alpha(j) \le 1$ and that $\alpha(1) \ge \alpha(2) \ge \cdots$. Let $j_1 \ge 1$ be large enough that $\alpha(j) < \varepsilon/9$ for all $j \ge j_1$. Let m_1 be sufficiently large that

$$|F^m(j)/m - F^m(j-1)/m - \alpha(j)| < \varepsilon/(9j_1 + 9)$$

for all $j \le j_1$ and all $m \ge m_1$. Set

$$U := \{x \in [0,1]\colon \exists\, j \le j_1\ |x - \alpha(j)| < \varepsilon/(9j_1 + 9)\} \cup [0, \varepsilon/3],$$

and let $U(\delta)$ denote the set of points $x \in \mathbb{R}$ within distance δ of U. For all clusters C that are visited by the random walk and all $m = 1, 2, \ldots$, there is some j such that $\alpha_0^m(C) = F^m(j)/m - F^m(j-1)/m$. If $j \le j_1$ and $m \ge m_1$, it follows that $\alpha_0^m(C) \in U$. The same also holds when $j \ge j_1$ and $m \ge m_1$, because the j_1th most frequently visited cluster in $[0, m-1]$, say C', satisfies $\alpha_0^m(C) \le \alpha_0^m(C') \le \alpha(j_1) + \varepsilon/9 \le \varepsilon/3$. Hence $\alpha_0^m(C) \in U$ for all clusters C and all $m \ge m_1$. Because $-1/m \le \alpha_0^m(C) - \alpha_0^{m+1}(C) \le 1/m$, it follows that for all C and all $n \ge m_1$, the set $\{\alpha_0^m(C)\colon m \ge n\}$ is contained in some connected component of $U(1/n)$. But when $n > \max\{m_1, 9(j_1+1)/\varepsilon\}$, the total length of $U(1/n)$ is less than ε, which implies that the diameter of each connected component is less than ε. This verifies (4.1).

Since

$$\begin{aligned} &2\max\bigl\{|\alpha_0^{2n}(C) - \alpha_0^n(C)|\colon C \text{ is a cluster}\bigr\} \\ &\quad = \max\bigl\{|\alpha_n^{2n}(C) - \alpha_0^n(C)|\colon C \text{ is a cluster}\bigr\} \end{aligned}$$

has the same law as $\max\{|\alpha_0^n(C) - \alpha_{-n}^0(C)|\colon C \text{ is a cluster}\}$ (by the $\mathscr{S}$-invariance noted above), it follows from (4.1) that a.s. $\lim_{n\to\infty} \alpha_0^n(C) = \lim_{n\to\infty} \alpha_{-n}^0(C)$ for every cluster C. When the cluster C is fixed, $\alpha_0^n(C)$ and $\alpha_{-n}^0(C)$ are independent, but both tend to $\alpha(C)$. Hence $\alpha(C)$ is an a.s. constant, which means that $\mathbf{P}$-a.s. we have $C \in \mathscr{Z}$ for every cluster C. This completes the proof. □

PROOF OF THEOREM 4.1. Let freq be as in Lemma 4.2. Since $\mathbf{P}$ is ergodic, Theorem 3.3 implies that there is a constant $c \in [0, 1]$ such that a.s. freq(C)

$= c$ for every infinite cluster C. Suppose that there is more than one infinite cluster with positive probability. Then there are infinitely many infinite clusters a.s. Since clearly $\sum_C \alpha(C) = \sum_C \mathrm{freq}(C) \le 1$, where the sum is over all clusters, it follows that $c = 0$. Since G is infinite, it is also immediate that $\mathrm{freq}(\{v\}) = 0$ for every $v \in \mathsf{V}$. Therefore, $\mathrm{freq}(C) = 0$ a.s. for all clusters, finite or infinite. In particular, $\mathrm{freq}(C(o)) = 0$ a.s., where $C(o)$ is the cluster of o.

Let $\tau_0 := \inf\{\tau(x, y)\colon x, y \in \mathsf{V}\}$. We have that

$$\mathbf{E}\big[\,\alpha_0^n(C(o))\big] \ge \tau_0,$$

whence $0 = \mathbf{E}[\mathrm{freq}(C(o))] = \mathbf{E}[\,\alpha(C(o))] \ge \tau_0$ by the bounded convergence theorem, as required. □

QUESTION 4.3. The proof actually shows that $(1/n)\sum_{k=0}^{n-1}\mathbf{E}[\tau(o, X(k))] \to 0$ as $n \to \infty$ when there are infinitely many infinite clusters in an invariant percolation process that has indistinguishable clusters. Do we have $\tau(o, X(n)) \to 0$ a.s.?

EXAMPLE 4.4. We give an example of an ergodic invariant deletion-tolerant percolation process with infinitely many infinite clusters and with τ bounded below. The percolation process ω will take place on $\mathbb{Z}^3$. Let $\{a(n)\colon n \in \mathbb{Z}\}$ be independent $\{0, 1\}$-valued random variables with $\mathbf{P}[a(n) = 1] = 1/2$, and let A be the set of vertices $(x_1, x_2, x_3) \in \mathbb{Z}^3$ with $a(x_1) = 1$. Let $\varepsilon > 0$ be small, and let z_e be i.i.d. uniformly distributed in $[0, 1]$ indexed by the edges of $\mathbb{Z}^3$ and independent of the $a(n)$'s. Let ω_1 be the set of edges e with both endpoints in A such that $z_e < 1 - \varepsilon$, and let ω_2 be the set of edges e with $z_e < \varepsilon$. Let ω_3 be the set of edges that have no vertex in common with $\omega_2 \cup A$ and satisfy $z_e < 1 - \varepsilon$. Set $\omega := \omega_1 \cup \omega_2 \cup \omega_3$.

It is immediate to verify that when ε is sufficiently small, a.s. ω has a single infinite component whose intersection with each component of A is infinite, and has a single infinite component in each component of the complement of A. The claimed properties of this example follow easily.

5. Ergodicity of delayed random walk. The following theorem is not needed for the rest of the paper. It is presented here because its proof uses some of the ideas from the proof of Lemma 4.2.

THEOREM 5.1 (Ergodicity of delayed random walk). *Let G be a graph with a transitive unimodular closed automorphism group $\Gamma \subset \mathrm{Aut}(G)$. Let $o \in \mathsf{V}(G)$ be some base point. Let $(\mathbf{P}, \omega)$ be a Γ-invariant ergodic bond percolation process on G with infinite clusters a.s. Also suppose that ω has indistinguishable infinite clusters a.s. Let $\hat{\mathbf{P}}_o$ be the joint law of ω and two-sided delayed simple random walk on ω, starting at o. Let $\mathscr{A}$ be an event that is Γ-invariant and $\mathscr{S}$-invariant. Then $\hat{\mathbf{P}}_o[\mathscr{A} \mid C(o)$ is infinite$]$ is either 0 or 1.*

PROOF. For each $m, n \in \mathbb{Z}$, let $\mathscr{F}_m^n$ be the σ-field of Γ-invariant sets generated by ω and the random variables $\langle W(j)\colon j \in [m, n]\rangle$, where $W(j)$ is the location of the delayed random walk at time j.

Let $\mathscr{C}$ be the event that $C(o)$ is infinite. Let $\varepsilon > 0$. Then there is an $n \in \mathbb{N}$ and an event $\mathscr{A}' \in \mathscr{F}_{-n}^{n}$ such that $\hat{\mathbf{P}}_o[\mathscr{A}' \triangle \mathscr{A}] < \varepsilon \mathbf{P}[\mathscr{C}]$, and therefore

$$\hat{\mathbf{P}}_o[\mathscr{A}' \triangle \mathscr{A} \mid \mathscr{C}] < \varepsilon. \tag{5.1}$$

Note that $\mathscr{S}\mathscr{C} = \mathscr{C}$. Hence we have for all $m \in \mathbb{Z}$,

$$\hat{\mathbf{P}}_o[\mathscr{S}^m \mathscr{A}' \triangle \mathscr{A} \mid \mathscr{C}] < \varepsilon.$$

Observe also that the two events $\mathscr{S}^{n+1}\mathscr{A}'$ and $\mathscr{S}^{-n-1}\mathscr{A}'$ are independent given ω by the Markov property for the delayed random walk. Consequently,

$$\begin{aligned}
\int_{\omega \in \mathscr{C}} \hat{\mathbf{P}}_o[\mathscr{A} \mid \omega]^2 \, d\mathbf{P}[\omega] &\geq \int_{\omega \in \mathscr{C}} \hat{\mathbf{P}}_o[\mathscr{S}^{n+1}\mathscr{A}' \mid \omega] \hat{\mathbf{P}}_o[\mathscr{S}^{-n-1}\mathscr{A}' \mid \omega] \, d\mathbf{P}[\omega] - 2\varepsilon \\
&= \int_{\omega \in \mathscr{C}} \hat{\mathbf{P}}_o[\mathscr{S}^{n+1}\mathscr{A}', \mathscr{S}^{-n-1}\mathscr{A}' \mid \omega] \, d\mathbf{P}[\omega] - 2\varepsilon \\
&\geq \int_{\omega \in \mathscr{C}} \hat{\mathbf{P}}_o[\mathscr{A} \mid \omega] \, d\mathbf{P}[\omega] - 4\varepsilon.
\end{aligned}$$

Since ε is arbitrary, it follows that

$$\int_{\mathscr{C}} \hat{\mathbf{P}}_o[\mathscr{A} \mid \omega]^2 \, d\mathbf{P}[\omega] \geq \int_{\mathscr{C}} \hat{\mathbf{P}}_o[\mathscr{A} \mid \omega] \, d\mathbf{P}[\omega],$$

which means that $\hat{\mathbf{P}}_o[\mathscr{A} \mid \omega]$ is 0 or 1 for almost every $\omega \in \mathscr{C}$.

Let $\mathscr{B}$ be the set of $\omega \in \mathscr{C}$ such that $\hat{\mathbf{P}}_o[\mathscr{A} \mid \omega] = 1$ and let

$$\hat{\mathscr{B}} := \{(C(o), \omega) \colon \omega \in \mathscr{B}\}.$$

Finally, let $\Gamma\hat{\mathscr{B}}$ denote the orbit of $\hat{\mathscr{B}}$ under Γ; that is, $\Gamma\hat{\mathscr{B}} := \{(C(\gamma o), \gamma\omega) \colon \omega \in \mathscr{B}\}$. Because $\mathscr{S}\mathscr{A} = \mathscr{A}$ and $\mathscr{A}$ is Γ-invariant, it follows as in the beginning of the proof of Lemma 4.2 that for every $\gamma \in \Gamma$ and every ω such that $\gamma o \in C(o)$, we have $\omega \in \mathscr{B}$ iff $\gamma^{-1}\omega \in \mathscr{B}$ (after possibly making a measure zero modification of $\mathscr{B}$). Therefore,

$$\Gamma\hat{\mathscr{B}} \cap \{(C, \omega) \colon o \in C\} = \hat{\mathscr{B}}. \tag{5.2}$$

Since ω is ergodic and has indistinguishable infinite clusters, a.s. all infinite clusters of ω are in $\Gamma\hat{\mathscr{B}}$ or a.s. all infinite clusters of ω are not in $\Gamma\hat{\mathscr{B}}$. If the latter is the case, then $\mathbf{P}[\mathscr{B}] = 0$ and hence $\hat{\mathbf{P}}_o[\mathscr{A} \mid \mathscr{C}] = 0$. Therefore, assume that

$$\mathbf{P}\big[(C(o), \omega) \in \Gamma\hat{\mathscr{B}} \mid \mathscr{C}\big] = 1.$$

Hence, by (5.2), $\mathbf{P}[\mathscr{B} \mid \mathscr{C}] = 1$, giving $\hat{\mathbf{P}}_o[\mathscr{A} \mid \mathscr{C}] = 1$, as required. □

6. Uniqueness for Bernoulli percolation. We now present some applications of Theorem 4.1 to Bernoulli percolation on Cayley graphs. Later in this section, we prove inequalities relating p_u based on Theorem 3.3.

We first note that the choice of generators does not influence whether $p_u < 1$.

THEOREM 6.1. *Let S_1 and S_2 be two finite generating sets for a countable group Γ, yielding corresponding Cayley graphs G_1 and G_2. Then $p_u(G_1) < 1$ iff $p_u(G_2) < 1$.*

PROOF. Left and right Cayley graphs with respect to a given set of generators are isomorphic via $x \mapsto x^{-1}$, so we consider only right Cayley graphs.

Write $\tau_p^i(x)$ for the probability under Bernoulli(p) percolation that o and x lie in the same cluster of G_i. Express each element $s \in S_1$ in terms of a word $\varphi(s) \in S_2$. Let ω_2 be Bernoulli(p) percolation on G_2 and define ω_1 on G_1 by letting $[x, xs] \in \omega_1$ iff the path from x to xs in G_2 given by $\varphi(s)$ lies in ω_2. Then ω_1 is a percolation such that if two edges are sufficiently far apart, then their presence in ω_1 is independent. Thus, Liggett, Schonmann and Stacey (1997) provide a function $f(p) \in (0,1)$ such that $f(p) \uparrow 1$ when $p \uparrow 1$ and such that ω_1 stochastically dominates Bernoulli($f(p)$) percolation on G_1. This implies that $\tau_{f(p)}^1(x) \le \tau_p^2(x)$ for each x. If $f(p) > p_u(G_1)$, then it follows from this and Theorem 4.1 that $p \ge p_u(G_2)$, showing that $p_u(G_1) < 1$ implies $p_u(G_2) < 1$. □

REMARK 6.2. For the situation used in the above proof, and for many similar applications, one does not need the full generality of the theorem of Liggett, Schonmann and Stacey. The following observation suffices. Suppose that $(X_i\colon i \in I)$ are i.i.d. random variables taking values in $\{0,1\}$ and with $\mathbf{P}[X_i = 1] = p$. Let $(I_j\colon j \in J)$ be an indexed collection of subsets of I. Set $Y_j := \min\{X_i\colon i \in I_j\}$ and $J_i := \{j \in J\colon i \in I_j\}$. Suppose that $n := \sup\{|I_j|\colon j \in J\}$ and $m := \sup\{|J_i|\colon i \in I\}$ are both finite. Then $(Y_j\colon j \in J)$ stochastically dominates independent random variables $(Z_j\colon j \in J)$ with $\mathbf{P}[Z_j = 1] \ge (1 - (1-p)^{1/m})^n$. Indeed, set $X_i' := \max\{Z_{i,j}\colon j \in J_i\}$ and $Z_j := \min\{Z_{i,j}\colon i \in I_j\}$, where $(Z_{i,j}\colon i \in I, j \in J)$ are independent $\{0,1\}$-variables with $\mathbf{P}[Z_{i,j} = 1] = 1 - (1-p)^{1/m}$. Then $(X_i\colon i \in I)$ stochastically dominates $(X_i'\colon i \in I)$, and hence $(Y_j\colon j \in J)$ stochastically dominates $(Z_j\colon j \in J)$.

REMARK 6.3. Schonmann (1999a) shows that for every quasi-transitive graph G, $p_u = p_{BB}$, where p_{BB} is the infimum of all p such that

$$\lim_{r \to \infty} \inf_{v,u \in V} \mathbf{P}\big[B(v,r) \leftrightarrow B(u,r)\big] = 1$$

in Bernoulli(p) percolation, where $B(v,r)$ denotes the ball in G of radius r and center v and $A \leftrightarrow A'$ is the event that there is a cluster C with $C \cap A \ne \varnothing$ and $C \cap A' \ne \varnothing$. Based on this and the proof of Theorem 6.1, one obtains the following generalization. Suppose that G and G' are quasi-transitive graphs and G' is quasi-isometric to G. Then $p_u(G) < 1$ iff $p_u(G') < 1$. This observation was also made independently by Y. Peres (private communication).

For our next result regarding p_u, we need the following construction. Let κ_0 be a real-valued random variable and suppose that $\mathbf{P}$ is a bond percolation

process on some graph G. We would like to color the clusters of $\mathbf{P}$ in such a way that conditioned on the configuration ω, the colors of the components are i.i.d. random variables with the same law as κ_0.

To construct this process, let $\langle v_1, v_2, \dots \rangle$ be an ordering of the vertices in G. Let $\langle \kappa_1, \kappa_2, \dots \rangle$ be i.i.d. random variables with the same law as κ_0. Given ω and $v \in \mathsf{V}$, set $\kappa_\omega(v) := \kappa_j$ if j is the least integer with $v_j \in C(v)$. It is not hard to see that $\kappa_\omega(v)$ is measurable, as a function on the product of 2^{E} and the sample spaces of the κ_j's. Let $\mathbf{P}^\kappa$ denote the law of (ω, κ_ω). Observe that $\mathbf{P}^\kappa$ satisfies the description of the previous paragraph, and therefore does not depend on the choice of the ordering of V. Consequently, if γ is an automorphism of G and $\mathbf{P}$ is γ-invariant, then $\mathbf{P}^\kappa$ is γ-invariant.

LEMMA 6.4 (Ergodicity of $\mathbf{P}^\kappa$). *Let Γ be a transitive closed subgroup of the automorphism group of a graph G. Suppose that $(\mathbf{P}, \omega)$ is a Γ-invariant ergodic insertion-tolerant percolation process on G. Let κ_0 be a real-valued random variable, and let $\mathbf{P}^\kappa$ be as above. Suppose that* $\inf\{\tau(o, x)\colon x \in \mathsf{V}(G)\} = 0$. *Then $\mathbf{P}^\kappa$ is Γ-invariant and ergodic.*

PROOF. To prove the ergodicity of $\mathbf{P}^\kappa$, let $\mathscr{A}$ be a Γ-invariant event in $2^{\mathsf{E}(G)} \times \mathbb{R}^{\mathsf{V}(G)}$ and $\varepsilon \in (0, 1/2)$. The probability of $\mathscr{A}$ conditioned on ω must be a constant, say a, by ergodicity of $\mathbf{P}$. We need to show that $a \in \{0, 1\}$.

There is a cylindrical event $\mathscr{A}'$ that depends only on the restriction of ω and κ_ω to some ball $B(o, r)$ about o and such that $\mathbf{P}^\kappa[\mathscr{A} \triangle \mathscr{A}'] < \varepsilon$. Then $\mathbf{E}[|\mathbf{P}^\kappa[\mathscr{A}' \mid \omega] - a|] < \varepsilon$.

Let $\mathscr{B}_x$ be the event that some vertex in $B(o, r)$ belongs to the same cluster as some vertex in $B(x, r)$. Since $\inf_x \tau(o, x) = 0$ and $\mathbf{P}$ is insertion tolerant, there is some x such that $\mathbf{P}[\mathscr{B}_x] < \varepsilon$. Indeed, suppose that $\mathbf{P}[\mathscr{B}_x] \geq \varepsilon$ for all x. Let $\mathscr{D}_x$ be the event that all the edges in $B(x, r)$ belong to ω, and for $A \subset G$, let $\mathscr{F}_{\neg A}$ denote the σ-field generated by the events $\{e \in \omega\}$ with $e \notin A$. Then $\mathscr{B}_x$ is $\mathscr{F}_{\neg (B(o,r) \cup B(x,r))}$-measurable. By insertion tolerance, for every $\mathscr{F}_{\neg B(o,r)}$-measurable event $\mathscr{C}$ with $\mathbf{P}[\mathscr{C}] > 0$, we have $\mathbf{P}[\mathscr{C} \cap \mathscr{D}_o] > 0$. Consequently, there is some $\delta > 0$ such that

$$\mathbf{P}\bigl[\mathbf{P}\bigl[\mathscr{D}_o \mid \mathscr{F}_{\neg B(o,r)}\bigr] > \delta\bigr] > 1 - \varepsilon/2.$$

It follows that for all $\mathscr{F}_{\neg B(o,r)}$-measurable events $\mathscr{C}$ with $\mathbf{P}[\mathscr{C}] \geq \varepsilon$, we have $\mathbf{P}[\mathscr{D}_o \mid \mathscr{C}] > \delta/2$. In particular, for all $x \in \mathsf{V}$, we have $\mathbf{P}[\mathscr{D}_o \mid \mathscr{B}_x] > \delta/2$, which gives $\mathbf{P}[\mathscr{D}_o \cap \mathscr{B}_x] > \varepsilon\delta/2$. Transitivity and insertion tolerance imply that there is a $\delta' > 0$ such that

$$\mathbf{P}\bigl[\mathbf{P}\bigl[\mathscr{D}_x \mid \mathscr{F}_{\neg B(x,r)}\bigr] > \delta'\bigr] > 1 - \varepsilon\delta/4$$

for all $x \in \mathsf{V}$. Hence, for all $\mathscr{F}_{\neg B(x,r)}$-measurable events $\mathscr{C}$ with $\mathbf{P}[\mathscr{C}] \geq \varepsilon\delta/2$, we have $\mathbf{P}[\mathscr{D}_x \mid \mathscr{C}] > \delta'/2$. Taking $\mathscr{C} := \mathscr{D}_o \cap \mathscr{B}_x$ gives for all $x \in \mathsf{V} - B(o, 2r)$,

$$\tau(o, x) \geq \mathbf{P}[\mathscr{D}_o \cap \mathscr{B}_x \cap \mathscr{D}_x] = \mathbf{P}[\mathscr{D}_x \mid \mathscr{D}_o \cap \mathscr{B}_x]\mathbf{P}[\mathscr{D}_o \cap \mathscr{B}_x] \geq (\delta'/2)(\varepsilon\delta/2),$$

which contradicts our assumption, and thereby verifies that $\mathbf{P}[\mathscr{B}_x] < \varepsilon$ for some $x \in \mathsf{V}$. Fix such an x.

Let $\gamma_x \in \Gamma$ be such that $\gamma_x o = x$. Then

$$\mathbf{E}\big[\,|\mathbf{P}^\kappa[\,\gamma_x\mathscr{A}' \mid \omega\,] - a|\big] < \varepsilon.$$

Since $\gamma_x\mathscr{A}'$ depends only on the colored configuration in $B(x, r)$, we have that on the complement of $\mathscr{B}_x$,

$$\mathbf{P}^\kappa[\,\mathscr{A}', \gamma_x\mathscr{A}' \mid \omega\,] = \mathbf{P}^\kappa[\,\mathscr{A}' \mid \omega\,]\mathbf{P}^\kappa[\,\gamma_x\mathscr{A}' \mid \omega\,],$$

whence $\mathbf{E}[\,|\mathbf{P}^\kappa[\mathscr{A}', \gamma_x\mathscr{A}' \mid \omega] - a^2|] = O(\varepsilon)$. Therefore,

$$a - a^2 = \mathbf{E}\big[\,|\mathbf{P}^\kappa[\,\mathscr{A}, \gamma_x\mathscr{A} \mid \omega\,] - a^2|\big] = O(\,\varepsilon).$$

Since ε was arbitrary, we get that $a \in \{0, 1\}$, as desired. □

We now prove that $p_u < 1$ for all Cayley graphs of Kazhdan groups. Let Γ be a countable group and S a finite subset of Γ. Let $\mathscr{U}(\mathscr{H})$ denote the set of unitary representations of Γ on a Hilbert space $\mathscr{H}$ that have no invariant vectors except 0. Set

$$\kappa(\Gamma, S) := \max\{\,\varepsilon\colon \forall\,\mathscr{H}\ \forall\ \pi \in \mathscr{U}(\mathscr{H})\ \forall\ v \in \mathscr{H},\ \exists s \in S\ \|\pi(\,s)v - v\| \geq \varepsilon\|v\|\}.$$

Then Γ is called *Kazhdan* [or has Kazhdan's property (T)] if $\kappa(T, S) > 0$ for all finite S. The only amenable Kazhdan groups are the finite ones. Examples of Kazhdan groups include $\mathrm{SL}(n, \mathbb{Z})$ for $n \geq 3$. See de la Harpe and Valette (1989) for background; in particular, every Kazhdan group is finitely generated (page 11), but not necessarily finitely presentable (as shown by examples of Gromov; see page 43). It can be shown directly, but also follows from our Corollary 6.6 below, that every infinite Kazhdan group has only one end. See Zuk (1996) for examples of Kazhdan groups arising as fundamental groups of finite simplicial complexes.

Rather than the definition, we shall use the following characterization of Kazhdan groups: Let $\mathbf{P}_*$ be the probability measure on subsets of Γ that is the empty set half the time and all of Γ half the time. Recall that Γ acts by translation on the probability measures on 2^Γ.

THEOREM 6.5 [Glasner and Weiss (1997)]. *A countable infinite group Γ is Kazhdan iff $\mathbf{P}_*$ is not in the weak* closure of the Γ-invariant ergodic probability measures on 2^Γ.*

COROLLARY 6.6. *If G is a Cayley graph of an infinite Kazhdan group Γ, then $p_u(G) < 1$. Moreover, $\mathbf{P}[\exists$ a unique infinite cluster$] = 0$ in Bernoulli$(\,p_u)$ percolation.*

In an older version of this manuscript, only the first statement appeared. We thank Yuval Peres for pointing out that a modification of the original proof produces the stronger second statement.

Schonmann (1999b) proved that Bernoulli$(\,p_u)$ on $T \times \mathbb{Z}$, where T is a regular tree of degree at least three, does not have a unique infinite cluster,

and Peres (1999) generalized this result to nonamenable products. On the other hand, Bernoulli(p_u) percolation on a planar nonamenable transitive graph has a unique infinite cluster. [See Lalley (1998) for the high genus case, and Benjamini and Schramm (1998) for the general case.] Little else is known, however, about the uniqueness of infinite clusters at p_u. For example, the case of lattices in hyperbolic 3-space is still open.

PROOF OF THEOREM 6.5. Suppose that $p_u(G) = 1$. Let o be the identity in Γ, regarded as a vertex in G. Write $\tau_p(x)$ for the probability under Bernoulli(p) percolation that o and x lie in the same cluster. Then by Theorem 4.1, for all $p < 1$ we have $\inf_x \tau_p(x) = 0$. Fix p and let ω be the open subgraph of a Bernoulli(p) percolation. Let η be the union of the sites of some of the clusters of ω, where each cluster is independently put in η with probability $1/2$. By Lemma 6.4, the law $\mathbf{Q}_p$ of η is Γ-invariant and ergodic. Furthermore, any fixed finite subset of Γ either is contained in η or is disjoint from η with high probability when p is sufficiently close to 1. That is, $\mathbf{P}_* = \text{weak}^*\text{-}\lim_{p \to 1} \mathbf{Q}_p$, whence Γ is not Kazhdan.

To prove the stronger statement, suppose that there is a unique infinite component $\mathbf{P}$-a.s. Let $(\omega, \tilde{\omega})$ be the standard coupling of Bernoulli(p) and Bernoulli(p_u) percolation on G. Let η be as above. For a vertex $x \in G$, write $A(x)$ for the set of clusters of ω that lie in the unique infinite cluster of $\tilde{\omega}$ and that are closest to x among those with this property (where distance is measured in G). Define η' to be the union of η with all sites x for which $A(x)$ contains only one cluster and that cluster lies in η. Since the law of $(\omega, \tilde{\omega}, \eta)$ is ergodic by an obvious extension of Lemma 6.4, so is the law $\mathbf{Q}'_p$ of its factor η'. Again, we obtain $\mathbf{P}_* = \text{weak}^*\text{-}\lim_{p \to p_u} \mathbf{Q}'_p$, whence Γ is not Kazhdan. □

REMARK 6.7. For probabilists, we believe that the proof we have presented of Corollary 6.6 is the most natural. For others, we note that one can avoid Lemma 6.4 and Theorem 6.5 by using a theorem of Delorme (1977) and Guichardet (1977) that characterizes Kazhdan groups in terms of positive semidefinite functions. This relies on the fact that $\tau(x, y)$ is positive semidefinite, as observed by Aizenman and Newman (1984).

Other groups that are not finitely presentable and that have provided interesting examples for probability theory are the so-called lamplighter groups [see Kaimanovich and Vershik (1983) and Lyons, Pemantle and Peres (1996); in these references, these groups are denoted G_d and are amenable, but we shall be interested here in nonamenable examples].

We first give a concrete description of a lamplighter graph and later generalize and use more algebraic language. Suppose that G is a graph. The *lamplighter graph* L_G over G is the graph whose vertices are pairs (A, v), where $A \subset \mathsf{V}(G)$ is finite and $v \in \mathsf{V}(G)$. We think of A as the locations of the lamps that are on, and consider v as the location of the lamplighter. One neighbor of (A, v) in L_G is the vertex $(A \triangle \{v\}, v)$ (the lamplighter switches the lamp off or on) and the other neighbors have the form (A, u), where $[v, u] \in \mathsf{E}(G)$ (the lamplighter walks one step).

In the algebraic context and language, lamplighter groups are particular wreath products: Let Γ be a group acting from the left on a set V. Let K be a group; K_*^{V} denotes the group of maps $f\colon \mathsf{V} \to K$ such that $f(x)$ is the identity element id_K of K for all but finitely many $x \in \mathsf{V}$ and with multiplication $(f_1 f_2)(x) := f_1(x) f_2(x)$. Then Γ acts on K_*^{V} by translation: $(\gamma f)(x) := f(\gamma^{-1} x)$. The (*restricted*) *wreath product* $K \wr \Gamma$ is the set $K_*^{\mathsf{V}} \times \Gamma$ with the multiplication

$$(f_1, \gamma_1)(f_2, \gamma_2) := (f_1(\gamma_1 f_2), \gamma_1 \gamma_2).$$

If Γ and K are finitely generated and Γ acts transitively on V, then $K \wr \Gamma$ is finitely generated. To see this, let $\gamma_1, \dots, \gamma_s$ generate Γ and $k_1, \dots, k_t$ generate K. Write id_Γ for the identity element of Γ and id_{V} for the identity element of K_*^{V}. Let $o \in \mathsf{V}$ be fixed. Write F_j for the element of K_*^{V} defined by $F_j(o) := k_j$ and $F_j(x) := \mathrm{id}_K$ for all $x \neq o$. Set

$$S_\Gamma := \{(\mathrm{id}_{\mathsf{V}}, \gamma_i)\colon 1 \le i \le s\},$$

$$S_K := \{(F_j, \mathrm{id}_\Gamma)\colon 1 \le j \le t\}.$$

Then $S := S_\Gamma \cup S_K$ is a finite generating set for $K \wr \Gamma$. Indeed, let $(f, \gamma) \in K \wr \Gamma$. For $x \in \mathsf{V}$, choose $\gamma_x \in \Gamma$ such that $\gamma_x o = x$ and write $h_x \in K_*^{\mathsf{V}}$ for the function $h_x(o) := f(x)$ and $h_x(y) := \mathrm{id}_K$ for $y \neq o$. Then

$$(f, \gamma) = \left(\prod_{\substack{x \in \mathsf{V} \\ f(x) \neq \mathrm{id}_K}} (\mathrm{id}_{\mathsf{V}}, \gamma_x)(h_x, \mathrm{id}_\Gamma)\big(\mathrm{id}_{\mathsf{V}}, \gamma_x^{-1}\big) \right) (\mathrm{id}_{\mathsf{V}}, \gamma), \tag{6.1}$$

where the product over x is taken in any order. If we then write each $(h_x, \mathrm{id}_\Gamma)$ as a product of $(F_j, \mathrm{id}_\Gamma)$, each $(\mathrm{id}_{\mathsf{V}}, \gamma_x)$ as a product of $(\mathrm{id}_{\mathsf{V}}, \gamma_i)$, each $(\mathrm{id}_{\mathsf{V}}, \gamma_x^{-1})$ as a product of $(\mathrm{id}_{\mathsf{V}}, \gamma_i)$, and $(\mathrm{id}_{\mathsf{V}}, \gamma)$ as a product of $(\mathrm{id}_{\mathsf{V}}, \gamma_i)$, we obtain a representation as a product of elements of S.

The lamplighter groups G_d are those where $\Gamma = \mathsf{V} = \mathbb{Z}^d$ and $K = \mathbb{Z}_2$. [By a theorem of Baumslag (1961), the groups G_d are not finitely presentable.]

COROLLARY 6.8. *Let K be a finite group with more than one element. Let Γ be a finitely generated group acting transitively on an infinite set V. If G is any Cayley graph of $K \wr \Gamma$, then $p_u(G) < 1$.*

To prove this, we borrow a technique from Benjamini, Pemantle and Peres (1998): If ϕ and ψ are two paths, denote by $|\phi \cap \psi|$ the number of edges they have in common (as sets of edges).

LEMMA 6.9. *Let G be a graph and x, y be any two vertices in G. Let $\theta \in (0,1)$ and $c > 0$ be constants. Suppose that μ is a probability measure on (possibly self-intersecting) paths ψ joining x to y such that*

$$(\mu \times \mu)\{(\phi, \psi)\colon |\phi \cap \psi| = n\} \le c\theta^n$$

for all $n \in \mathbb{N}$. Let $\theta < p < 1$. Then for Bernoulli(p) percolation on G, we have

$$\tau(x, y) \ge c^{-1}(1 - \theta/p).$$

PROOF. Define the random variable

$$Z := \sum_{\psi} \mu(\psi) \frac{\mathbf{1}_{\{\psi \text{ is open}\}}}{\mathbf{P}[\,\psi \text{ is open}]}.$$

Then $\mathbf{E}[Z] = 1$ and

$$\begin{aligned}\mathbf{E}[Z^2] &= \sum_{\phi,\psi} \mu(\phi)\mu(\psi) \frac{\mathbf{P}[\,\phi, \psi \text{ are both open}]}{\mathbf{P}[\,\phi \text{ is open}]\mathbf{P}[\,\psi \text{ is open}]} \\ &= \sum_{\phi,\psi} \mu(\phi)\mu(\psi)\, p^{-|\phi \cap \psi|} \\ &= \sum_{n} p^{-n} \sum_{|\phi \cap \psi| = n} \mu(\phi)\mu(\psi) \\ &\le \sum_{n} p^{-n} c\theta^n = c(1 - \theta/p)^{-1}.\end{aligned}$$

Therefore, the Cauchy–Schwarz inequality yields

$$\tau(x, y) \ge \mathbf{P}[Z > 0] \ge \mathbf{E}[Z]^2/\mathbf{E}[Z^2] \ge c^{-1}(1 - \theta/p). \qquad \square$$

PROOF OF COROLLARY 6.8. Since K is assumed to be finite, we take the generating set $\{k_1, \dots, k_t\}$ to be all of K. We use the Cayley graph given by the generating set S. It suffices to exhibit a measure on paths connecting $(\mathrm{id}_{\mathsf{V}}, \mathrm{id}_{\Gamma})$ to (f, γ) that satisfies the condition of Lemma 6.9 with θ and c not depending on (f, γ), since then Theorem 4.1 implies that $p_u \le \theta$.

Define edges joining pairs in V by

$$\mathsf{E} := \{[x, \gamma_i^{\varepsilon} x]\colon x \in \mathsf{V},\ 1 \le i \le s,\ \varepsilon = \pm 1\}.$$

The graph $G' := (\mathsf{V}, E)$ will be called the *base graph*. Note that the base graph is not the same as the Cayley graph, $G = (\mathsf{V}(G), \mathsf{E}(G))$. Let $\pi\colon \mathsf{V}(G) \to \mathsf{V}$ be the projection $\pi(g, \beta) := \beta o$. Let $v_1, v_2, \dots$ and $u_1, u_2, \dots$ be infinite simple paths in G' starting at $v_1 = o$ and $u_1 = \gamma o$, respectively, that are disjoint, except that v_1 may equal u_1. Because G' is an infinite, connected, transitive graph, it is easy to show that such paths exist. For each $j = 1, 2, \dots$, fix an $\alpha_j \in S_{\Gamma}$ such that $\alpha_j v_j = v_{j+1}$ and fix a $\beta_j \in S_{\Gamma}$ such that $\beta_j u_j = u_{j+1}$.

Given a word $W = w_1 w_2 \cdots w_n$ with letters from S, let W^{-1} denote the word $w_n^{-1} w_{n-1}^{-1} \cdots w_1^{-1}$, and let $\tilde{W}(j)$, $j \in \{0, 1, \dots, n\}$, denote the group element $w_1 w_2 \cdots w_j$.

Let W_{γ} be a word in S_{Γ} representing $(\mathrm{id}_{\mathsf{V}}, \gamma)$ with the property that for every $v \in \mathsf{V}$ such that $f(v) \ne \mathrm{id}_K$, there is a j such that $\pi\tilde{W}(j) = v$. Let n be the length of W_{γ}. Let W_{α} be the word $\alpha_1 \alpha_2 \cdots \alpha_n$, and let W_{β} be the word $\beta_1 \beta_2 \cdots \beta_n$. Let W be the concatenation

$$W := W_{\alpha} W_{\alpha}^{-1} W_{\gamma} W_{\beta} W_{\beta}^{-1} W_{\gamma}^{-1} W_{\alpha} W_{\alpha}^{-1} W_{\gamma} W_{\beta} W_{\beta}^{-1},$$

and let N be the length of W. Let the letters in W be $W = w_1 w_2 \cdots w_N$. Consider words of the form $\phi(X) := w_1 X_1 w_2 X_2 \cdots w_N X_N$, where $X =$

$(X_1, \dots, X_N) \in S_K^N = (S_K)^N$. For any $v \in \mathsf{V}$, let J_v be the set of $j \in \{1, 2, \dots, N\}$ such that $\pi \tilde{W}(j) = v$. Let $\pi_K \colon K \wr \Gamma \to K_*^{\mathsf{V}}$ be the projection onto the first coordinate. The uniform measure μ_0 on the set of $X \in S_K^N$ such that $\phi(X)$ is a word representing (f, γ) can be described as follows: $\prod_{j \in J_v} \pi_K X_j(o) = f(v)$ (where the order of multiplication is the order of J_v as a subset of $\mathbb{N}$) and for every $j_1 \in J_v$, the random variables $(X_j \colon j \in J_v,\ j \neq j_1)$ are independent, uniform in S_K and independent of $(X_j \colon j \notin J_v)$. Note that $\phi(X)$ can also be thought of as a random path in G from $(\mathrm{id}_{\mathsf{V}}, \mathrm{id}_\Gamma)$ to (f, γ).

Let X and Y be i.i.d. with law μ_0. We want to bound the probability that there are k edges shared by the paths $\phi(X)$ and $\phi(Y)$. In fact, we shall bound the probability that these paths share at least k vertices. Given any $j \in \{1, \dots, N\}$, let $H(j)$ be the set of $v \in \mathsf{V}$ such that $\min J_v < j < \max J_v$, and let $h(j) := |H(j)|$. The choice of W ensures that

$$h(j) \geq \min\{j, N - j, n\} - 1.$$

Because $N = O(n)$, this gives

$$h(j) \geq c_1 \min\{j, N - j\} - 1$$

for some universal constant $c_1 > 0$. If (g, δ) is the element of $\hat{\mathsf{V}}$ represented by the word $w_1 X_1 w_2 \cdots w_j X_j$, then $(g(v) \colon v \in H(j))$, are i.i.d. uniform in K. Consequently, the probability that $w_1 X_1 w_2 \cdots w_j X_j = w_1 Y_1 w_2 \cdots w_{j'} Y_{j'}$, as elements of $K \wr \Gamma$, is at most $|K|^{-\max\{h(j),\, h(j')\}}$. If there are more than $8k$ vertices common to $\phi(X)$ and $\phi(Y)$, then there must be a pair of indices $j, j' \in \{k, k+1, \dots, N-k\}$ such that $w_1 X_1 w_2 \cdots w_j X_j = w_1 Y_1 w_2 \cdots w_{j'} Y_{j'}$, as elements of $K \wr \Gamma$, or with a similar equality when the rightmost letter is dropped from either or both sides. The probability for that is at most

$$4 \sum_{j=k}^{N-k} \sum_{j'=k}^{N-k} |K|^{-\max\{h(j),\, h(j')\}} \leq 4 \sum_{j=k}^{N-k} \sum_{j'=k}^{N-k} |K|^{-(h(j)+h(j'))/2} \leq c_2 |K|^{-c_1 k},$$

where c_2 is a constant depending only on $|K|$. Consequently, as appeal to Lemma 6.9 completes the proof. $\square$

Benjamini and Schramm (1996) conjectured that for any nonamenable Cayley graph G, we have $p_c(G) < p_u(G)$. This is still open, but using our main result on cluster indistinguishability, we may show that a comparable statement fails for invariant percolation processes.

COROLLARY 6.10. *There is an invariant deletion-tolerant percolation process* $(\mathbf{P}, \omega)$ *on a nonamenable Cayley graph that has only finite components a.s., but such that for every* $\varepsilon > 0$, *the union of* ω *with an independent Bernoulli*(ε) *percolation process produces a unique infinite cluster a.s.*

PROOF. Recall that Bernoulli(p) bond percolation on $\mathbb{Z}^2$ produces a.s. no infinite cluster iff $p \leq 1/2$ [see Grimmett (1989)]. Let $\mathbb{F}_2$ be the Cayley graph of the free group on two letters with the usual generating set. Let $G := \mathbb{F}_2 \times \mathbb{Z}^2$. Each edge of G joins vertices (x_1, y_1) to (x_2, y_2), whether either $x_1 = x_2$ and $y_1 \sim y_2$, or $x_1 \sim x_2$ and $y_1 = y_2$. The former type of edge will be called a

$\mathbb{Z}^2$ edge, while the latter will be called an $\mathbb{F}_2$ edge. For each $x \in \mathbb{F}_2$, the set of all edges $[(x, y_1), (x, y_2)]$ is called the $\mathbb{Z}^2$ fiber on x. Let ω be Bernoulli$(1/2)$ bond percolation on all the $\mathbb{Z}^2$ edges; it contains no $\mathbb{F}_2$ edge. Then ω is invariant and deletion tolerant. When we take the union with an independent Bernoulli(ε) process η, the intersection of $\omega \cup \eta$ with each $\mathbb{Z}^2$ fiber contains a.s. a unique infinite cluster. These clusters a.s. connect to each other in G. Thus, $\omega \cup \eta$ contains a unique infinite cluster whose intersection with each $\mathbb{Z}^2$ fiber has an infinite component. Since $\omega \cup \eta$ is insertion tolerant, cluster indistinguishability of $\omega \cup \eta$ implies that it has no other infinite cluster. □

QUESTION 6.11. Is there an insertion-tolerant process with the property exhibited in Corollary 6.10?

The same line of thought is also used to prove the following theorem, which is an answer to a question posed by Yuval Peres (private communication) and motivated by the work of Salzano and Schonmann (1997, 1999) on contact processes.

THEOREM 6.12. *Let G, H and H' be unimodular transitive graphs. Assume that $H' \subset H$ and G is infinite. Then*

$$p_u(G \times H') \geq p_u(G \times H).$$

In particular, when H' consists of a single vertex of H we get

$$p_u(G) \geq p_u(G \times H).$$

PROOF. Let ω be Bernoulli(p) bond percolation on $G \times H$, where $p > p_u(G \times H')$, and let Γ be the product of a unimodular transitive automorphism group on G with a unimodular transitive automorphism group on H. Then Γ is a unimodular transitive automorphism group on $G \times H$.

Let $\hat{Z} := \{\gamma(G \times H') : \gamma \in \Gamma\}$. By Theorem 1.2 applied to $G \times H'$, for every $Z \in \hat{Z}$ a.s. there is a unique infinite cluster, say Q_Z, of $\omega \cap Z$. Now let $Z, Z' \in \hat{Z}$. We claim that a.s. Q_Z and $Q_{Z'}$ are in the same cluster of ω. Indeed, since G is infinite, there are infinite sequences of vertices $v_j \in Z$, $v_j' \in Z'$ such that $\operatorname{dist}(v_j, v_j') = \inf\{\operatorname{dist}(v, v') : v \in Z, v' \in Z'\}$ for all j. Because $\inf_j \mathbf{P}[v_j \in Q_Z, v_j' \in Q_{Z'}] > 0$, with positive probability $v_j \in Q_Z$ and $v_j' \in Q_{Z'}$ for infinitely many j, and by Kolmogorov's 0–1 law this holds a.s. It is then clear that for some such j there is a connection in ω between v_j and v_j'. It follows that a.s. there is a unique cluster Q of ω that contains every Q_Z, $Z \in \hat{Z}$.

Let $\mathscr{A}$ be the set of all pairs $(C, W) \in 2^{\mathsf{V}(G \times H)} \times 2^{\mathsf{E}(G \times H)}$ such that C meets an infinite component of $Z \cap W$ for every $Z \in \hat{Z}$. We know that Q is the only cluster of ω with $(Q, \omega) \in \mathscr{A}$. By Theorem 3.3, it follows then that Q is the only infinite cluster of ω. □

REMARK 6.13. The same theorem holds without the assumption of unimodularity and with "transitive" replaced by "quasi-transitive." This follows similarly from the indistinguishability of robust properties proved by Häggström, Peres and Schonmann (1999).

REMARK 6.14. There are infinite Cayley graphs $H' \subset H$ with $p_u(H') < p_u(H)$. For example, one may take $H' = \mathbb{Z}^2$ and let H be the free product $\mathbb{Z}^2 * \mathbb{Z}_2$, as in Remark 1.3.

The work of Schonmann (1999a) was motivated by an analogy between percolation and contact processes. More specifically, the property of having complete convergence with survival for a contact process is closely analogous to having uniqueness of the infinite cluster for percolation. However, Remark 6.14 describes an instance where this analogy fails, because complete convergence with survival is a property which is monotone in the graph [Salzano and Schonmann (1997)].

Acknowledgments. We are grateful to Itai Benjamini, Olle Häggström, Yuval Peres and Roberto Schonmann for fruitful conversations and helpful advice.

REFERENCES

AIZENMAN, M., KESTEN, H. and NEWMAN, C. M. (1987). Uniqueness of the infinite cluster and continuity of connectivity functions for short- and long-range percolation. *Comm. Math. Phys.* **111** 505–532.

AIZENMAN, M. and NEWMAN, C. M. (1984). Tree graph inequalities and critical behavior in percolation models. *J. Statist. Phys.* **36** 107–143.

ALEXANDER, K. (1995). Simultaneous uniqueness of infinite clusters in stationary random labeled graphs. *Comm. Math. Phys.* **168** 39–55.

BABSON, E. and BENJAMINI, I. (1999). Cut sets and normed cohomology with applications to percolation. *Proc. Amer. Math. Soc.* **127** 589–597.

BAUMSLAG, G. (1961). Wreath products and finitely presented groups. *Math. Z.* **75** 22–28.

BENJAMINI, I., LYONS, R., PERES, Y. and SCHRAMM, O. (1999). Group-invariant percolation on graphs. *Geom. Funct. Anal.* **9** 29–66.

BENJAMINI, I., LYONS, R. and SCHRAMM, O. (1999). Percolation perturbations in potential theory and random walks. In *Random Walks and Discrete Potential Theory* (H. Picardello and W. Woess, eds.) 56–84. Cambridge Univ. Press.

BENJAMINI, I., PEMANTLE, R. and PERES, Y. (1998). Unpredictable paths and percolation. *Ann. Probab.* **26** 1198–1211.

BENJAMINI, I. and SCHRAMM, O. (1996). Percolation beyond $\mathbb{Z}^d$, many questions and a few answers. *Electronic Comm. Probab.* **1** 71–82.

BENJAMINI, I. and SCHRAMM, O. (1998). Percolation in the hyperbolic plane. Unpublished manuscript.

BURTON, R. M. and KEANE, M. (1989). Density and uniqueness in percolation. *Comm. Math. Phys.* **121** 501–505.

DE LA HARPE, P. and VALETTE, A. (1989). La propriété (T) de Kazhdan pour les groupes localement compacts (avec un appendice de Marc Burger). *Astérisque* **175**.

DELORME, P. (1977). 1-cohomologie des représentations unitaires des groupes de Lie semisimples et résolubles. Produits tensoriels continus et représentations. *Bull. Soc. Math. France* **105** 281–336.

GANDOLFI, A., KEANE, M. S. and NEWMAN, C. M. (1992). Uniqueness of the infinite component in a random graph with applications to percolation and spin glasses. *Probab. Theory Related Fields* **92** 511–527.

GLASNER, E. and WEISS, B. (1997). Kazhdan's property T and the geometry of the collection of invariant measures. *Geom. Funct. Anal.* **7** 917–935.

GRIMMETT, G. R. (1989). *Percolation*. Springer, New York.

GRIMMETT, G. R. and NEWMAN, C. M. (1990). Percolation in $\infty + 1$ dimensions. In *Disorder in Physical Systems* (G. R. Grimmett and D. J. A. Welsh, eds.) 167–190. Clarendon Press, Oxford.

GUICHARDET, A. (1977). Étude de la 1-cohomologie et de la topologie du dual pour les groupes de Lie à radical abélien. *Math. Ann.* **228** 215–232.

HÄGGSTRÖM, O. (1997). Infinite clusters in dependent automorphism invariant percolation on trees. *Ann. Probab.* **25** 1423–1436.

HÄGGSTRÖM, O. and PERES, Y. (1999). Monotonicity of uniqueness for percolation on Cayley graphs: all infinite clusters are born simultaneously. *Probab. Theory Related Fields* **113** 273–285.

HÄGGSTRÖM, O., PERES, Y. and SCHONMANN, R. H. (1999). Percolation on transitive graphs as a coalescent process: relentless merging followed by simultaneous uniqueness. In *Perplexing Probability Problems*: *Papers in Honor of H. Kesten* (M. Bramson and R. Durrett, eds.) 69–90. Birkhäuser, Boston.

KAIMANOVICH, V. A. and VERSHIK, A. M. (1983). Random walks on discrete groups: boundary and entropy. *Ann. Probab.* **11** 457–490.

LALLEY, S. P. (1998). Percolation on Fuchsian groups. *Ann. Inst. H. Poincaré Probab. Statist.* **34** 151–177.

LIGGETT, T. M., SCHONMANN, R. H. and STACEY, A. M. (1997). Domination by product measures. *Ann. Probab.* **25** 71–95.

LYONS, R. (1990). Random walks and percolation on trees. *Ann. Probab.* **18** 931–958.

LYONS, R., PEMANTLE, R. and PERES, Y. (1996). Random walks on the Lamplighter Group. *Ann. Probab.* **24** 1993–2006.

LYONS, R. and PERES, Y. (1998). *Probability on Trees and Networks*. Cambridge Univ. Press. To appear. Available at http://php.indiana.edu/~rdlyons/.

LYONS, R. and SCHRAMM, O. (1999). Stationary measures for random walks in a random environment with random scenery. *New York J. Math.* **5** 107–113.

NEWMAN, C. M. and SCHULMAN, L. S. (1981). Infinite clusters in percolation models. *J. Statist. Phys.* **26** 613–628.

PERES, Y. (1999). Percolation on nonamenable products at the uniqueness threshold. *Ann. Inst. J. Poincaré Probab. Statist.* To appear.

SALZANO, M. and SCHONMANN, R. H. (1997). The second lowest extremal invariant measure of the contact process. *Ann. Probab.* **25** 1846–1871.

SALZANO, M. and SCHONMANN, R. H. (1999). The second lowest extremal invariant measure of the contact process II. *Ann. Probab.* **27** 845–875.

SCHONMANN, R. H. (1999a). Stability of infinite clusters in supercritical percolation. *Probab. Theory Related Fields* **113** 287–300.

SCHONMANN, R. H. (1999b). Percolation in $\infty + 1$ dimensions at the uniqueness threshold. In *Perplexing Probability Problems*: *Papers in Honor of H. Kesten* (M. Bramson and R. Durrett, eds.) 53–67. Birkhäuser, Boston.

TROFIMOV, V. I. (1985). Groups of automorphisms of graphs as topological groups. *Math. Notes* **38** 717–720.

ZUK, A. (1996). La propriété (T) de Kazhdan pour les groupes agissant sur les polyèdres. *C. R. Acad. Sci. Paris Sér. I Math.* **323** 453–458.

DEPARTMENT OF MATHEMATICS
INDIANA UNIVERSITY
BLOOMINGTON, INDIANA 47405-5701
E-MAIL: rdlyons@indiana.edu

DEPARTMENT OF MATHEMATICS
WEIZMANN INSTITUTE OF SCIENCE
REHOVOT 76100
ISRAEL
E-MAIL: schramm@wisdom.weizmann.ac.il

JOURNAL OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 14, Number 2, Pages 487–507
S 0894-0347(00)00362-3
Article electronically published on December 28, 2000

PERCOLATION IN THE HYPERBOLIC PLANE

ITAI BENJAMINI AND ODED SCHRAMM

1. Introduction

The purpose of this paper is to study percolation in the hyperbolic plane and in transitive planar graphs that are quasi-isometric to the hyperbolic plane.

There are several sources available which the reader may consult for background on percolation on $\mathbb{Z}^d$ [Gri89] and $\mathbb{R}^d$ [MR96] and for background on percolation on more general graphs [BS96], [Lyo00], [BS99]. For this reason, we will be quite brief here. Background on hyperbolic geometry may be found in [CFKP97] and the references sited there.

Percolation on planar hyperbolic graphs. A graph G is *transitive* if the automorphism group of G acts transitively on the vertices $\mathsf{V}(G)$. An *invariant percolation* on a graph G is a probability measure on the space of subgraphs of G, which is invariant under the automorphisms of G. The connected components of the random subgraph are often called *clusters*. Of special interest are the *Bernoulli* site and bond percolation. The Bernoulli(p) bond percolation is the random subgraph with vertices $\mathsf{V}(G)$, and where each edge is in the percolation subgraph with probability p, independently. In Bernoulli(p) site percolation, each vertex is in the percolation subgraph with probability p, independently, and an edge appears in the percolation subgraph iff its endpoints are present.

Let G be an infinite, connected, planar, transitive graph, with finite vertex degree. Each such graph is quasi-isometric with one and only one of the following spaces: $\mathbb{Z}$, the 3-regular tree, the Euclidean plane $\mathbb{R}^2$, and the hyperbolic plane $\mathbb{H}^2$ [Bab97]. Each of these classes has its distinct geometry, and random processes behave similarly on graphs within each class. On trees, Bernoulli percolation is quite simple, but there are some interesting results concerning invariant percolation [Häg97]. Bernoulli percolation on $\mathbb{Z}^2$ and planar lattices in $\mathbb{R}^2$ has an extensive theory. (See [Gri89].) Burton and Keane [BK91] also obtained a theory of invariant percolation in $\mathbb{Z}^2$.

A transitive graph G has one end if for every finite set of vertices $V_0 \subset \mathsf{V}(G)$ there is precisely one infinite connected component of $G \backslash V_0$. A transitive planar graph G with one end is quasi-isometric to $\mathbb{R}^2$ or to $\mathbb{H}^2$ (see, e.g., [Bab97] or the proof of Proposition 2.1(b), below). Amenability is a simple geometric property which distinguishes these two possibilities. G is *amenable* if for every $\epsilon > 0$ there

Received by the editors January 18, 2000 and, in revised form, November 9, 2000.
2000 *Mathematics Subject Classification.* Primary 82B43; Secondary 60K35, 60D05.
Key words and phrases. Poisson process, Voronoi, isoperimetric constant, nonamenable, planarity, Euler formula, Gauss-Bonnet theorem.
The second author's research was partially supported by the Sam and Ayala Zacks Professorial Chair at the Weizmann Institute.

©2000 American Mathematical Society

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_24, © Springer Science+Business Media, LLC 2011

is a finite set of vertices V_0 such that $|\partial V_0| < \epsilon |V_0|$. A graph quasi-isometric with $\mathbb{H}^2$ must be nonamenable. Grimmett and Newman [GN90] have shown that for some parameter values, Bernoulli percolation on the cartesian product of $\mathbb{Z}$ with a regular tree of sufficiently high degree has infinitely many infinite components. This cannot happen on amenable graphs, by the Burton-Keane argument [BK89].

Let $p_u = p_u(G)$ be the infimum of the set of $p \in [0,1]$ such that Bernoulli(p) percolation on G has a unique infinite cluster a.s. The critical parameter $p_c = p_c(G)$ is defined as the infimum of the set of $p \in [0,1]$ such that Bernoulli(p) percolation on G has an infinite cluster a.s. It has been conjectured [BS96] that $p_c(G) < p_u(G)$ for every nonamenable transitive graph G. It has been recently proven by Pak and Smirnova-Nagnibeda [PSN00] that every nonamenable group has some Cayley graph G satisfying $p_c(G) < p_u(G)$. Lalley [Lal98] proved $p_c(G) < p_u(G)$ for planar Cayley graphs of Fuchsian groups with sufficiently high genus. We now show

Theorem 1.1. *Let G be a transitive, nonamenable, planar graph with one end. Then $0 < p_c(G) < p_u(G) < 1$, for Bernoulli bond or site percolation on G.*

The hyperbolic plane seems to be a very good testing ground for conjectures about nonamenable graphs. The planarity and hyperbolic geometry help to settle questions that may be more difficult in general.

Theorem 1.2. *Let G be a transitive, nonamenable, planar graph with one end. Then Bernoulli(p_u) percolation on G has a unique infinite cluster a.s.*

This contrasts with a result of Schonmann [Sch99a], which shows that a.s. the number of infinite components of Bernoulli(p_u) percolation on $T \times \mathbb{Z}$ is either 0 or ∞ when T is a regular tree. (If T is a regular tree of sufficiently high degree, then $p_u(T \times \mathbb{Z}) > p_c(T \times \mathbb{Z})$, by [GN90], and hence there are infinitely many infinite components at p_u on $T \times \mathbb{Z}$.) Peres [Per00] has generalized this result of [Sch99a] (with a different proof) to nonamenable products.

It is known that on a transitive graph G when $p > p_u(G)$, Bernoulli(p) percolation a.s. has a unique infinite cluster. This has been proven by Häggström and Peres [HP99] in the unimodular setting, and in full generality by Schonmann [Sch99b]. The proof of Theorem 1.2 will also prove this "uniqueness monotonicity" in this restricted setting.

The situation at p_c is also known. The following theorem is from [BLPS99a]; see also [BLPS99b].

Theorem 1.3. *Let G be a nonamenable graph with a vertex-transitive unimodular automorphism group. Then a.s. critical Bernoulli bond or site percolation on G has no infinite components.*

Note that a planar transitive graph with one end has a unimodular automorphism group (Proposition 2.1). Hence the above theorem applies to the graphs under consideration here.

Percolation in $\mathbb{H}^2$. Though percolation is usually studied on graphs, and many interesting phenomena already appear in the graph setup, there are some special properties that only appear in the continuous setting.

We therefore consider the Poisson-Voronoi-Bernoulli percolation model in the hyperbolic plane $\mathbb{H}^2$. There are two parameters needed to specify the model, $\lambda > 0$ and $p \in [0,1]$. Given such a pair (p, λ), consider a Poisson point process with intensity λ in the hyperbolic plane. Such a process gives rise to a Voronoi tessallation of

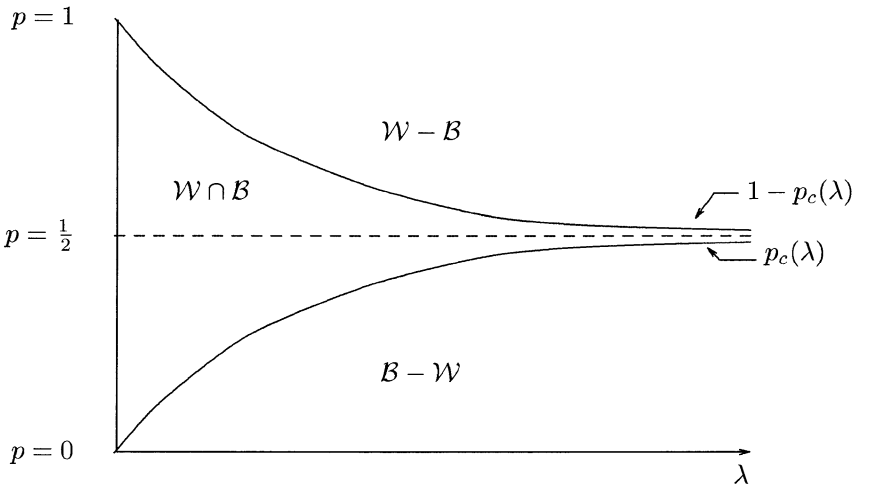


FIGURE 1.1. The phase diagram of Voronoi percolation

$\mathbb{H}^2$. Each tile of this tessallation is colored white (respectively, black) with probability p (respectively, $1-p$), independently. The union of all the white (resp. black) tiles is denoted by $\widehat{W}$ (resp. $\widehat{B}$). Let $\mathcal{W}$ (resp. $\mathcal{B}$) denote the set of pairs (p, λ) such that $\widehat{W}$ (resp. $\widehat{B}$) has an unbounded component a.s. For $\lambda > 0$, let

$$p_c(\lambda) := \inf\{p \in [0,1] : (p, \lambda) \in \mathcal{W}\}.$$

We prove the following (see Figure 1.1)

Theorem 1.4. *Consider (p, λ)-Voronoi percolation in $\mathbb{H}^2$.*

(a) *If $(p, \lambda) \in \mathcal{W} \cap \mathcal{B}$, then there are a.s. infinitely many unbounded components of $\widehat{W}$ and there are infinitely many unbounded components of $\widehat{B}$.*
(b) *If $(p, \lambda) \in \mathcal{W} \backslash \mathcal{B}$, then a.s. there is a unique unbounded components of $\widehat{W}$, and no unbounded components of $\widehat{B}$.*
(c) *If $(p, \lambda) \in \mathcal{B} \backslash \mathcal{W}$, then a.s. there is a unique unbounded component of $\widehat{B}$, and no unbounded components of $\widehat{W}$.*

Note that, by symmetry,

$$\mathcal{B} = \{(p, \lambda) : (1 - p, \lambda) \in \mathcal{W}\}.$$

Theorem 1.5. (a) $0 < p_c(\lambda) \leq \frac{1}{2} - \frac{1}{4\lambda\pi + 2} < \frac{1}{2}$.
(b) $\lim_{\lambda \to 0} p_c(\lambda) = 0$.
(c) $p_c(\lambda)$ *is continuous.*
(d) $(p_c(\lambda), \lambda) \notin \mathcal{W}$ *for all* $\lambda > 0$, *and hence* $\mathcal{W} = \{(p, \lambda) : p > p_c(\lambda)\}$.

We conjecture that $p_c(\lambda) \to 1/2$ as $\lambda \to \infty$. Note that percolation with parameters (p, λ) on $\mathbb{H}^2$ is equivalent to percolation with parameters $(p, 1)$ on $\mathbb{H}^2$ with the metric rescaled by $\sqrt{\lambda}$. Hence, taking $\lambda \to \infty$ amounts to the same as letting the curvature tend to zero. This means that Voronoi percolation on $\mathbb{R}^2$ can be seen as a limit of Voronoi percolation on $\mathbb{H}^2$. See Question 7.5 for further discussion of this issue.

We generalize the Mass Transport Principle ([Häg97], [BLPS99a]) to the hyperbolic plane, and use it as a tool for our investigations. One consequence of the Mass Transport Principle that we derive and use is a generalization of Euler's formula $|V|-|E|+|F|=2$ relating the number of vertices, edges, and faces in a (finite) tiling of the sphere to random (infinite) tilings of the hyperbolic plane with invariant law.

A collection of open problems is presented at the end of the paper. Most of these are related to the Voronoi percolation model. We believe that this process deserves further study.

2. Terminology and preliminaries

All the graphs that we shall consider in this paper are locally finite; that is, each vertex has finitely many incident edges. The vertices of a graph G will be denoted by $\mathsf{V}(G)$, and the edges by $\mathsf{E}(G)$.

Given a graph G, let $\mathrm{Aut}(G)$ denote the group of automorphisms of G. G is *transitive* if $\mathrm{Aut}(G)$ acts transitively on the vertices $\mathsf{V}(G)$. G is *quasi-transitive* if $\mathsf{V}(G)/\mathrm{Aut}(G)$ is finite; that is, there are finitely many $\mathrm{Aut}(G)$ orbits in $\mathsf{V}(G)$. A graph G is *unimodular* if $\mathrm{Aut}(G)$ is a unimodular group (which means that the left-invariant Haar measure is also right-invariant). Cayley graphs are unimodular, and any graph such that $\mathrm{Aut}(G)$ is discrete is unimodular. See [BLPS99a] for a further discussion of unimodularity and its relevance to percolation.

An *invariant percolation* on G is a probability measure on the space of subgraphs of G, which is $\mathrm{Aut}(G)$-invariant. A *cluster* is a connected component of the percolation subgraph.

Let $X=\mathbb{R}^2$ or $X=\mathbb{H}^2$. We say that an embedded graph $G\subset X$ in X is *properly* embedded if every compact subset of X contains finitely many vertices of G and intersects finitely many edges. Suppose that G is an infinite connected graph with one end, properly embedded in X. Let $G^\dagger$ denote the dual graph of G. We assume that $G^\dagger$ is embedded in X in the standard way relative to G; that is, every vertex $v^\dagger$ of $G^\dagger$ lies in the corresponding face of G, and every edge $e\in\mathsf{E}(G)$ intersects only the dual edge $e^\dagger\in\mathsf{E}(G^\dagger)$, and only in one point. If ω is a subset of the edges $\mathsf{E}(G)$, then $\omega^\dagger$ will denote the set

$$\omega^\dagger := \{e^\dagger : e\notin\omega\}.$$

Given $p\in[0,1]$ and a graph G, we often denote the percolation graph of Bernoulli(p) bond percolation on G by ω_p.

Proposition 2.1. *Let G be a transitive, nonamenable, planar graph with one end, and let Γ be the group of automorphisms of G.*

(a) *Γ is discrete (and hence unimodular).*

(b) *G can be embedded as a graph G' in the hyperbolic plane $\mathbb{H}^2$ in such a way that the action of Γ on G' extends to an isometric action on $\mathbb{H}^2$. Moreover, the embedding can be chosen in such a way that the edges of G' are hyperbolic line segments.*

A sketch of the proof of (b) appears in [Bab97]. We include a proof here, for completeness.

Proof of Proposition 2.1. By [Mad70] or by [Wat70] it follows that G is 3-vertex connected; that is, every finite nonempty set of vertices $V_0\subset\mathsf{V}(G)$, $V_0\neq\varnothing$, neighbors with at least 3 vertices in $\mathsf{V}(G)\backslash V_0$. Therefore, by the extension of Imrich

to Whitney's Theorem [Imr75], the embedding of G in the plane is unique, in the sense that in any two embeddings of G in the plane, the cyclic orientation of the edges going out of the vertices is either identical for all the vertices, or reversed for all the vertices. This implies that an automorphism of G that fixes a vertex and all its neighbors is the identity, and therefore $\mathrm{Aut}(G)$ is discrete. For a discrete group, the counting measure is Haar measure, and is both left- and right-invariant. Hence $\mathrm{Aut}(G)$ is unimodular. This proves part (a).

Think of G as embedded in the plane. Call a component of $S^2 - G$ a *face* if its boundary consists of finitely many edges in G. In each face f put a new vertex v_f, and connect it by edges to the vertices on the boundary of f. If this is done appropriately, then the resulting graph $\widehat{G}$ is still embedded in the plane. Note that $\widehat{G}$ together with all its faces forms a triangulation T of a simply connected domain in S^2. To prove (b) it is enough to produce a triangulation T' of $\mathbb{H}^2$ isomorphic with T such that the elements of $\mathrm{Aut}(T')$ extend to isometries of $\mathbb{H}^2$ and the edges of T' are hyperbolic line segments. There are various ways to do this; one of them is with circle packing theory. See, for example, [BSt90], [HS95], or [Bab97]. □

3. The number of components

Theorem 3.1. *Let G be a transitive, nonamenable, planar graph with one end, and let ω be an invariant bond percolation on G. Let k be the number of infinite components of ω, and let $k^\dagger$ be the number of infinite components of $\omega^\dagger$. Then a.s.*

$$(k, k^\dagger) \in \{(1,0), (0,1), (1,\infty), (\infty,1), (\infty,\infty)\}.$$

Remark 3.2. Each of these possibilities can happen. The case $(k, k^\dagger) = (1, \infty)$ appears when ω is the free spanning forest of G, while $(\infty, 1)$ is the situation for the wired spanning forest. See [BLPS00]. The other possibilities occur for Bernoulli percolation, as we shall see.

Lemma 3.3. *Let G be a transitive, nonamenable, planar graph with one end. Let ω be an invariant percolation on G. If ω has only finite components a.s., then $\omega^\dagger$ has infinite components a.s.*

The proof will use a result from [BLPS99a], which says that when the expected degree $\mathbf{E}\deg_\omega v$ of a vertex v in an invariant percolation on a unimodular nonamenable graph G is sufficiently close to $\deg_G v$, there are infinite clusters in ω with positive probability. (The case of trees was established earlier in [Häg97].)

Proof. Suppose that both ω and $\omega^\dagger$ have only finite components a.s. Then a.s. given a component K of ω, there is a unique component K' of $\omega^\dagger$ that surrounds it. Similarly, for every component K of $\omega^\dagger$, there is a unique component K' of ω that surrounds it. Let $\mathcal{K}_0$ denote the set of all components of ω. Inductively, set

$$\mathcal{K}_{j+1} := \{K'' : K \in \mathcal{K}_j\}.$$

For $K \in \mathcal{K}_0$ let $r(K) := \sup\{j : K \in \mathcal{K}_j\}$ be the *rank* of K, and define $r(v) := r(K)$ if K is the component of v in ω. For each r let ω^r be the set of edges in $\mathsf{E}(G)$ incident with vertices $v \in \mathsf{V}(G)$ with $r(v) \le r$. Then ω^r is an invariant bond percolation and

$$\lim_{r\to\infty} \mathbf{E}[\deg_{\omega^r} v] = \deg_G v.$$

Consequently, by the above result from [BLPS99a], we find that ω^r has with positive probability infinite components for all sufficiently large r. This contradicts the assumption that ω and $\omega^\dagger$ have only finite components a.s. □

The following has been proven in [BLPS99a] and [BLPS99b] in the transitive case. The extension to the quasi-transitive case is straightforward.

Theorem 3.4. *Let G be a nonamenable, quasi-transitive, unimodular graph, and let ω be an invariant percolation on G which has a single component a.s. Then $p_c(\omega) < 1$ a.s.*

The following has been proven in [BLPS99a].

Lemma 3.5. *Let G be a quasi-transitive, nonamenable, planar graph with one end, and let ω be an invariant percolation on G. Then a.s. the number of infinite components of ω is 0, 1, or ∞.*

For the sake of completeness, we present a (somewhat different) proof here.

Proof. In order to reach a contradiction, assume that with positive probability ω has a finite number $k > 1$ of infinite components, and condition on that event. Select at random, uniformly, a pair of distinct infinite components ω_1, ω_2 of ω. Let ω_1^c be the subgraph of G spanned by the vertices outside of ω_1, and let τ be the set of edges of G that connect vertices in ω_1 to vertices in the component of ω_1^c containing ω_2. Set

$$\tau^\dagger := \{e^\dagger : e \in \tau\}.$$

Then $\tau^\dagger$ is an invariant bond percolation in the dual graph $G^\dagger$. Using planarity, it is easy to verify that $\tau^\dagger$ is a.s. a bi-infinite path. This contradicts Theorem 3.4, and thereby completes the proof. □

Corollary 3.6. *Let G be a transitive, nonamenable, planar graph with one end. Let ω be an invariant percolation on G. Suppose that both ω and $\omega^\dagger$ have infinite components a.s. Then a.s. at least one among ω and $\omega^\dagger$ has infinitely many infinite components.*

Proof. Draw G and $G^\dagger$ in the plane in such a way that every edge e intersects $e^\dagger$ in one point, v_e, and there are no other intersections of G and $G^\dagger$. This defines a new graph $\widehat{G}$, whose vertices are $\mathsf{V}(G) \cup \mathsf{V}(G^\dagger) \cup \{v_e : e \in \mathsf{E}(G)\}$. Note that $\widehat{G}$ is quasi-transitive.

Set

$$\widehat{\omega} := \{[v, v_e] \in \mathsf{E}(\widehat{G}) : v \in \mathsf{V}(G),\ e \in \omega\} \cup \{[v^\dagger, v_e] \in \mathsf{E}(\widehat{G}) : v^\dagger \in \mathsf{V}(G^\dagger),\ e \notin \omega\}.$$

Then $\widehat{\omega}$ is an invariant percolation on $\widehat{G}$. Note that the number of infinite components of $\widehat{\omega}$ is the number of infinite components of ω plus the number of infinite components of $\omega^\dagger$. By Lemma 3.5 applied to $\widehat{\omega}$, we find that $\widehat{\omega}$ has infinitely many infinite components. □

Proof of Theorem 3.1. Each of k, $k^\dagger$ is in $\{0, 1, \infty\}$ by Lemma 3.5. The case $(k, k^\dagger) = (0, 0)$ is ruled out by Lemma 3.3. Since every two infinite components of ω must be separated by some component of $\omega^\dagger$, the situation $(k, k^\dagger) = (\infty, 0)$ is impossible. The same reasoning shows that $(k, k^\dagger) = (0, \infty)$ cannot happen. The case $(k, k^\dagger) = (1, 1)$ is ruled out by Corollary 3.6. □

Theorem 3.7. *Let G be a transitive, nonamenable, planar graph with one end, and let ω be Bernoulli(p) bond percolation on G. Let k be the number of infinite components of ω, and let $k^\dagger$ be the number of infinite components of $\omega^\dagger$. Then a.s.*

$$(k, k^\dagger) \in \{(1,0), (0,1), (\infty, \infty)\}.$$

Proof. By Theorem 3.1, it is enough to rule out the cases $(1, \infty)$ and $(\infty, 1)$. Let K be a finite connected subgraph of G. If K intersects two distinct infinite components of ω, then $\omega^\dagger - \{e^\dagger : e \in \mathsf{E}(K)\}$ has more than one infinite component. If $k > 1$ with positive probability, then there is some finite subgraph K such that K intersects two infinite components of ω with positive probability. Therefore, we find that $k^\dagger > 1$ with positive probability (since the distribution of $\omega^\dagger - \{e^\dagger : e \in \mathsf{E}(K)\}$ is absolutely continuous to the distribution of $\omega^\dagger$). By ergodicity, this gives $k^\dagger > 1$ a.s. An entirely dual argument shows that $k > 1$ a.s. when $k^\dagger > 1$ with positive probability. □

Theorem 3.8. *Let G be a transitive, nonamenable, planar graph with one end. Then $p_c(G^\dagger) + p_u(G) = 1$ for Bernoulli bond percolation.*

Proof. Let ω_p be Bernoulli(p) bond percolation on G. Then $\omega_p^\dagger$ is Bernoulli($1-p$) bond percolation on $G^\dagger$. It follows from Theorem 3.7 that the number of infinite components $k^\dagger$ of $\omega^\dagger$ is 1 when $p < p_c(G)$, ∞ when $p \in (p_c(G), p_u(G))$, and 0 when $p > p_u(G)$. □

Proof of Theorem 1.1. We start with the proof for bond percolation. The easy inequality $p_c(G) \geq 1/(d-1)$, where d is the maximal degree of the vertices in G, is well known.

Set $p_c = p_c(G)$. By Theorem 1.3, ω_{p_c} has only finite components a.s. By Theorem 3.7, $(\omega_{p_c})^\dagger$ has a unique infinite component a.s. Consequently, by Theorem 1.3 again, $(\omega_{p_c})^\dagger$ is supercritical; that is, $p_c(G^\dagger) < 1 - p_c(G)$. An appeal to Theorem 3.8 now establishes the inequality $p_c(G) < p_u(G)$.

Since $p_u(G) = 1 - p_c(G^\dagger) \leq 1 - 1/(d^\dagger - 1)$, where $d^\dagger$ is the maximal degree of the vertices in $G^\dagger$, we get $p_u(G) < 1$, and the proof for bond percolation is complete.

If ω is site percolation on G, let ω^b be the set of edges of G with both endpoints in ω. Then ω^b is a bond percolation on G. In this way, results for bond percolation can be adapted to site percolation. However, even if ω is Bernoulli, ω^b is not. Still, it is easy to check that the above proof applies also to ω^b. The details are left to the reader. □

Proof of Theorem 1.2. By Proposition 2.1, $\mathrm{Aut}(G)$ is discrete and unimodular. By Theorem 3.8, $(\omega_{p_u})^\dagger$ is critical Bernoulli bond percolation on $G^\dagger$. Hence, by Theorem 1.3, $(\omega_{p_u})^\dagger$ has a.s. no infinite components. Therefore, it follows from Theorem 3.1 that ω_{p_u} has a single infinite component. □

4. Geometric consequences

We now investigate briefly the geometry of percolation clusters on vertex-transitive tilings of $\mathbb{H}^2$. Recall that the *ideal boundary* $\partial\mathbb{H}^2$ of $\mathbb{H}^2$ is homeomorphic to the circle S^1. Given a point $o \in \mathbb{H}^2$, $\partial\mathbb{H}^2$ can be identified with the space of infinite geodesic rays starting from o. Let z_n be a sequence in $\mathbb{H}^2$. We say that z_n converges to a point z in $\partial\mathbb{H}^2$, if the geodesic segments $[o, z_n]$ converge to the ray corresponding to z. (That is, the length of $[o, z_n]$ tends to infinity and the angle at

o between the ray from o corresponding to z and the segment $[o, z_n]$ tends to zero.) One can show that the convergence of z_n does not depend on the choice of o.

Theorem 4.1. *Let T be a vertex-transitive tiling of $\mathbb{H}^2$ with finite sided faces, let G be the graph of T, and let ω be Bernoulli percolation on G. Almost surely, every infinite component of ω contains a path that has a unique limit point in the ideal boundary of $\mathbb{H}^2$.*

In [BLS99] it is shown that for any unimodular transitive graph G and every $p \in [0,1]$, a.s. every infinite component of Bernoulli(p) percolation on G is transient and simple random walk on it has positive speed.

Proof. Suppose that ω has infinite components with positive probability. Let $X(t)$ be a simple random walk on an infinite component of ω. Then, by the above result of [BLS99], a.s. there is a $\Lambda > 0$ such that

$$\operatorname{dist}(X(t), X(0)) \geq t\Lambda, \tag{4.1}$$

for all sufficiently large t, where dist is the distance in the hyperbolic metric. Clearly, we also have

$$\operatorname{dist}(X(t), X(s)) \leq L|t-s|, \tag{4.2}$$

for some constant L. These inequalities are enough to conclude that $X(t)$ tends to a limit in the ideal boundary. Indeed, fix a polar coordinate system (r, θ), where $r = r(p)$ is the hyperbolic distance $\operatorname{dist}(X(0), p)$ from p to $X(0)$ and θ is the angle between the segment $[X(0,), p]$ and some fixed ray starting at $X(0)$. Note that there are constants $c < 1 < C$ such that for points $p, q \in \mathbb{H}^2$

$$d(\theta(p), \theta(q)) \leq C c^{r(p) - \operatorname{dist}(p,q)},$$

where the distance $d(\theta(p), \theta(q))$ refers to the arclength distance on the unit circle. It therefore follows immediately from (4.1) and (4.2) that $\{\theta(X(t))\}$ is a Cauchy sequence; that is, $\lim_t \theta(X(t))$ exists. This implies that $X(t)$ tends to a limit in the ideal boundary. □

There is also a proof that does not use speed, but instead uses the easier result from [BLS99] that the infinite components are transient. We now give that proof.

Recall that a metric d_0 on the vertices of a graph H is proper, if every ball of finite radius contains finitely many vertices.

Lemma 4.2. *Let G be an infinite connected (locally finite) graph. Then G is transient iff every proper metric on G satisfies*

$$\sum_{[v,u] \in \mathsf{E}(G)} d_0(v,u)^2 = \infty.$$

This easy observation is certainly not new.

Proof. A function $f : \mathsf{V}(G) \to \mathbb{R}$ is *proper* if $f^{-1}(K)$ is finite for every compact K. The *Dirichlet energy* of f is $\mathcal{D}(f) := \sum_{[v,u] \in \mathsf{E}(G)} (f(u) - f(v))^2$. It is well known that G is recurrent iff there is a proper function $f : \mathsf{V}(G) \to [0, \infty)$ with finite Dirichlet energy. (See [DS84] or [Lyo01].)

Suppose that f is proper and has finite Dirichlet energy. Without loss of generality, we may assume that $f(v) \neq f(u)$ if $v \neq u$. Then define $d_0(v,u) := |f(v) - f(u)|$. In the other direction, given a proper metric d_0, set $f(u) = d_0(o, u)$, where $o \in \mathsf{V}(G)$ is arbitrary. By the triangle inequality it follows that $\mathcal{D}(f) < \infty$. □

Second proof of Theorem 4.1. We first prove that every transient connected subgraph $H \subset G$ has a path with a limit in the ideal boundary. Indeed, consider the Poincaré (disk) model for the hyperbolic plane $\mathbb{H}^2$. Given two points, $x, y \in \mathbb{H}^2$, let $d_E(x, y)$ be the *Euclidean* distance from x to y. Let d_E^H be the maximal metric on $\mathsf{V}(H)$ such that $d_E^H(v, u) \leq d_E(v, u)$ whenever $[v, u] \in \mathsf{E}(H)$ (in other words, $d_E^H(v, u) = \inf \sum_{j=1}^n d_E(v_{j-1}, v_j)$, where the infimum is taken with respect to all finite paths $v_0, v_1, \ldots, v_n$ in H from v to u).

We now show that

$$\sum_{[v,u]\in\mathsf{E}(G)} d_E(v, u)^2 < \infty. \tag{4.3}$$

Let $o \in \mathsf{V}(G)$ be a basepoint, and for every neighbor u of o let D_u be a compact disk in $\mathbb{H}^2$ whose boundary contains the points o and u. Let P be the collection of all disks of the form γD_u where u neighbors with o and $\gamma \in \Gamma$ is an automorphism of G acting on $\mathbb{H}^2$ as an isometry. Since Γ acts discretely on $\mathbb{H}^2$, there is a finite upper bound M for the number of disks in P that contain any point $z \in \mathbb{H}^2$. Since each disk in P is contained in $\mathbb{H}^2$, it follows that the sum of the Euclidean areas of the disks in P is finite. As the square of the Euclidean diameter of a disk is linearly related to the Euclidean area, this proves (4.3).

Because H is transient, Lemma 4.2 implies that the metric d_E^H is not proper. Hence there is an infinite simple path in H with finite d_E-length. Theorem 4.1 follows, because a.s. the infinite components of ω are transient, by [BLS99]. □

Lemma 4.3. *Let T be a vertex-transitive tiling of $\mathbb{H}^2$ with finite sided faces, let G be the graph of T, and let ω be an invariant percolation on G. Let Z be the set of points z in the ideal boundary $\partial\mathbb{H}^2$ such that there is a path in ω with limit z. Then a.s. $Z = \varnothing$ or Z is dense in $\partial\mathbb{H}^2$.*

Proof. Given a vertex $v \in \mathsf{V}(G)$, we may consider a polar coordinate system with v as the origin, and with respect to this coordinate system $\partial\mathbb{H}^2$ can be thought of as a metric circle of circumference 2π. Let d_v denote this metric of $\partial\mathbb{H}^2$, and let $a(v)$ be the length of the largest component of $\partial\mathbb{H}^2 \backslash Z$, with respect to d_v. Note that for vertices $v \in \mathsf{V}(G)$, the law of the random variable $a(v)$ does not depend on v. Let $o \in \mathsf{V}(G)$, let $\epsilon \in (0, 1)$, and let δ be the probability that $\epsilon < a(o) < 2\pi - \epsilon$.

Suppose that $\delta > 0$. Let $R > 0$ be very large, and let x be a random-uniform point on the circle of radius R about o in $\mathbb{H}^2$. Let v_x be the vertex of G closest to x. On the event $\epsilon < a(o) < 2\pi - \epsilon$, there is probability greater than $\epsilon/(4\pi)$ that the geodesic ray from o containing x hits $\partial\mathbb{H}^2$ at a point x' with $d_o(x', Z) \geq \epsilon/4$; this happens when x' is within the inner half of the largest arc of $\partial\mathbb{H}^2 \backslash Z$ with respect to d_o. On that event, if R is very large (as a function of ϵ), we have $a(v_x)$ as close as we wish to 2π, and $a(v_x) \neq 2\pi$. However, since

$$\mathbf{P}[a(v_x) \in (2\pi - t, 2\pi)] = \mathbf{P}[a(o) \in (2\pi - t, 2\pi)] \to 0, \quad \text{as } t \searrow 0,$$

it follows that $\mathbf{P}[a(o) \in (\epsilon, 2\pi - \epsilon)] = 0$. Consequently, $a(o) \in \{0, 2\pi\}$ a.s.

It remains to deal with the case that Z is a single point with positive probability. But this is impossible, since the law of this point would be a finite positive measure on $\partial\mathbb{H}^2$ which is invariant under the automorphisms of T, and such a measure does not exist. (To verify this, observe that such a measure can have no atoms, since the orbit of any point in $\partial\mathbb{H}^2$ is infinite. Given any finite positive atomless measure on $\mathbb{H}^2$ and a small $\epsilon > 0$, there is a bounded set of points $x \in \mathbb{H}^2$ with the property that

for every half-plane which contains x, the arc of $\partial\mathbb{H}^2$ associated with it has measure at least ϵ. However, the orbit of every point $x \in \mathbb{H}^2$ under the automorphisms of T is infinite.) □

Corollary 4.4. *Let T be a vertex-transitive tiling of $\mathbb{H}^2$ with finite sided faces, let G be the graph of T, and let ω be Bernoulli percolation on G. If ω has infinite components a.s., then for every half-space $W \subset \mathbb{H}^2$, a.s. $\omega \cap W$ has infinite components.*

The significance of percolation in hyperbolic half-spaces was noted by [Lal98].

Proof. This follows immediately from Theorem 4.1 and Lemma 4.3. □

Corollary 4.5 (Connectivity decay)**.** *Let T be a vertex-transitive tiling of $\mathbb{H}^2$ with finite sided faces, let G be the graph of T, let $p < p_u(G)$, and let $\omega = \omega_p$ be Bernoulli(p) percolation on G. Let $\tau(v,u)$ denote the probability that v and u are in the same cluster of ω. There is some $a = a(G,p) < 1$ such that $\tau(v,u) \le a^{d(v,u)}$ for every $v, u \in \mathsf{V}(G)$.*

Proof. There are constants $b = b(G,p)$ and $c = c(G,p) > 0$ such that if L and L' are two hyperbolic lines and the minimal distance between L and L' is at least b, then with probability at least c there is an infinite path in $\omega^\dagger$ which separates L from L' in $\mathbb{H}^2$ and does not intersect $L \cup L'$. This follows from Corollary 4.4 (or rather its generalization to the quasi-transitive setting), since, when b is large enough, for given L and L', there are hyperbolic half-spaces H and H', disjoint from $L \cup L'$, such that $H \cup H'$ separates L from L'. Moreover, the automorphisms of G act co-compactly on $\mathbb{H}^2$, so it is enough to consider a compact collection of pairs (L, L').

If $v, u \in \mathsf{V}(G)$ and $d(v,u)$ is large, then one can find a collection of hyperbolic lines $L_0, L_1, \ldots, L_m$, with $d(v,u) \le O(1)m$, such that the distance between L_j and L_{j+1} is b, and for each $j = 1, 2, \ldots, m$, the line L_j separates $\{v\} \cup L_0 \cup \cdots \cup L_{j-1}$ from $L_{j+1} \cup \cdots \cup L_m \cup \{u\}$. The corollary immediately follows, since, by independence, with probability at least $1 - (1-c)^m$ there is some $j = 1, 2, \ldots, m$ and a path $\gamma \subset \omega^\dagger$ which separates L_{j-1} from L_j, and thereby separates u from v. □

5. Mass transport in the hyperbolic plane and some applications

The Mass Transport Principle [Häg97] has an important role in the study of percolation on nonamenable transitive graphs. See, e.g., [BLPS99a]. We now develop a continuous version of this principle, in the setting of the hyperbolic plane, and later produce several applications.

Definition 5.1. A *diagonally-invariant* measure μ on $\mathbb{H}^2 \times \mathbb{H}^2$ is a measure satisfying

$$\mu(gA \times gB) = \mu(A \times B)$$

for all measurable $A, B \subset \mathbb{H}^2$ and $g \in \mathrm{Isom}(\mathbb{H}^2)$.

Theorem 5.2 (Mass Transport Principle in $\mathbb{H}^2$)**.** *Let μ be a nonnegative diagonally-invariant Borel measure on $\mathbb{H}^2 \times \mathbb{H}^2$. Suppose that $\mu(A \times \mathbb{H}^2) < \infty$ for some open $A \subset \mathbb{H}^2$. Then*

$$\mu(B \times \mathbb{H}^2) = \mu(\mathbb{H}^2 \times B)$$

for all measurable $B \subset \mathbb{H}^2$. Moreover, there is a constant c such that $\mu(B \times \mathbb{H}^2) = c\,\mathrm{area}(B)$ for all measurable $B \subset \mathbb{H}^2$.

The same conclusions apply in the case where μ is a diagonally-invariant Borel signed measure on $\mathbb{H}^2 \times \mathbb{H}^2$, provided that $|\mu|(K \times \mathbb{H}^2) < \infty$ for every compact $K \subset \mathbb{H}^2$.

Recall that the subgroup $\mathrm{Isom}_+(\mathbb{H}^2)$ of orientation preserving isometries is a simple group. Consequently, $\mathrm{Isom}_+(\mathbb{H}^2)$ must be contained in the kernel of the modular function, which is a homomorphism from $\mathrm{Isom}(\mathbb{H}^2)$ to the multiplicative group $\mathbb{R}_+$. It follows that the modular function is identically 1; that is, $\mathrm{Isom}(\mathbb{H}^2)$ is unimodular. Let η denote Haar measure of $\mathrm{Isom}(\mathbb{H}^2)$. The particular consequence of unimodularity that we shall need is that $\eta(A) = \eta\{g^{-1} : g \in A\}$ for every measurable $A \subset \mathrm{Isom}(\mathbb{H}^2)$. (Indeed, set $\nu(A) := \eta\{g^{-1} : g \in A\}$. Then it is easy to see that ν is left-invariant, and by the uniqueness of Haar measure it must be a multiple of η. By choosing A symmetric with respect to $g \to g^{-1}$, it follows that $\nu = \eta$.)

Proof. We first prove the nonnegative case. Note that $\nu(B) = \mu(B \times \mathbb{H}^2)$ is an $\mathrm{Isom}(\mathbb{H}^2)$-invariant Borel measure on $\mathbb{H}^2$. Hence $\nu = c\,\mathrm{area}$, for some constant c. Suppose for the moment that $\mu(\mathbb{H}^2 \times B)$ is finite for some open B. Then $\mu(\mathbb{H}^2 \times B) = c'\,\mathrm{area}(B)$, and it remains to show that $c = c'$.

Let B be some open ball in $\mathbb{H}^2$. Fix a point $o \in \mathbb{H}^2$, and let η be Haar measure on $\mathrm{Isom}(\mathbb{H}^2)$. Note that $\eta\{g : z \in gB\} = \eta\{g : o \in gB\}$ when $z \in \mathbb{H}^2$. Hence, Fubini gives

$$\int \mu((gB) \times B)\, d\eta(g) = \int_{(z,g)} 1_{z \in gB}\, d\mu(\{z\} \times B)\, d\eta(g) = \mu(\mathbb{H}^2 \times B)\eta\{g : o \in gB\}.$$

Therefore,

$$\mu(\mathbb{H}^2 \times B) = \frac{\int \mu((gB) \times B)\, d\eta(g)}{\eta\{g : o \in gB\}} = \frac{\int \mu(B \times (g^{-1}B))\, d\eta(g)}{\eta\{g : o \in gB\}} = \mu(B \times \mathbb{H}^2).$$

To complete the proof for the nonnegative case, we only need to show that $\mu(\mathbb{H}^2 \times B) < \infty$ for some open B. Indeed, for every $r > 0$ let F_r be the set of $(x, y) \in \mathbb{H}^2 \times \mathbb{H}^2$ such that $d(x, y) < r$. Set $\mu_r(K) = \mu(K \cap F_r)$. Then μ_r is diagonally invariant and the above proof applies to it. Therefore,

$$\mu(\mathbb{H}^2 \times B) = \sup_{r>0} \mu_r(\mathbb{H}^2 \times B) = \sup_{r>0} \mu_r(B \times \mathbb{H}^2) = \mu(B \times \mathbb{H}^2).$$

This completes the proof in the nonnegative case, and the signed version easily follows by decomposing the measure as a difference of two nonnegative measures. □

Of course, there is nothing special about the hyperbolic plane in this case; there is a similar version of the Mass Transport Principle in any symmetric space.

Also, the Mass Transport Principle holds when the assumption that μ is invariant under the diagonal action of $\mathrm{Isom}(\mathbb{H}^2)$ is replaced by the weaker assumption that μ is invariant under the diagonal action of $\mathrm{Isom}_+(\mathbb{H}^2)$, the group of orientation preserving isometries. Similarly, the results stated below assuming invariance under $\mathrm{Isom}(\mathbb{H}^2)$ are equally valid assuming $\mathrm{Isom}_+(\mathbb{H}^2)$-invariance.

We now illustrate the Mass Transport Principle with an application involving a relation between the densities of vertices, edges, and faces in tilings.

Let T be a random tiling of $\mathbb{H}^2$, whose law is invariant under $\mathrm{Isom}(\mathbb{H}^2)$, and with the property that a.s. the edges in T are piecewise smooth, and each face and each vertex has a finite number of edges incident with it.

Definition 5.3 (Vertex, face, and edge densities)**.** How does one measure the abundance of vertices in T? Given a measurable set $A \subset \mathbb{H}^2$, let $N_v(A)$ be the number of vertices of T in A. If $\mathbf{E}N_v(A)$ is finite for every measurable bounded $A \subset \mathbb{H}^2$, then we say that T has *finite vertex density*. In this case, $\mu(A) := \mathbf{E}N_v(A)$ is clearly an invariant Borel measure; hence there is a constant D_{V} such that $\mathbf{E}N_v = D_{\mathsf{V}}$ area. This number D_{V} will be called the *vertex density* of T. The definitions of the face and edge densities are a bit more indirect. Fix a point $o \in \mathbb{H}^2$. The *face density* is defined by $D_{\mathsf{F}} = \mathbf{E}[A_o^{-1}]$, where A_o is the area of the tile of T that contains o. Given a measurable set $A \subset \mathbb{H}^2$, let $Y_e(A)$ be the sum of the degrees of the vertices of T in A. If $\mathbf{E}Y_e(A)$ is finite for bounded measurable A, and $\mathbf{E}Y_e = 2D_{\mathsf{E}}$ area, then we say that T has *finite edge density* D_{E}. It is left to the reader to convince herself or himself that these definitions are reasonable.

Given a set $B \subset \mathbb{H}^2$ with reasonably smooth boundary (piecewise C^2 is enough), let $\kappa_{\partial B}$ denote the associated curvature measure. That is, for all open $A \subset \mathbb{H}^2$, $\kappa_{\partial B}(A)$ is the (signed) curvature of $A \cap \partial B$. Note that if p is a point on ∂B and the internal angle of B at p is α, then $\kappa_{\partial B}(\{p\}) = \pi - \alpha$. The following instance of the Gauss Bonnet Theorem will be very useful for us.

Theorem 5.4 (Gauss-Bonnet in $\mathbb{H}^2$)**.** *Let A be a closed topological disk in $\mathbb{H}^2$, whose boundary is a piecewise smooth simple closed path. Then*

$$2\pi + \mathrm{area}(A) = \kappa_{\partial A}(\partial A).$$

Suppose that $K \subset \mathbb{H}^2$ is a compact set whose boundary ∂K is a 1-manifold. If ∂K has finite unsigned curvature, namely $|\kappa_{\partial K}|(\mathbb{H}^2) < \infty$, then set

$$\mu_K(A \times B) := \frac{\mathrm{area}(A \cap K)}{\mathrm{area}(K)} \kappa_{\partial K}(B), \tag{5.1}$$

for every measurable $A, B \subset \mathbb{H}^2$. This is a signed measure on $\mathbb{H}^2 \times \mathbb{H}^2$, and will be very useful below. Note that $\mu_K(\mathbb{H}^2 \times B) = \kappa_{\partial K}(B)$.

Given a tiling T of $\mathbb{H}^2$, let $\mathsf{F}(T)$ denote the set of faces (that is, tiles) of T, and set

$$\mu_T = \sum_{f \in \mathsf{F}(T)} \mu_f, \qquad \widehat{\mu}_T = \sum_{f \in \mathsf{F}(T)} |\mu_f|.$$

If T is a random tiling such that

$$\mathbf{E}\widehat{\mu}_T(\mathbb{H}^2 \times A) < \infty$$

for bounded open sets A, then we say that T has *locally integrable curvature*.

Theorem 5.5 (Euler formula for random tilings in $\mathbb{H}^2$)**.** *Let T be a random tiling of $\mathbb{H}^2$, whose distribution is $\mathrm{Isom}(\mathbb{H}^2)$-invariant. Suppose that a.s. each face of T is a closed topological disk with piecewise smooth boundary, and each vertex has degree at least 3. Further suppose that T has locally integrable curvature. Then T has finite vertex, edge, and face densities, and these densities satisfy the relation*

$$2\pi(D_{\mathsf{F}} - D_{\mathsf{E}} + D_{\mathsf{V}}) = -1. \tag{5.2}$$

Proof. Suppose that γ is an arc on the boundary of two tiles, f and f', of T, and that γ is disjoint from the vertices of T. Then $\kappa_f(\gamma) = -\kappa_{f'}(\gamma)$, as the curvature negates under a change of orientation. Consequently, if $A \subset \mathbb{H}^2$ is disjoint from the vertices of T, then $\mu_T(\mathbb{H}^2 \times A) = 0$. On the other hand, if v is a vertex of a face f, then $\mu_f(\mathbb{H}^2 \times \{v\})$ is equal to π minus the interior angle of f at v. Consequently, $\mu_T(\mathbb{H}^2 \times \{v\}) = \pi(\deg_v -2)$. Since T has locally integrable curvature and every vertex has degree at least 3, this shows that T has finite vertex and edge density, and that

$$\mathbf{E}\mu_T(\mathbb{H}^2 \times A) = 2\pi(D_{\mathsf{E}} - D_{\mathsf{V}}) \operatorname{area}(A).$$

By the Mass Transport Principle,

$$\mathbf{E}\mu_T(A \times \mathbb{H}^2) = 2\pi(D_{\mathsf{E}} - D_{\mathsf{V}}) \operatorname{area}(A).$$

On the other hand, Gauss-Bonnet shows that when A is contained in a face $f \in \mathsf{F}(T)$, we have $\mu_T(A \times \mathbb{H}^2) = \operatorname{area}(A)(2\pi + \operatorname{area}(f))/\operatorname{area}(f)$. Consequently,

$$\operatorname{area}(A)(2\pi D_{\mathsf{F}} + 1) = \mathbf{E}\mu_T(A \times \mathbb{H}^2) = 2\pi(D_{\mathsf{E}} - D_{\mathsf{V}}) \operatorname{area}(A).$$

□

Remark 5.6. As the proof shows, the -1 on the right side of (5.2) comes from the curvature of $\mathbb{H}^2$. An analogous equation holds in $\mathbb{R}^2$ and in S^2, with the -1 changed to 0 and 1, respectively. (For S^2 this is just the classical Euler formula.) A similar proof applies to S^2 and $\mathbb{R}^2$, but for these spaces one can also replace the use of the Mass Transport Principle with amenability.

Remark 5.7. When a.s. all the vertices in T have degree 3, we have $2D_{\mathsf{E}} = 3D_{\mathsf{V}}$, and therefore (5.2) simplifies to

$$D_{\mathsf{V}} = 2D_{\mathsf{F}} + \frac{1}{\pi}. \tag{5.3}$$

6. Voronoi percolation in the hyperbolic plane

We now describe a very natural continuous percolation process in the hyperbolic plane.

Given a discrete nonempty set of points $X \subset \mathbb{H}^2$, the associated *Voronoi tiling* of $\mathbb{H}^2$ is defined by $T = T(X) = \{T_x : x \in X\}$, where

$$T_x := \{y \in \mathbb{H}^2 : \operatorname{dist}(y, x) = \operatorname{dist}(y, X)\}.$$

The point $x \in X$ is called the nucleus of T_x. Fix some parameters $p \in [0, 1]$ and $\lambda \geq 0$. Let W be a Poisson point process in $\mathbb{H}^2$ with intensity $\lambda_W := p\lambda$, and let B be an independent Poisson point process with intensity $\lambda_B := (1 - p)\lambda$. Note that $X := W \cup B$ is a Poisson point process with intensity λ, and, given X, each point $x \in X$ is in W with probability p, independently. Let $T = T(X)$ be the Voronoi tiling associated with X, and set

$$\widehat{W} := \bigcup_{w \in W} T_w, \qquad \widehat{B} := \bigcup_{b \in B} T_b.$$

Observe that a.s. each vertex of the tiling $T(X)$ has degree 3. A.s. $\mathbb{H}^2$ is the union of $\widehat{W}$ and $\widehat{B}$, and $\widehat{W} \cap \widehat{B}$ is a 1-manifold. This model will be referred to as (p, λ)-*Poisson-Voronoi-Bernoulli* percolation in $\mathbb{H}^2$, or just Voronoi percolation, for short. The connected components of $\widehat{W}$ and of $\widehat{B}$ will be called *clusters*.

It is clear that if $\widehat{W}$ has infinite components with positive probability, then it has infinite components a.s. Set

$$\mathcal{W} := \{(p,\lambda) \in [0,1] \times (0,\infty) : \widehat{W} \text{ has infinite components a.s.}\},$$
$$\mathcal{B} := \{(p,\lambda) \in [0,1] \times (0,\infty) : \widehat{B} \text{ has infinite components a.s.}\}.$$

For every $\lambda > 0$ define

$$p_c(\lambda) := \inf\{p \in [0,1] : (p,\lambda) \in \mathcal{W}\}.$$

It is clear that $(p,\lambda) \in \mathcal{W}$ if $p > p_c(\lambda)$.

Proof of Theorem 1.4. It is easy to adapt the proof of Theorem 3.7 to this setting. The details are left to the reader. □

As we shall see, $[0,1] \times (0,\infty) = \mathcal{W} \cup \mathcal{B}$, so Theorem 1.4 covers all possibilities.

It is easy to see that in $\mathcal{W} \cap \mathcal{B}$ a.s. all unbounded components of $\widehat{W}$ have a cantor set of limit points in $\partial\mathbb{H}^2$, with dimension smaller than 1.

Remark 6.1. One can easily show that the face density D_{F} of T is λ, using the following mass transport. For each tile f of T with nucleus x let $\nu_f(A \times A') = \operatorname{area}(A \cap f)/\operatorname{area}(f)$ if $x \in A'$ and 0 otherwise. Set $\nu_T = \sum_f \nu_f$. Then the Mass Transport Principle with $\mathbf{E}\nu_T$ gives $\lambda = D_{\mathsf{F}}$.

Theorem 6.2. *$(1/2,\lambda) \in \mathcal{W} \cap \mathcal{B}$ for every $\lambda \in (0,\infty)$.*

We present two proofs of this theorem, one uses the Mass Transport Principle, and the other uses hyperbolic surfaces. We start with the latter.

Hyperbolic Surfaces Proof. Fix some large $d > 0$. Let S be a compact hyperbolic surface, whose injectivity radius is greater than $5d$; that is, any disk of radius $5d$ in S is isometric to a disk in the hyperbolic plane. It is well known that such surfaces exist. (See, e.g., Proposition 1 and Lemma 2 from [SS97].)

Consider $(1/2,\lambda)$ percolation in S. Let K be the union of all white or black clusters of diameter less than d. We claim that each component of K has diameter less than d. Indeed, if A is a white or black cluster with diameter less than d, then A is contained in a disk in S which is isometric to a disk in the hyperbolic plane. It follows that the complement of A consists of one component of diameter greater than d, and possibly several components of diameter smaller than d. So, if a black and a white cluster have diameters $< d$ and are adjacent, then one of them 'surrounds' the other, in the sense that the latter is contained in a component of the complement of the first which has diameter $< d$. It follows that each component of K indeed has diameter $< d$.

For all $t > 0$ let K_t be the set of points in K with distance at least t to $S - K$. Because each component of K is isometric to a set in the hyperbolic plane, the linear isoperimetric inequality for the hyperbolic plane implies that

$$\operatorname{length}(\partial K_\delta) \geq c \operatorname{area} K_\delta$$

holds for all $\delta \geq 0$, where $c > 0$ is some fixed constant. Consequently,

$$\begin{aligned}\operatorname{area}(S) - \operatorname{area}(K_1) \geq \operatorname{area}(K - K_1) &= \int_0^1 \left|\frac{d}{dt}\operatorname{area}(K_t)\right| dt \\ &= \int_0^1 \operatorname{length} \partial K_t \, dt \geq c \int_0^1 \operatorname{area}(K_t)\, dt \geq c \operatorname{area}(K_1),\end{aligned}$$

which implies

$$\operatorname{area}(K_1) \le (1+c)^{-1} \operatorname{area}(S).$$

Therefore, a uniform-random point in S has probability at least $1-(1+c)^{-1}$ to be within distance 1 of a cluster with diameter $\ge d$. Because the injectivity radius of S is $\ge 5d$, the same would be true for an arbitrary point in the hyperbolic plane. Letting $d \to \infty$, we see that for any fixed point in $\mathbb{H}^2$, the probability that it is within distance 1 of an unbounded cluster is at least $1-(1+c)^{-1}$. This implies that $(1/2,\lambda) \in \mathcal{W} \cup \mathcal{B}$, and hence $(1/2,\lambda) \in \mathcal{W} \cap \mathcal{B}$, by symmetry. □

Mass Transport Proof. Consider Voronoi percolation with parameters $(1/2,\lambda)$. Let $d>0$ be large, and let F_d be the union of all black or white components of diameter less than d. Then each component K of F_d is a.s. a topological disk with diameter bounded by d.

Given a component K of F_d, as above, let μ_K be the signed measure on $\mathbb{H}^2 \times \mathbb{H}^2$ defined by

$$\mu_K(A \times B) = \frac{\operatorname{area}(A \cap K)}{\operatorname{area}(K)} \kappa_{\partial K}(B),$$

where $\kappa_{\partial K}$ is the curvature measure on ∂K. Given the percolation configuration X, let

$$\mu^X := \sum_K \mu_K, \qquad \mu := \mathbf{E}\mu^X,$$

where the sum extends over all components K of F_d. Note that $|\mu^X(\mathbb{H}^2 \times A)|$ is bounded by 2π times the number of vertices of the Voronoi tiling that are in the intersection of A with the boundary of F_d. Consequently, $|\mu|(\mathbb{H}^2 \times A)$ is finite for every bounded A. The measure μ is clearly invariant under the diagonal action of $\mathrm{Isom}(\mathbb{H}^2)$ on $\mathbb{H}^2 \times \mathbb{H}^2$. Therefore, the Mass Transport Principle applies to μ.

Fix some point $o \in \mathbb{H}^2$. Fubini and the Gauss Bonnet Theorem 5.4 show that $\mu(A \times \mathbb{H}^2) \ge \operatorname{area}(A)\mathbf{P}[o \in F_d]$. By the Mass Transport Principle, we therefore have

$$\mu(\mathbb{H}^2 \times A) \ge \operatorname{area}(A)\mathbf{P}[o \in F_d]. \tag{6.1}$$

Fix some A with $\operatorname{area}(A)>0$. Let $F_\infty = \bigcup_{d>0} F_d$. From (6.1) we find that

$$2\pi\mathbf{E}|\{\text{Voronoi vertices in } A \cap \partial F_d\}| \ge \operatorname{area}(A)\mathbf{P}[o \in F_d].$$

Because $\mathbf{E}|\{\text{Voronoi vertices in } A\}| < \infty$, by letting $d \to \infty$ it follows that

$$2\pi\mathbf{E}|\{\text{Voronoi vertices in } A \cap \partial F_\infty\}| \ge \operatorname{area}(A)\mathbf{P}[o \in F_\infty].$$

This shows that ∂F_∞ is not empty with positive probability. On this event, $\widehat{W}$ has an unbounded component or $\widehat{B}$ has an unbounded component. This gives $(1/2,\lambda) \in \mathcal{W} \cup \mathcal{B}$. Symmetry then implies that $(1/2,\lambda) \in \mathcal{W} \cap \mathcal{B}$, which completes the proof. □

Note that the latter proof does not require that the tile colors be independent. In fact, it gives the following generalization.

Theorem 6.3. *Suppose that $Z \subset \mathbb{H}^2$ is a closed random subset whose distribution is $\mathrm{Isom}(\mathbb{H}^2)$-invariant, such that ∂Z is a.s. a 1-manifold and $\mathbf{E}[|\kappa_{\partial Z}|(A)] < \infty$ for some nonempty open $A \subset \mathbb{H}^2$. Then a.s. there is an infinite component in Z or in $\mathbb{H}^2 \backslash Z$.* □

We now work a bit harder to get the following explicit upper bound on $p_c(\lambda)$. The bound will also show that $p_c(\lambda) \to 0$ as $\lambda \searrow 0$.

Theorem 6.4.

$$\forall \lambda > 0, \quad p_c(\lambda) \le \frac{1}{2} - \frac{1}{4\lambda\pi + 2}.$$

Note that the right hand side is zero at $\lambda = 0$ and its derivative there is π.

Proof. Let $\lambda > 0$, and let $p < p_c(\lambda)$. Then $p < 1/2$, by Theorem 6.2. Consider three independent Poisson point processes W, B, R in $\mathbb{H}^2$, with (positive) intensities $p\lambda$, $p\lambda$, and $(1-2p)\lambda$, respectively. Then the intensity of $W \cup B \cup R$ is λ. Let $X = (W, B, R)$, and consider the Voronoi tiling $T = T(W \cup B \cup R)$. Color the tiles with white, black, or red, depending on whether the nucleus is in W, B, or R, respectively. Let

$$p_R := 1 - 2p, \quad p_W := p, \quad p_B := p.$$

Then, given $W \cup B \cup R$, p_r, p_W, and p_B are the probabilities that any given tile of $T(W \cup B \cup R)$ is red, white, and black, respectively.

Because $p < p_c$, a.s. there are no unbounded white or black clusters. Given a cluster K, let the *hull* of K be the union of K with the bounded components of $\mathbb{H}^2 - K$. Call a white or black cluster K an *empire* if there is no black or white cluster K' such that the hull of K' contains K. Let $\mathcal{K}$ be the set of all hulls of empires.

Condition on the triplet $X = (W, B, R)$, and let $K \in \mathcal{K}$. Let $\kappa_{\partial K}$ denote the curvature measure on ∂K. Let μ_K be the signed measure on $\mathbb{H}^2 \times \mathbb{H}^2$ defined by

$$\mu_K(A \times A') = \frac{\text{area}(A \cap K)}{\text{area}(K)} \kappa_{\partial K}(A'),$$

and let

$$\mu^X = \sum_{K \in \mathcal{K}} \mu_K, \qquad \mu = \mathbf{E}\mu^X,$$

as above. As before, it is not hard to verify that $|\mu|(\mathbb{H}^2 \times A)$ is finite for every bounded measurable A.

Fix a point $o \in \mathbb{H}^2$. The Gauss Bonnet Theorem 5.4 shows that

$$\mu(A \times \mathbb{H}^2) \ge \mathbf{P}\left[o \in \bigcup \mathcal{K}\right] \text{area}(A). \tag{6.2}$$

Note that the measure $\mu^X(\mathbb{H}^2 \times \cdot)$ on $\mathbb{H}^2$ is supported on the vertices of the Voronoi tiling that are on the outer boundaries of the empires. Let z be such a vertex. Note that if z does not belong to a red tile, then z is in the interior of the union of two empires. In that case, $\mu^X(\mathbb{H}^2 \times \{z\}) = 0$, since the contributions to $\mu^X(\mathbb{H}^2 \times \{z\})$ from both empires cancel. Consequently, $\mu^X(\mathbb{H}^2 \times \cdot)$ is supported on the vertices that are on the boundaries of red tiles. Suppose that v is a Voronoi vertex and there are three tiles T_1, T_2, T_3 meeting at v, and α_1, α_2, α_3 are their angles at v, respectively. Since the Voronoi tiles are convex, we have $\alpha_1, \alpha_2, \alpha_3 \in (0, \pi]$.

If T_1 is red, T_2 is white, and T_3 is black, then $\mu^X(\mathbb{H}^2 \times \{v\}) = \pi - \alpha_2 + \pi - \alpha_3 = \alpha_1$ or $\mu^X(\mathbb{H}^2 \times \{v\}) = 0$ (the latter happens if $v \notin \partial \cup \mathcal{K}$). If T_1 is red, and T_2 and T_3 are both of the same color, then $\mu^X(\mathbb{H}^2 \times \{v\}) = \pi - \alpha_2 - \alpha_3 = \alpha_1 - \pi \le 0$ or $\mu^X(\mathbb{H}^2 \times \{v\}) = 0$. If T_1 and T_2 are red, then $\mu^X(\mathbb{H}^2 \times \{v\}) = \pi - \alpha_3 = \alpha_1 + \alpha_2 - \pi$

or $\mu^X(\mathbb{H}^2 \times \{v\}) = 0$. If there is no red tile among T_1, T_2, T_3 or if all of them are red, then $\mu^X(\mathbb{H}^2 \times \{v\}) = 0$. Consequently, given the tiling (but not the colors), the expected value of $\mu^X(\mathbb{H}^2 \times \{v\})$ is bounded by

$$2p_R p_B p_W(\alpha_1 + \alpha_2 + \alpha_3) + p_R^2(1 - p_R)(2\alpha_1 + 2\alpha_2 + 2\alpha_3 - 3\pi) = 2\pi p(1 - 2p).$$

Hence,

$$\mu(\mathbb{H}^2 \times A) \leq D_{\mathsf{V}} 2\pi p(1 - 2p)\,\mathrm{area}(A).$$

Using the Mass Transport Principle and (6.2), this gives

$$\mathbf{P}\left[o \in \bigcup \mathcal{K}\right] \leq D_{\mathsf{V}} 2\pi p(1 - 2p). \tag{6.3}$$

We now show

$$\mathbf{P}\left[o \in \bigcup \mathcal{K}\right] \geq p_B + p_W = 2p. \tag{6.4}$$

For this, we prove that every white or black tile is contained in the hull of an empire. If not, there is a sequence of black or white clusters $K_1, K_2, \ldots$ such that each K_j is contained in the hull of K_{j+1}. If infinitely many of these clusters are white, say, then when all the red tiles are changed to black, there is still no unbounded black cluster. However, $p_R + p_B > 1/2 \geq p_c(\lambda)$, which is a contradiction. A similar contradiction is obtained if infinitely many of the clusters K_j are black. Hence (6.4) holds.

Since $D_{\mathsf{F}} = \lambda$, by Remark 6.1, combining (6.3), (6.4), and (5.3) gives

$$(2\lambda\pi + 1)(1 - 2p) \geq 1.$$

This inequality must be satisfied for every $p < p_c(\lambda)$, which proves the theorem. □

Lemma 6.5.

$$\forall \lambda > 0, \quad p_c(\lambda) > 0.$$

Proof. Let Q be a set of points in the hyperbolic plane which is maximal with the property that the distance between any two points in Q is at least 1. Then every open ball of radius 1 contains a point in Q.

Consider Voronoi percolation with some parameters (p, λ). As always, let W and B denote the Poisson point processes of the white and black nuclei, respectively. Fix some large $R > 0$. Given $y \in \mathbb{H}^2$, consider the Voronoi tiling $T(y)$ with nuclei $W \cup \{y\} \cup B$, and let $\widehat{W}(y)$ denote the union of the tiles of $T(y)$ with nuclei in $W \cup \{y\}$. Let $\mathcal{A}(y)$ be the event that some component of $\widehat{W}(y)$ intersects the hyperbolic circle of radius R with center y and the hyperbolic circle of radius 1 with center y. The reason for introducing $T(y)$ and $\widehat{W}(y)$ is that the events $\mathcal{A}(y)$ and $\mathcal{A}(y')$ are independent when $\mathrm{dist}(y, y') > 4R$, because $\mathcal{A}(y)$ depends only on the intersections of B and W with the closed ball of radius $2R$ about y. The analogous property does not hold for $\widehat{W}$ in place of $\widehat{W}(y)$. (That is, the event that there is a component of $\widehat{W}$ which intersects the circles of radii 1 and R about y is not independent from the corresponding event for y', even if the distance between y and y' is large.)

Let $\alpha(R)$ be the probability that the tile S with nucleus y in $T(y)$ intersects the circle of radius R about y. We now estimate $\alpha(R)$ from above. Indeed, if S intersects this circle at a point z, then the open ball of radius R about z does not contain any point in $B \cup W$. Then there is a point $x \in Q$ with $\mathrm{dist}(x, z) < 1$ such

that the ball $B(x, R-1)$ does not intersect $B \cup W$. For a given x, the probability for that is $\exp(-\lambda \operatorname{area}(B(x, R-1)))$, which is less than $\exp(-c\lambda e^R)$ for some constant $c > 0$ and all R sufficiently large (recall that $\operatorname{area}(B(x, R))/e^R$ tends to a positive constant as $R \to \infty$). Consequently,

(6.5)
$$\begin{aligned}\alpha(R) &\le |B(y, R+1) \cap Q| \exp(-c\lambda e^R) \le O(1) \operatorname{area}(B(y, R+1)) \exp(-c\lambda e^R) \\ &\le O(1) \exp(-c'\lambda e^R),\end{aligned}$$

for large R. Clearly, $\mathbf{P}[\mathcal{A}(y)] \ge \alpha(R)$. However, there is some very small $p(R) > 0$ such that $\mathbf{P}[\mathcal{A}(y)] \le 2\alpha(R)$ if $p < p(R)$.

Assume that $p < p(R)$. Let $o \in \mathbb{H}^2$ be some basepoint, and let $\widehat{W}_o$ be the component of o in $\widehat{W}$ (set $\widehat{W}_o = \varnothing$ if $o \notin \widehat{W}$). Assuming that $\widehat{W}_o$ is unbounded, there is a.s. a path $\gamma : [0, \infty) \to \widehat{W}_o$ starting at $\gamma(0) = o$ and with $\operatorname{dist}(\gamma(t), o) \to \infty$ as $t \to \infty$. In this case, let $t_0 := 0$, $y_0 := o$, and inductively set $t_n := \sup\{t : \operatorname{dist}(\gamma(t), y_{n-1}) = 5R\}$ and $y_n := \gamma(t_n)$. Let y'_n be a point in Q closest to y_n. Then for each n, $\operatorname{dist}(y'_n, y'_{n+1}) \le 5R + 2$, and for each $j \ne k$, $\operatorname{dist}(y'_k, y'_j) \ge 5R - 2$. Observe that the events $\mathcal{A}(y'_k)$ all hold. Consequently, for every $n = 1, 2, \ldots$, we may bound the event that o is in an unbounded component of $\widehat{W}$ by

$$\sum_{x_0, x_1, \ldots, x_n} \mathbf{P}[\mathcal{A}(x_0) \cap \mathcal{A}(x_1) \cap \cdots \cap \mathcal{A}(x_n)],$$

where the sum is over all sequences $x_0, x_1, \ldots, x_n$ in Q such that x_0 is within distance 1 of o, each x_j is within distance $5R + 2$ of x_{j-1}, and $\operatorname{dist}(x_j, x_k) \ge 5R - 2$ when $j \ne k$. Assuming $R > 2$, these events $\mathcal{A}(x_0), \mathcal{A}(x_1), \ldots, \mathcal{A}(x_n)$ are independent, and we get the bound

$$\mathbf{P}[\widehat{W}_o \text{ is unbounded}] \le (2\alpha(R))^{n+1} |\{\text{such sequences } x_0, \ldots, x_n\}|. \tag{6.6}$$

Now, the number of such sequences is at most

$$\max\{|Q \cap B(z, 5R+2)|^n : z \in \mathbb{H}^2\}$$

times the number of possible choices of x_0. This is at most $\exp(c_1 R n)$, for some constant c_1. By our estimate (6.5) for $\alpha(R)$, it is clear that there is some large $R_0 > 0$ such that the right hand side of (6.6) goes to zero as $n \to \infty$. Then for $p < p(R_0)$ the probability that $\widehat{W}_o$ is unbounded is zero, and hence $p_c(\lambda) \ge p(R_0)$. □

Lemma 6.6. *$p_c(\lambda)$ is continuous on $(0, \infty)$.*

Proof. Note that given $\lambda, h > 0$, a union of two independent Poisson point processes with intensities λ and h is a Poisson point process with intensity $\lambda + h$. Fix some $o \in \mathbb{H}^2$, and let θ be the probability that o is in an unbounded component of $\widehat{W}$. Consider $\theta = \theta(\lambda_W, \lambda_B)$ as a function of $\lambda_W = p\lambda$ and $\lambda_B = (1-p)\lambda$. It is clear that θ is monotone increasing (weakly) in λ_W and monotone decreasing (weakly) in λ_B. Consequently, if $\lambda' > \lambda$, we must have $p_c(\lambda')\lambda' \ge p_c(\lambda)\lambda$ and $(1 - p_c(\lambda'))\lambda' \ge (1 - p_c(\lambda))\lambda$. Hence,

$$p_c(\lambda)\frac{\lambda}{\lambda'} \le p_c(\lambda') \le 1 - (1 - p_c(\lambda))\frac{\lambda}{\lambda'},$$

which implies continuity. □

Proof of Theorem 1.5. Part (a) follows from Theorem 6.4 and Lemma 6.5. Part (b) follows from (a). Part (c) is Lemma 6.6, above. The proof from [BLPS99b] of Theorem 1.3 can easily be adapted to prove (d). □

7. Open problems

Question 7.1. In the absence of planarity, most of the proofs in this paper are invalid. Which results can be extended to transitive graphs that are quasi-isometric to the hyperbolic plane, but are not planar?

Conjecture 7.2. $\lim_{\lambda\to\infty} p_c(\lambda) = 1/2$.

Question 7.3. What are the asymptotics of $p_c(\lambda)$ as $\lambda \to \infty$ and as $\lambda \to 0$? In particular, what is $p_c'(0)$?

Conjecture 7.4. *$p_c(\lambda)$ is strictly monotone increasing.*

Given $a > 0$, let $a\mathbb{H}^2$ denote $\mathbb{H}^2$ with the metric scaled by a. It is clear that (p, λ)-Voronoi percolation on $\mathbb{H}^2$ is the same as $(p, 1)$-Voronoi percolation on $\sqrt{\lambda}\mathbb{H}^2$. As $a \to \infty$, the space $a\mathbb{H}^2$ looks more and more like $\mathbb{R}^2$ ($a\mathbb{H}^2$ has constant curvature $-1/a^2$). This leads to the following question.

Question 7.5. It is known that for Voronoi percolation in the Euclidean plane $p_c \geq 1/2$ [Zva96] (in the Euclidean setting, p_c clearly does not depend on λ). However, the conjecture $p_c = 1/2$ is still open. Lacking is a proof that there are unbounded components at $p > 1/2$. Is it possible to prove that $p_c = 1/2$ in the Euclidean setting, by taking a limit as $\lambda \to \infty$ in the hyperbolic setting?

This direction can also lead to a plausible guess as to the asymptotics of $p_c(\lambda)$ as $\lambda \to \infty$. Recall the notion of the correlation length $\xi(p)$ (see [Gri89]), which roughly measures the length scale at which Bernoulli(p) percolation is substantially different from Bernoulli(p_c) percolation. Similarly, at lengths on the order of a or higher, the space $a\mathbb{H}^2$ appears substantially different from $\mathbb{R}^2$. On balls of size smaller than a, $(p, 1)$-Voronoi percolation on $a\mathbb{H}^2$ does not look too different from $(p, 1)$-Voronoi percolation on $\mathbb{R}^2$. This suggests that if $\xi(p)$ is significantly smaller than a (here $\xi(p)$ is the correlation length for $(p, 1)$-Voronoi percolation on $\mathbb{R}^2$), then there will be no unbounded white clusters for $(p, 1)$-Voronoi percolation on $a\mathbb{H}^2$. Conversely, if $\xi(p)$ is significantly larger than a, the hyperbolicity of $a\mathbb{H}^2$ appearing on the scale of a will help to create unbounded white clusters. Since $(p, 1)$-Voronoi percolation on $\sqrt{\lambda}\mathbb{H}^2$ is the same as (p, λ)-Voronoi percolation on $\mathbb{H}^2$, this suggests that $\xi(p_c(\lambda))$ should grow roughly at the same rate as $\sqrt{\lambda}$ when $\lambda \to \infty$. It is conjectured that $\xi(p)$ grows like $|p - p_c|^{-4/3}$, which leads to the guess that $p_c(\lambda)$ is asymptotic to $\frac{1}{2} - \lambda^{-2/3}$.

Acknowledgement

We wish to express gratitude to Harry Kesten, Russ Lyons, Bojan Mohar, Igor Pak, Yuval Peres, Roberto Schonmann, and Bill Thurston for fruitful conversations and useful advice.

References

[Bab97] L. Babai, *The growth rate of vertex-transitive planar graphs*, in Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms (New Orleans, LA, 1997), ACM, New York, 1997, pp. 564–573. MR **97k:**68011

[BK89] R. M. Burton and M. Keane, *Density and uniqueness in percolation*, Comm. Math. Phys. **121** (1989), no. 3, 501–505. MR **90g:**60090

[BK91] ———, *Topological and metric properties of infinite clusters in stationary two-dimensional site percolation*, Israel J. Math. **76** (1991), no. 3, 299–316. MR **93i:**60182

[BLPS99a] I. Benjamini, R. Lyons, Y. Peres, and O. Schramm, *Group-invariant percolation on graphs*, Geom. Funct. Anal. **9** (1999), no. 1, 29–66. MR **99m:**60149

[BLPS99b] ———, *Critical percolation on any nonamenable group has no infinite clusters*, Ann. Probab. **27** (1999), no. 3, 1347–1356. MR **2000k:**60197

[BLPS00] ———, *Uniform spanning forests*, Ann. Probab. (2000), to appear.

[BLS99] I. Benjamini, R. Lyons, and O. Schramm, *Percolation perturbations in potential theory and random walks*, in M. Picardello and W. Woess, editors, Random Walks and Discrete Potential Theory, Sympos. Math., Cambridge University Press, Cambridge, 1999, pp. 56–84. Papers from the workshop held in Cortona, 1997.

[BS96] I. Benjamini and O. Schramm, *Percolation beyond $\mathbb{Z}^d$, many questions and a few answers*, Electron. Comm. Probab. **1** (1996), no. 8, 71–82 (electronic). MR **97j:**60179

[BS99] ———, *Recent progress on percolation beyond $\mathbb{Z}^d$*, `http://www.wisdom.weizmann.ac.il/~schramm/papers/pyond-rep/`, 1999.

[BSt90] A. F. Beardon and K. Stephenson, *The uniformization theorem for circle packings*, Indiana Univ. Math. J. **39** (1990), no. 4, 1383–1425. MR **92b:**52038

[CFKP97] J. W. Cannon, W. J. Floyd, R. Kenyon, and W. R. Parry, *Hyperbolic geometry*, in Flavors of Geometry, Cambridge University Press, Cambridge, 1997, pp. 59–115. MR **99c:**57036

[DS84] P. G. Doyle and J. L. Snell, *Random walks and electric networks*, Mathematical Association of America, Washington, DC, 1984. MR **89a:**94023

[GN90] G. R. Grimmett and C. M. Newman, *Percolation in $\infty + 1$ dimensions*, in G. R. Grimmett and D. J. A. Welsh, editors, Disorder in Physical Systems, Oxford University Press, New York, 1990, pp. 167–190. MR **92a:**60207

[Gri89] G. Grimmett, *Percolation*, Springer-Verlag, New York, 1989. MR **90j:**60109

[Häg97] O. Häggström, *Infinite clusters in dependent automorphism invariant percolation on trees*, Ann. Probab. **25** (1997), no. 3, 1423–1436. MR **98f:**60207

[HP99] O. Häggström and Y. Peres, *Monotonicity of uniqueness for percolation on Cayley graphs: all infinite clusters are born simultaneously*, Probab. Theory Related Fields **113** (1999), no. 2, 273–285. MR **99k:**60253

[HS95] Z.-X. He and O. Schramm, *Hyperbolic and parabolic packings*, Discrete Comput. Geom. **14** (1995), no. 2, 123–149. MR **96h:**52017

[Imr75] W. Imrich, *On Whitney's theorem on the unique embeddability of 3-connected planar graphs*, in Recent Advances in Graph Theory (Proc. Second Czechoslovak Sympos., Prague, 1974), Academia, Prague, 1975, pp. 303–306. MR **52:**5462

[Lal98] S. P. Lalley, *Percolation on Fuchsian groups*, Ann. Inst. H. Poincaré Probab. Statist. **34** (1998), no. 2, 151–177. MR **99g:**60190

[Lyo00] R. Lyons, *Phase transitions on nonamenable graphs*, J. Math. Phys. **41** (2000), no. 3, 1099–1126. Probabilistic techniques in equilibrium and nonequilibrium statistical physics. CMP 2000:13

[Lyo01] ———, *Probability on trees and networks*, Cambridge University Press, 2001. Written with the assistance of Y. Peres, in preparation. Current version available at `http://php.indiana.edu/~rdlyons/`.

[Mad70] W. Mader, *Über den Zusammenhang symmetrischer Graphen*, Arch. Math. (Basel) **21** (1970), 331–336. MR **44:**6534

[MR96] R. Meester and R. Roy, *Continuum percolation*, Cambridge University Press, Cambridge, 1996. MR **98d:**60193

[PSN00] I. Pak and T. Smirnova-Nagnibeda, *On non-uniqueness of percolation on nonamenable Cayley graphs*, C. R. Acad. Sci. Paris Sér. I Math. **330** (2000), no. 6, 495–500. MR **2000m:**60116

[Per00] Y. Peres, *Percolation on nonamenable products at the uniqueness threshold*, Ann. Inst. H. Poincaré Probab. Statist. **36** (2000), no. 3, 395–406. CMP 2000:15

[Sch99a] R. H. Schonmann, *Percolation in $\infty+1$ dimensions at the uniqueness threshold*, in M. Bramson and R. Durrett, editors, Perplexing Probability Problems: Papers in Honor of Harry Kesten, Birkhäuser, Boston, 1999, pp. 53–67. CMP 99:16

[Sch99b] ———, *Stability of infinite clusters in supercritical percolation*, Probab. Theory Related Fields **113** (1999), no. 2, 287–300. MR **99k:**60252

[SS97] P. Schmutz Schaller, *Extremal Riemann surfaces with a large number of systoles*, in Extremal Riemann Surfaces (San Francisco, CA, 1995), Amer. Math. Soc., Providence, RI, 1997, pp. 9–19. MR **98c:**11064

[Wat70] M. E. Watkins, *Connectivity of transitive graphs*, J. Combinatorial Theory **8** (1970), 23–29. MR **42:**1707

[Zva96] A. Zvavitch, *The critical probability for Voronoi percolation*, MSc. thesis, Weizmann Institute of Science, 1996.

Department of Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel
E-mail address: itai@wisdom.weizmann.ac.il
URL: http://www.wisdom.weizmann.ac.il/~itai/

Department of Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel
E-mail address: schramm@microsoft.com

Annals of Mathematics, **160** (2004), 465–491

Geometry of the uniform spanning forest: Transitions in dimensions $4, 8, 12, \ldots$

By Itai Benjamini, Harry Kesten, Yuval Peres, and Oded Schramm*

Abstract

The uniform spanning forest (USF) in $\mathbb{Z}^d$ is the weak limit of random, uniformly chosen, spanning trees in $[-n, n]^d$. Pemantle [11] proved that the USF consists a.s. of a single tree if and only if $d \leq 4$. We prove that any two components of the USF in $\mathbb{Z}^d$ are adjacent a.s. if $5 \leq d \leq 8$, but not if $d \geq 9$. More generally, let $N(x, y)$ be the minimum number of edges outside the USF in a path joining x and y in $\mathbb{Z}^d$. Then

$$\max\{N(x, y) : x, y \in \mathbb{Z}^d\} = \lfloor (d-1)/4 \rfloor \text{ a.s.}$$

The notion of stochastic dimension for random relations in the lattice is introduced and used in the proof.

1. Introduction

A *uniform spanning tree* (UST) in a finite graph is a subgraph chosen uniformly at random among all spanning trees. (A spanning tree is a subgraph such that every pair of vertices in the original graph are joined by a unique simple path in the subgraph.) The *uniform spanning forest* (USF) in $\mathbb{Z}^d$ is a random subgraph of $\mathbb{Z}^d$, that was defined by Pemantle [11] (following a suggestion of R. Lyons), as follows: The USF is the weak limit of uniform spanning trees in larger and larger finite boxes. Pemantle showed that the limit exists, that it does not depend on the sequence of boxes, and that every connected component of the USF is an infinite tree. See Benjamini, Lyons, Peres and Schramm [2] (denoted BLPS below) for a thorough study of the construction and properties of the USF, as well as references to other works on the subject. Let $T(x)$ denote the tree in the USF which contains the vertex x.

*Research partially supported by NSF grants DMS-9625458 (Kesten) and DMS-9803597 (Peres), and by a Schonbrunn Visiting Professorship (Kesten).

Key words and phrases. Stochastic dimension, Uniform spanning forest.

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_25, © Springer Science+Business Media, LLC 2011

Also define

$$N(x,y) = \min\{\text{number of edges outside the USF in a path from } x \text{ to } y\}$$

(the minimum here is over all paths in $\mathbb{Z}^d$ from x to y).

Pemantle [11] proved that for $d \le 4$, almost surely $T(x) = T(y)$ for all $x, y \in \mathbb{Z}^d$, and for $d > 4$, almost surely $\max_{x,y} N(x,y) > 0$. The following theorem shows that a.s. $\max N(x,y) \le 1$ for $d = 5, 6, 7, 8$, and that $\max N(x,y)$ increases by 1 whenever the dimension d increases by 4.

THEOREM 1.1.

$$\max\{N(x,y) : x, y \in \mathbb{Z}^d\} = \left\lfloor \frac{d-1}{4} \right\rfloor \qquad \text{a.s.}$$

Moreover, a.s. on the event $\{T(x) \ne T(y)\}$, *there exist infinitely many disjoint simple paths in* $\mathbb{Z}^d$ *which connect* $T(x)$ *and* $T(y)$ *and which contain at most* $\lfloor (d-1)/4 \rfloor$ *edges outside the* USF.

It is also natural to study

$$D(x,y) := \lim_{n\to\infty} \inf\{|u - v| : u \in T(x), v \in T(y), |u|, |v| \ge n\},$$

where $|u| = \|u\|_1$ is the l^1 norm of u. The following result is a consequence of Pemantle [11] and our proof of Theorem 1.1.

THEOREM 1.2. *Almost surely, for all* $x, y \in \mathbb{Z}^d$,

$$D(x,y) = \begin{cases} 0 & \text{if } d \le 4, \\ 1 & \text{if } 5 \le d \le 8 \text{ and } T(x) \ne T(y), \\ \infty & \text{if } d \ge 9 \text{ and } T(x) \ne T(y). \end{cases}$$

When $5 \le d \le 8$, this provides a natural example of a translation invariant random partition of $\mathbb{Z}^d$, into infinitely many components, each pair of which comes infinitely often within unit distance from each other.

The lower bounds on $N(x,y)$ follow readily from standard random walk estimates (see Section 5), so the bulk of our work will be devoted to the upper bounds.

Part of our motivation comes from the conjecture of Newman and Stein [10] that invasion percolation clusters in $\mathbb{Z}^d$, $d \ge 6$, are in some sense 4-dimensional and that two such clusters, which are formed by starting at two different vertices, will intersect with probability 1 if $d < 8$, but not if $d > 8$. A similar phenomenon is expected for minimal spanning trees on the points of a homogeneous Poisson process in $\mathbb{R}^d$. These problems are still open, as the tools presently available to analyze invasion percolation and minimal

spanning forests are not as sharp as those available for the uniform spanning forest.

In the next section the notion of stochastic dimension is introduced. A random relation $\mathcal{R} \subset \mathbb{Z}^d \times \mathbb{Z}^d$ has *stochastic dimension* $d - \alpha$, if there is some constant $c > 0$ such that for all $x \neq z$ in $\mathbb{Z}^d$,

$$c^{-1}|x - z|^{-\alpha} \leq \mathbf{P}[x\mathcal{R}z] \leq c|x - z|^{-\alpha},$$

and if a natural correlation inequality (2.2) (an upper bound for $\mathbf{P}[x\mathcal{R}z,\ y\mathcal{R}w]$) holds. The results regarding stochastic dimension are formulated and proved in this generality, to allow for future applications.

The bulk of the paper is devoted to the proof of the upper bound on $\max N(x, y)$ in (1.1). We now present an overview of this proof. Let $\mathcal{U}^{(n)}$ be the relation $N(x, y) \leq n - 1$. Then $x\mathcal{U}^{(1)}y$ means that x and y are in the same USF tree, and $x\mathcal{U}^{(2)}y$ means that $T(x)$ is equal to or adjacent to $T(y)$. We show that $\mathcal{U}^{(1)}$ has stochastic dimension 4 when $d \geq 4$. When $\mathcal{R}, \mathcal{L} \subset \mathbb{Z}^d \times \mathbb{Z}^d$ are independent relations with stochastic dimensions $\dim_S(\mathcal{R})$ and $\dim_S(\mathcal{L})$, respectively, it is proven that the composition $\mathcal{L}\mathcal{R}$ (defined by $x\mathcal{L}\mathcal{R}y$ if and only if there is a z such that $x\mathcal{L}z$ and $z\mathcal{R}y$) has stochastic dimension $\min\{\dim_S(\mathcal{R}) + \dim_S(\mathcal{L}), d\}$. It follows that the composition of $m + 1$ independent copies of $\mathcal{U}^{(1)}$ has stochastic dimension d, where m is equal to the right hand of (1.1). By proving that $\mathcal{U}^{(m+1)}$ stochastically dominates the composition of $m + 1$ independent copies of $\mathcal{U}^{(1)}$, we conclude that $\dim_S(\mathcal{U}^{(m+1)}) = d$, which implies $\inf_{x,y\in\mathbb{Z}^d} \mathbf{P}[N(x, y) \leq m] > 0$. Nonobvious tail-triviality arguments then give $\mathbf{P}[N(x, y) \leq m] = 1$ for every x and y in $\mathbb{Z}^d$, which proves the required upper bound.

In Section 4 we present the relevant USF properties needed; in particular, we obtain a tight upper bound, Theorem 4.3, for the probability that a finite set of vertices in $\mathbb{Z}^d$ is contained in one USF component. Fundamental for these results is a method from BLPS [2] for generating the USF in any transient graph, which is based on an algorithm by Wilson [15] for sampling uniformly from the spanning trees in finite graphs. (We recall this method in Section 4.)

Our main results are established in Section 5. Section 6 describes several examples of relations having a stochastic dimension, including long-range percolation, and suggests some conjectures. We note that proving $D(x, y) \in \{0, 1\}$ for $5 \leq d \leq 8$, is easier than the higher dimensional result. (The full power of Theorem 2.4 is not needed; Corollary 2.9 suffices.)

2. Stochastic dimension and compositions

Definition 2.1. When $x, y \in \mathbb{Z}^d$, we write $\langle xy\rangle := 1 + |x - y|$, where $|x-y| = \|x-y\|_1$ is the distance from x to y in the graph metric on $\mathbb{Z}^d$. Suppose that $W \subset \mathbb{Z}^d$ is finite, and τ is a tree on the vertex set W (τ need not be a

subgraph of $\mathbb{Z}^d$). Then let $\langle\tau\rangle := \prod_{\{x,y\}\in\tau}\langle xy\rangle$ denote the product of $\langle xy\rangle$ over all undirected edges $\{x,y\}$ in τ. Define the *spread* of W by $\langle W\rangle := \min_\tau\langle\tau\rangle$, where τ ranges over all trees on the vertex set W.

For three vertices, $\langle xyz\rangle = \min\{\langle xy\rangle\langle yz\rangle, \langle yz\rangle\langle zx\rangle, \langle zx\rangle\langle xy\rangle\}$. More generally, for n vertices, $\langle x_1 \dots x_n\rangle$ is a minimum of n^{n-2} products (since this is the number of trees on n labeled vertices); see Remark 2.7 for a simpler equivalent expression.

Definition 2.2 (Stochastic dimension). Let $\mathcal{R}$ be a random subset of $\mathbb{Z}^d \times \mathbb{Z}^d$. We think of $\mathcal{R}$ as a relation, and usually write $x\mathcal{R}y$ instead of $(x,y)\in\mathcal{R}$. Let $\alpha \in [0,d)$. We say that $\mathcal{R}$ has *stochastic dimension* $d-\alpha$, and write $\dim_S(\mathcal{R}) = d-\alpha$, if there is a constant $C = C(\mathcal{R}) < \infty$ such that

$$C\,\mathbf{P}[x\mathcal{R}z] \geq \langle xz\rangle^{-\alpha}, \tag{2.1}$$

and

$$\mathbf{P}[x\mathcal{R}z,\ y\mathcal{R}w] \leq C\,\langle xz\rangle^{-\alpha}\langle yw\rangle^{-\alpha} + C\,\langle xzyw\rangle^{-\alpha}, \tag{2.2}$$

hold for all $x,y,z,w\in\mathbb{Z}^d$.

Observe that (2.2) implies

$$\mathbf{P}[x\mathcal{R}z] \leq 2C\,\langle xz\rangle^{-\alpha}, \tag{2.3}$$

since we may take $x=y$ and $z=w$. Also, note that $\dim_S(\mathcal{R}) = d$ if and only if $\inf_{x,z\in\mathbb{Z}^d}\mathbf{P}[x\mathcal{R}z] > 0$.

To motivate (2.2), focus on the special case in which $\mathcal{R}$ is a random equivalence relation. Then heuristically, the first summand in (2.2) represents an upper bound for the probability that x,z are in one equivalence class and y,w are in another, while the second summand, $C\,\langle xzyw\rangle^{-\alpha}$, represents an upper bound for the probability that x,z,y,w are all in the same class. Indeed, when the equivalence classes are the components of the USF, we will make this heuristic precise in Section 4.

Several examples of random relations that have a stochastic dimension are described in Section 6. The main result of Section 4, Theorem 4.2, asserts that the relation determined by the components of the USF has stochastic dimension 4.

Definition 2.3 (Composition). Let $\mathcal{L},\mathcal{R}\subset\mathbb{Z}^d\times\mathbb{Z}^d$ be random relations. The *composition* $\mathcal{L}\mathcal{R}$ of $\mathcal{L}$ and $\mathcal{R}$ is the set of all $(x,z)\in\mathbb{Z}^d\times\mathbb{Z}^d$ such that there is some $y\in\mathbb{Z}^d$ with $x\mathcal{L}y$ and $y\mathcal{R}z$.

Composition is clearly an associative operation, that is, $(\mathcal{L}\mathcal{R})\mathcal{Q} = \mathcal{L}(\mathcal{R}\mathcal{Q})$. Our main goal in this section is to prove,

THEOREM 2.4. *Let $\mathcal{L}, \mathcal{R} \subset \mathbb{Z}^d \times \mathbb{Z}^d$ be independent random relations. Then*

$$\dim_S(\mathcal{L}\mathcal{R}) = \min\Big\{\dim_S(\mathcal{L}) + \dim_S(\mathcal{R}), d\Big\},$$

when $\dim_S(\mathcal{L})$ and $\dim_S(\mathcal{R})$ exist.

Notation. We write $\phi \preccurlyeq \psi$ (or equivalently, $\psi \succcurlyeq \phi$), if $\phi \le C\psi$ for some constant $C > 0$, which may depend on the laws of the relations considered. We write $\phi \asymp \psi$ if $\phi \preccurlyeq \psi$ and $\phi \succcurlyeq \psi$. For $v \in \mathbb{Z}^d$ and $0 \le n < N$, define the *dyadic shells*

$$H_n^N(v) := \{x \in \mathbb{Z}^d : 2^n \le \langle vx \rangle < 2^N\}.$$

Remark 2.5. As the proof will show, the composition rule of Theorem 2.4 for random relations in $\mathbb{Z}^d$ is valid for any graph where the shells $H_k^{k+1}(v)$ satisfy $|H_k^{k+1}(v)| \asymp 2^{dk}$.

For sets $V, W \subset \mathbb{Z}^d$ let

$$\rho(V, W) := \min\{\langle vw \rangle : v \in V,\ w \in W\}.$$

In particular, if V and W have nonempty intersection, then $\rho(V, W) = 1$. We write $\langle Wx \rangle$ as an abbreviation for $\langle W \cup \{x\} \rangle$.

LEMMA 2.6. *For every $M > 0$, every $x \in \mathbb{Z}^d$ and every $W \subset \mathbb{Z}^d$ with $|W| \le M$,*

$$\langle Wx \rangle \le \langle W \rangle \, \rho(x, W) \preccurlyeq \langle Wx \rangle ,$$

where the constant implicit in the $\preccurlyeq$ notation depends only on M.

Proof. We may assume without loss of generality that $x \notin W$. The inequality $\langle Wx \rangle \le \langle W \rangle \, \rho(x, W)$ holds because given a tree on W we may obtain a tree on $W \cup \{x\}$ by adding an edge connecting x to the closest vertex in W. For the second inequality, consider some tree τ with vertices $W \cup \{x\}$. Let W' denote the neighbors of x in τ, and let $u \in W'$ be such that $\langle xu \rangle = \rho(x, W')$. Let τ' be the tree on W obtained from τ by replacing each edge $\{w', x\}$ where $w' \in W' \setminus \{u\}$, by the edge $\{w', u\}$. (See Figure 2.1.) It is easy to verify that τ' is a tree. For each $w' \in W'$, we have $\langle uw' \rangle \le \langle ux \rangle + \langle xw' \rangle \le 2\langle xw' \rangle$. Hence $\langle \tau' \rangle \, \langle ux \rangle \le 2^M \langle \tau \rangle$, and the inequality $\langle W \rangle \, \rho(x, W) \le 2^M \langle Wx \rangle$ follows. □

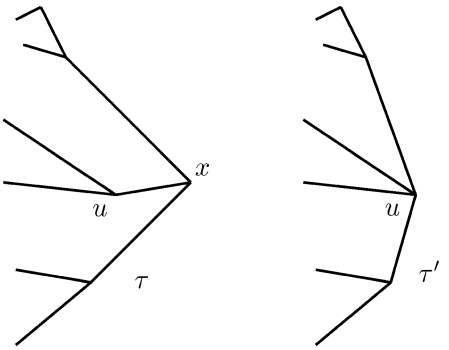


Figure 2.1: The trees τ and τ'.

Remark 2.7. Repeated application of Lemma 2.6 yields that for any set $\{x_1, \ldots, x_n\}$ of n vertices in $\mathbb{Z}^d$,

$$\langle x_1 \ldots x_n \rangle \asymp \prod_{i=1}^{n-1} \rho(x_i, \{x_{i+1}, \ldots, x_n\}), \tag{2.4}$$

where the implied constants depend only on n.

Our next goal in the proof of Theorem 2.4 is to establish (2.1) for the composition $\mathcal{LR}$. For this, the following lemma will be essential.

LEMMA 2.8. *Let $\mathcal{L}$ and $\mathcal{R}$ be independent random relations in $\mathbb{Z}^d$. Suppose that $\dim_S(\mathcal{L}) = d-\alpha$ and $\dim_S(\mathcal{R}) = d-\beta$ exist, and denote $\gamma := \alpha+\beta-d$. For $u, z \in \mathbb{Z}^d$ and $1 \le n \le N$, let*

$$S_{uz} = S_{uz}(n, N) := \sum_{x \in H_n^{N+1}(u)} \mathbf{1}_{u\mathcal{L}x} \mathbf{1}_{x\mathcal{R}z}\,.$$

If $\langle uz \rangle < 2^{n-1}$ and $N \ge n$, then

$$\mathbf{P}[S_{uz} > 0] \succcurlyeq \frac{\sum_{k=n}^{N} 2^{-k\gamma}}{\sum_{k=0}^{N} 2^{-k\gamma}}\,. \tag{2.5}$$

Proof. For $k \ge n$ and $x \in H_k^{k+1}(u)$, we have $\mathbf{P}[u\mathcal{L}x] \asymp 2^{-k\alpha}$ (use (2.1) and (2.3)). Also, for $x \in H_k^{k+1}(u)$, $k \ge n$,

$$\frac{1}{2}\langle xu \rangle \le \langle xu \rangle - \langle zu \rangle \le \langle xz \rangle \le \langle xu \rangle + \langle zu \rangle \le 2\langle xu \rangle$$

and $\mathbf{P}[x\mathcal{R}z] \asymp 2^{-k\beta}$.

Because $\mathcal{L}$ and $\mathcal{R}$ are independent and $|H_k^{k+1}(u)| \asymp 2^{dk}$, we have

$$\mathbf{E}[S_{uz}] \asymp \sum_{k=n}^{N} 2^{kd} 2^{-k\alpha} 2^{-k\beta} = \sum_{k=n}^{N} 2^{-k\gamma}\,. \tag{2.6}$$

To estimate the second moment, observe that if $2\langle uz\rangle \le \langle ux\rangle \le \langle uy\rangle$, then

$$\langle uy\rangle \le \langle uz\rangle + \langle zy\rangle \le \frac{1}{2}\langle uy\rangle + \langle zy\rangle\,,$$

so that $\langle xy\rangle \le \langle xu\rangle + \langle uy\rangle \le 2\langle uy\rangle \le 4\langle yz\rangle$. Applying the first two inequalities here, (2.2) for $\mathcal{L}$ and Lemma 2.6, we obtain that

$$\begin{aligned}\mathbf{P}[u\mathcal{L}x, u\mathcal{L}y] &\preccurlyeq \langle ux\rangle^{-\alpha}\langle uy\rangle^{-\alpha} + \langle uxy\rangle^{-\alpha} \\ &\preccurlyeq \langle ux\rangle^{-\alpha}\big(\langle uy\rangle^{-\alpha} + \langle xy\rangle^{-\alpha}\big) \preccurlyeq \langle ux\rangle^{-\alpha}\langle xy\rangle^{-\alpha}.\end{aligned}$$

Similarly, from Lemma 2.6 and the inequality $\langle xy\rangle \le 4\langle yz\rangle$ above, it follows that

$$\mathbf{P}[x\mathcal{R}z, y\mathcal{R}z] \preccurlyeq \langle xz\rangle^{-\beta}\big(\langle xy\rangle^{-\beta} + \langle yz\rangle^{-\beta}\big) \preccurlyeq \langle xz\rangle^{-\beta}\langle xy\rangle^{-\beta}\,.$$

Since

$$\mathbf{E}[S_{uz}^2] \le 2 \sum_{x\in H_n^{N+1}(u)} \sum_{y\in H_n^{N+1}(u)} \mathbf{1}_{\{\langle uy\rangle \ge \langle ux\rangle\}} \mathbf{P}[u\mathcal{L}x, u\mathcal{L}y]\mathbf{P}[x\mathcal{R}z, y\mathcal{R}z],$$

we deduce by breaking the inner sum up into sums over $y \in H_j^{j+1}(x)$ for various j that

(2.7)

$$\mathbf{E}[S_{uz}^2] \preccurlyeq \sum_{x\in H_n^{N+1}(u)} \langle ux\rangle^{-\alpha}\langle xz\rangle^{-\beta} \sum_{y\in H_0^{N+2}(x)} \langle xy\rangle^{-\alpha-\beta} \preccurlyeq \mathbf{E}[S_{uz}] \sum_{j=0}^{N} 2^{j(d-\alpha-\beta)}\,.$$

Finally, recall that $\gamma = \alpha + \beta - d$, and apply the Cauchy-Schwarz inequality in the form

$$\mathbf{P}[S_{uz} > 0] \ge \frac{\mathbf{E}[S_{uz}]^2}{\mathbf{E}[S_{uz}^2]}\,.$$

The estimates (2.6) and (2.7) then yield the assertion of the lemma. □

Corollary 2.9. *Under the assumptions of Theorem* 2.4, $\mathbf{P}[u\mathcal{L}\mathcal{R}z] \succcurlyeq \langle uz\rangle^{-\gamma'}$ *for all* $u, z \in \mathbb{Z}^d$, *where* $\gamma' := \max\{0, d - \dim_S(\mathcal{L}) - \dim_S(\mathcal{R})\}$.

Proof. Let $n := \lceil \log_2\langle uz\rangle\rceil + 2$ and $\gamma := \alpha + \beta - d$ with $\alpha := d - \dim_S(\mathcal{L})$, $\beta := d - \dim_S(\mathcal{R})$. Apply the lemma with $N := n$ if $\gamma \ne 0$ and $N := 2n$ if $\gamma = 0$. □

The proof of (2.2) for the composition $\mathcal{L}\mathcal{R}$ requires some further preparation.

Lemma 2.10. *Let* $\alpha, \beta \in [0, d)$ *satisfy* $\alpha + \beta > d$. *Let* $\gamma := \alpha + \beta - d$. *Then for* $v, w \in \mathbb{Z}^d$,

$$\sum_{x\in\mathbb{Z}^d} \langle vx\rangle^{-\alpha}\langle xw\rangle^{-\beta} \asymp \langle vw\rangle^{-\gamma}\,.$$

Proof. Suppose that $2^N \le \langle vw\rangle \le 2^{N+1}$. By symmetry, it suffices to sum over vertices x such that $\langle xv\rangle \le \langle xw\rangle$. For such x in the shell $H_n^{n+1}(v)$, we have by the triangle inequality

$$\langle xv\rangle^{-\alpha}\langle xw\rangle^{-\beta} \asymp 2^{-n\alpha}2^{-\max\{n,N\}\beta}.$$

Multiplying by 2^{dn} (for the volume of the shell) and summing over all n prove the lemma. □

LEMMA 2.11. *Let $M > 0$ be finite and let $V, W \subset \mathbb{Z}^d$ satisfy $|V|, |W| \le M$. Let $\alpha, \beta \in [0, d)$ satisfy $\alpha + \beta > d$. Denote $\gamma := \alpha + \beta - d$. Then*

$$\sum_{x\in\mathbb{Z}^d} \langle Vx\rangle^{-\alpha}\langle Wx\rangle^{-\beta} \preccurlyeq \langle V\rangle^{-\alpha}\rho(V,W)^{-\gamma}\langle W\rangle^{-\beta} \le \langle VW\rangle^{-\gamma}, \tag{2.8}$$

where the constant implicit in the $\preccurlyeq$ relation may depend only on M.

Proof. Using Lemma 2.6 we see that

$$\langle Vx\rangle^{-\alpha} \preccurlyeq \langle V\rangle^{-\alpha}\rho(x,V)^{-\alpha} \le \langle V\rangle^{-\alpha}\sum_{v\in V}\langle vx\rangle^{-\alpha}$$

and similarly $\langle Wx\rangle^{-\beta} \preccurlyeq \langle W\rangle^{-\beta}\sum_{w\in W}\langle wx\rangle^{-\beta}$. Therefore, by Lemma 2.10,

$$\begin{aligned}\sum_{x\in\mathbb{Z}^d}\langle Vx\rangle^{-\alpha}\langle Wx\rangle^{-\beta} &\preccurlyeq \langle V\rangle^{-\alpha}\langle W\rangle^{-\beta}\sum_{v\in V}\sum_{w\in W}\langle vw\rangle^{-\gamma}\\ &\le M^2\langle V\rangle^{-\alpha}\langle W\rangle^{-\beta}\rho(V,W)^{-\gamma}.\end{aligned}$$

The rightmost inequality in (2.8) holds because $\gamma \le \min\{\alpha, \beta\}$ and given trees on V and on W, a tree on $V \cup W$ can be obtained by adding an edge $\{v, w\}$ with $v \in V, w \in W$ and $\langle vw\rangle = \rho(V, W)$ (unless $V \cap W \ne \varnothing$, in which case $|V \cap W| - 1$ edges have to be deleted to obtain a tree on $V \cup W$). □

LEMMA 2.12. *Let $\alpha, \beta \in [0, d)$ satisfy $\gamma := \alpha + \beta - d > 0$. Then*

$$\sum_{a\in\mathbb{Z}^d}\langle ax\rangle^{-\alpha}\langle ay\rangle^{-\beta}\langle az\rangle^{-\gamma} \preccurlyeq \langle xyz\rangle^{-\gamma}$$

holds for $x, y, z \in \mathbb{Z}^d$.

Proof. Without loss of generality, we may assume that $\langle zx\rangle \le \langle zy\rangle$. Consider separately the sum over $A := \{a : \langle az\rangle \le \frac{1}{2}\langle zx\rangle\}$ and its complement. For $a \in A$ we have $\langle ax\rangle \ge \frac{1}{2}\langle zx\rangle$ and $\langle ay\rangle \ge \frac{1}{2}\langle zy\rangle$. Therefore,

$$\begin{aligned}\sum_{a\in A}\langle ax\rangle^{-\alpha}\langle ay\rangle^{-\beta}\langle az\rangle^{-\gamma} &\preccurlyeq \langle zx\rangle^{-\alpha}\langle zy\rangle^{-\beta}\sum_{a\in A}\langle az\rangle^{-\gamma}\\ &\preccurlyeq \langle zx\rangle^{-\alpha}\langle zy\rangle^{-\beta}\langle zx\rangle^{d-\gamma}\\ &= \langle zx\rangle^{-\gamma}\langle zy\rangle^{-\gamma}\frac{\langle zx\rangle^{d-\alpha}}{\langle zy\rangle^{d-\alpha}} \le \langle xyz\rangle^{-\gamma},\end{aligned}$$

because of the assumption $\langle zx\rangle \le \langle zy\rangle$ and $\gamma > 0$. Passing to the complement of A, we have,

$$\sum_{a\in\mathbb{Z}^d\setminus A} \langle az\rangle^{-\gamma}\langle ax\rangle^{-\alpha}\langle ay\rangle^{-\beta} \preccurlyeq \langle zx\rangle^{-\gamma}\textstyle\sum_{a\in\mathbb{Z}^d}\langle ax\rangle^{-\alpha}\langle ay\rangle^{-\beta}$$
$$\preccurlyeq \langle zx\rangle^{-\gamma}\langle xy\rangle^{-\gamma} \le \langle xyz\rangle^{-\gamma},$$

where Lemma 2.10 was used in the next to last inequality. Combining these two estimates completes the proof of the lemma. □

The following slight extension of Lemma 2.12 will also be needed. Under the same assumptions on α, β, γ,

$$\sum_{a\in\mathbb{Z}^d} \langle axw\rangle^{-\alpha}\langle ay\rangle^{-\beta}\langle az\rangle^{-\gamma} \preccurlyeq \langle xwyz\rangle^{-\gamma}. \tag{2.9}$$

This is obtained from Lemma 2.12 by using $\langle axw\rangle \succcurlyeq \langle xw\rangle \min\{\langle ax\rangle, \langle aw\rangle\}$, which is an application of Lemma 2.6, and $\langle xw\rangle\langle wyz\rangle \ge \langle xwyz\rangle$, which holds by the definition of the spread.

Proof of Theorem 2.4. If $\dim_S(\mathcal{L}) + \dim_S(\mathcal{R}) \ge d$, then Corollary 2.9 shows that

$$\inf_{x,y\in\mathbb{Z}^d} \mathbf{P}[x\mathcal{L}\mathcal{R}y] > 0,$$

which is equivalent to $\dim_S(\mathcal{L}\mathcal{R}) = d$. Therefore, assume that $\dim_S(\mathcal{L}) + \dim_S(\mathcal{R}) < d$. Let $\alpha := d - \dim_S(\mathcal{L})$, $\beta := d - \dim_S(\mathcal{R})$ and $\gamma := \alpha + \beta - d$. Since Corollary 2.9 verifies (2.1) for the composition $\mathcal{L}\mathcal{R}$, it suffices to prove (2.2) for $\mathcal{L}\mathcal{R}$ with γ in place of α. Independence of the relations $\mathcal{L}$ and $\mathcal{R}$, together with (2.2) for $\mathcal{L}$ and for $\mathcal{R}$ with β in place of α imply

$$\mathbf{P}[x\mathcal{L}\mathcal{R}z,\, y\mathcal{L}\mathcal{R}w] \le \sum_{a,b\in\mathbb{Z}^d} \mathbf{P}[x\mathcal{L}a,\, y\mathcal{L}b]\mathbf{P}[a\mathcal{R}z,\, b\mathcal{R}w] \tag{2.10}$$
$$\preccurlyeq \sum_{a,b\in\mathbb{Z}^d} \Big(\langle xa\rangle^{-\alpha}\langle yb\rangle^{-\alpha} + \langle xayb\rangle^{-\alpha}\Big)\Big(\langle az\rangle^{-\beta}\langle bw\rangle^{-\beta} + \langle azbw\rangle^{-\beta}\Big).$$

Opening the parentheses gives four sums, which we deal with separately. First,

$$\sum_{a,b\in\mathbb{Z}^d} \langle xa\rangle^{-\alpha}\langle yb\rangle^{-\alpha}\langle az\rangle^{-\beta}\langle bw\rangle^{-\beta} \preccurlyeq \langle xz\rangle^{-\gamma}\langle yw\rangle^{-\gamma}, \tag{2.11}$$

by Lemma 2.10 applied twice. Second, by two applications of Lemma 2.11,

$$\sum_{a\in\mathbb{Z}^d}\sum_{b\in\mathbb{Z}^d} \langle xayb\rangle^{-\alpha}\langle azbw\rangle^{-\beta} \tag{2.12}$$
$$\preccurlyeq \sum_{a\in\mathbb{Z}^d} \rho\big(\{x,a,y\},\{z,a,w\}\big)^{-\gamma}\langle xay\rangle^{-\alpha}\langle zaw\rangle^{-\beta}$$
$$= \sum_{a\in\mathbb{Z}^d} \langle xay\rangle^{-\alpha}\langle zaw\rangle^{-\beta} \preccurlyeq \langle xyzw\rangle^{-\gamma}.$$

Third, by Lemma 2.11 and (2.9),

(2.13)

$$\begin{aligned}\sum_{a\in\mathbb{Z}^d}\sum_{b\in\mathbb{Z}^d}&\langle xayb\rangle^{-\alpha}\langle az\rangle^{-\beta}\langle bw\rangle^{-\beta}\\ &\preccurlyeq \sum_{a\in\mathbb{Z}^d}\rho\big(w,\{x,y,a\}\big)^{-\gamma}\langle az\rangle^{-\beta}\langle xay\rangle^{-\alpha}\\ &\le \sum_{a\in\mathbb{Z}^d}\langle wa\rangle^{-\gamma}\langle az\rangle^{-\beta}\langle xay\rangle^{-\alpha}+\rho\big(w,\{x,y\}\big)^{-\gamma}\sum_{a\in\mathbb{Z}^d}\langle az\rangle^{-\beta}\langle xay\rangle^{-\alpha}\\ &\preccurlyeq \langle wzxy\rangle^{-\gamma}+\rho\big(w,\{x,y\}\big)^{-\gamma}\langle zxy\rangle^{-\gamma}\le 2\langle wzxy\rangle^{-\gamma}.\end{aligned}$$

By symmetry, we also have

$$\sum_{a,b\in\mathbb{Z}^d}\langle xa\rangle^{-\alpha}\langle yb\rangle^{-\alpha}\langle azbw\rangle^{-\beta}\preccurlyeq\langle xywz\rangle^{-\gamma}.$$

This, together with (2.10), (2.11), (2.12) and (2.13), implies that $\mathcal{LR}$ satisfies the correlation inequality (2.2) with γ in place of α, and completes the proof. □

3. Tail triviality

Consider the USF on $\mathbb{Z}^d$. Given $n=1,2,\dots$, let $\mathcal{R}_n$ be the relation consisting of all pairs $(x,y)\in\mathbb{Z}^d\times\mathbb{Z}^d$ such that y may be reached from x by a path which uses no more than $n-1$ edges outside of the USF. We show below that $\mathcal{R}_1$ has stochastic dimension 4 (Theorem 4.2) and that $\mathcal{R}_n$ stochastically dominates the composition of n independent copies of $\mathcal{R}_1$ (Theorem 4.1). By Theorem 2.4, $\mathcal{R}_n$ dominates a relation with stochastic dimension $\min\{4n,d\}$. When $4n\ge d$, this says that $\inf_{x,y\in\mathbb{Z}^d}\mathbf{P}[x\mathcal{R}_n y]>0$. For Theorem 1.1, the stronger statement that $\inf_{x,y\in\mathbb{Z}^d}\mathbf{P}[x\mathcal{R}_n y]=1$ is required. For this purpose, tail triviality needs to be discussed.

Definition 3.1 (Tail triviality). Let $\mathcal{R}\subset\mathbb{Z}^d\times\mathbb{Z}^d$ be a random relation with law $\mathbf{P}$. For a set $\Lambda\subset\mathbb{Z}^d\times\mathbb{Z}^d$, let $\mathfrak{F}_\Lambda$ be the σ-field generated by the events $x\mathcal{R}y$, $(x,y)\in\Lambda$. Let the *left tail field* $\mathfrak{F}_L(v)$ corresponding to a vertex v be the intersection of all $\mathfrak{F}_{\{v\}\times K}$ where $K\subset\mathbb{Z}^d$ ranges over all subsets such that $\mathbb{Z}^d\setminus K$ is finite. Let the *right tail field* $\mathfrak{F}_R(v)$ be the intersection of all $\mathfrak{F}_{K\times\{v\}}$ where $K\subset\mathbb{Z}^d$ ranges over all subsets such that $\mathbb{Z}^d\setminus K$ is finite. Let the *remote tail field* $\mathfrak{F}_W$ be the intersection of all $\mathfrak{F}_{K_1\times K_2}$, where $K_1,K_2\subset\mathbb{Z}^d$ range over all subsets of $\mathbb{Z}^d$ such that $\mathbb{Z}^d\setminus K_1$ and $\mathbb{Z}^d\setminus K_2$ are finite. The random relation $\mathcal{R}$ with law $\mathbf{P}$ is said to be *left tail trivial* if $\mathbf{P}[A]\in\{0,1\}$ for every $A\in\mathfrak{F}_L(v)$ and every $v\in\mathbb{Z}^d$. Analogously, define *right tail triviality* and *remote tail triviality.*

We will need the following known lemma, which is a corollary of the main result of von Weizsäcker (1983). Its proof is included for the reader's convenience.

LEMMA 3.2. *Let* $\{\mathfrak{F}_n\}$ *and* $\{\mathfrak{G}_n\}$ *be two decreasing sequences of complete* σ*-fields in a probability space* $(X, \mathfrak{F}, \mu)$*, with* $\mathfrak{G}_1$ *independent of* $\mathfrak{F}_1$*, and let* $\mathfrak{T}$ *denote the trivial* σ*-field, consisting of events with probability* 0 *or* 1*. If* $\cap_{n\geq 1}\mathfrak{F}_n = \mathfrak{T} = \cap_{n\geq 1}\mathfrak{G}_n$*, then* $\cap_{n\geq 1}(\mathfrak{F}_n \vee \mathfrak{G}_n) = \mathfrak{T}$ *as well.*

Proof. Let $h \in \bigcap_{n=1}^{\infty} L^2(\mathfrak{F}_n \vee \mathfrak{G}_n)$. It suffices to show that h is constant. Suppose that $f_n \in L^2(\mathfrak{F}_n)$, $g_n \in L^\infty(\mathfrak{G}_n)$ and $f \in L^\infty(\mathfrak{F}_1)$. Then

$$\mathbf{E}[f_n g_n f|\mathfrak{G}_1] = g_n \mathbf{E}[f_n f|\mathfrak{G}_1] = g_n \mathbf{E}[f_n f] \qquad \text{a.s.}$$

As the linear span of such products $f_n g_n$ is dense in $L^2(\mathfrak{F}_n \vee \mathfrak{G}_n)$, it follows that

$$\mathbf{E}[hf|\mathfrak{G}_1] \in L^2(\mathfrak{G}_n)\,. \tag{3.1}$$

By our assumption $\mathfrak{T} = \cap_{n\geq 1}\mathfrak{G}_n$, we infer from (3.1) that

$$\mathbf{E}[hf|\mathfrak{G}_1] = \mathbf{E}[hf] \text{ for } f \in L^\infty(\mathfrak{F}_1). \tag{3.2}$$

In particular, $\mathbf{E}[h|\mathfrak{G}_1] = \mathbf{E}[h]$. By symmetry, $\mathbf{E}[h|\mathfrak{F}_1] = \mathbf{E}[h]$, whence

$$\forall f \in L^\infty(\mathfrak{F}_1), \qquad \mathbf{E}[hf] = \mathbf{E}\big[f\mathbf{E}[h|\mathfrak{F}_1]\big] = \mathbf{E}[h]\mathbf{E}[f]\,.$$

Inserting this into (3.2), we conclude that for $f \in L^\infty(\mathfrak{F}_1)$ and $g \in L^\infty(\mathfrak{G}_1)$ (so that f and g are independent),

$$\mathbf{E}[hfg] = \mathbf{E}\big[g\mathbf{E}[hf|\mathfrak{G}_1]\big] = \mathbf{E}[h]\mathbf{E}[f]\mathbf{E}[g] = \mathbf{E}[h]\mathbf{E}[fg]\,.$$

Thus $h - \mathbf{E}[h]$ is orthogonal to all such products fg, so it must vanish a.s. □

Suppose that $\mathcal{R}, \mathcal{L} \subset \mathbb{Z}^d \times \mathbb{Z}^d$ are independent random relations which have trivial remote tails and trivial left tails, and have stochastic dimensions. We do not know whether it follows that $\mathcal{L}\mathcal{R}$ is left tail trivial. For that reason, we introduce the notion of the *restricted composition* $\mathcal{L} \diamond \mathcal{R}$, which is the relation consisting of all pairs $(x, z) \in \mathbb{Z}^d \times \mathbb{Z}^d$ such that there is some $y \in \mathbb{Z}^d$ with $x\mathcal{L}y$ and $y\mathcal{R}z$ and

$$\langle xz\rangle \leq \min\{\langle xy\rangle, \langle yz\rangle\}\,. \tag{3.3}$$

THEOREM 3.3. *Let* $\mathcal{R}, \mathcal{L} \subset \mathbb{Z}^d \times \mathbb{Z}^d$ *be independent random relations.*

(i) *If* $\mathcal{L}$ *has trivial left tail and* $\mathcal{R}$ *has trivial remote tail, then the restricted composition* $\mathcal{L} \diamond \mathcal{R}$ *has trivial left tail.*

(ii) *If* $\dim_S(\mathcal{L})$ *and* $\dim_S(\mathcal{R})$ *exist, then*

$$\dim_S(\mathcal{L} \diamond \mathcal{R}) = \min\{\dim_S(\mathcal{L}) + \dim_S(\mathcal{R}),\, d\}\,.$$

(iii) *If* $\dim_S(\mathcal{L})$ *and* $\dim_S(\mathcal{R})$ *exist,* $\mathcal{L}$ *has trivial left tail,* $\mathcal{R}$ *has trivial right tail and* $\dim_S(\mathcal{L}) + \dim_R(\mathcal{R}) \geq d$, *then* $\mathbf{P}[x\mathcal{L} \diamond \mathcal{R}z] = 1$ *for all* $x, z \in \mathbb{Z}^d$.

Proof. (i) This is a consequence of Lemma 3.2.

(ii) Denote $\gamma' := \max\{0, d - \dim_S(\mathcal{L}) - \dim_R(\mathcal{R})\}$. The inequality $\mathbf{P}[x\mathcal{L} \diamond \mathcal{R}z] \succcurlyeq \langle xz \rangle^{-\gamma'}$ follows from the proof of Corollary 2.9; this concludes the proof if $\gamma' = 0$. If $\gamma' > 0$, then the required upper bound for $\mathbf{P}[x\,\mathcal{L} \diamond \mathcal{R}\,z,\ y\,\mathcal{L} \diamond \mathcal{R}\,w]$ follows from Theorem 2.4, since $\mathcal{L} \diamond \mathcal{R}$ is a subrelation of $\mathcal{LR}$.

(iii) Let $\mathfrak{F}_n$ (respectively, $\mathfrak{G}_n$) be the (completed) σ-field generated by the events $u\mathcal{L}x$ (respectively, $x\mathcal{R}z$) as x ranges over $H_n^\infty(u)$. The event

$$A_n := \{\exists x \in H_n^{2n}(u) \,:\, u\mathcal{L}x,\ x\mathcal{R}z\}$$

is clearly in $\mathfrak{F}_n \vee \mathfrak{G}_n$. By Lemma 2.8, there is a constant C such that $\mathbf{P}[A_n] \geq 1/C > 0$, provided that n is sufficiently large. Let $A = \cap_{k\geq 1} \cup_{n\geq k} A_n$ be the event that there are infinitely many $x \in \mathbb{Z}^d$ satisfying $u\mathcal{L}x$ and $x\mathcal{R}z$. Then $\mathbf{P}[A] \geq 1/C$. By Lemma 3.2, $\mathbf{P}[A] = 1$. □

COROLLARY 3.4. *Let* $m \geq 2$, *and let* $\{\mathcal{R}_i\}_{i=1}^m$ *be independent random relations in* $\mathbb{Z}^d$ *such that* $\dim_S(\mathcal{R}_i)$ *exists for each* $i \leq m$. *Suppose that*

$$\sum_{i=1}^m \dim_S(\mathcal{R}_i) \geq d\,,$$

and in addition, $\mathcal{R}_1$ *is left tail trivial, each of* $\mathcal{R}_2, \ldots, \mathcal{R}_{m-1}$ *has a trivial remote tail, and* $\mathcal{R}_m$ *is right tail trivial. Then* $\mathbf{P}[u\mathcal{R}_1\mathcal{R}_2 \cdots \mathcal{R}_m z] = 1$ *for all* $u, z \in \mathbb{Z}^d$.

Proof. Let $\mathcal{L}_1 = \mathcal{R}_1$ and define inductively $\mathcal{L}_k = \mathcal{L}_{k-1} \diamond \mathcal{R}_k$ for $k = 2, \ldots, m$, so that $\mathcal{L}_k$ is a subrelation of $\mathcal{R}_1\mathcal{R}_2 \ldots \mathcal{R}_k$. It follows by induction from Theorem 3.3 that $\mathcal{L}_k$ has a trivial left tail for each $k < m$, and

$$\dim_S(\mathcal{L}_k) = \min\Big\{\sum_{i=1}^k \dim_S(\mathcal{R}_i), d\Big\}\,.$$

By Theorem 3.3(iii), the restricted composition $\mathcal{L}_m = \mathcal{L}_{m-1} \diamond \mathcal{R}_m$ satisfies

$$\mathbf{P}[u\mathcal{L}_m z] = 1 \text{ for all } u, z \in \mathbb{Z}^d\,. \tag{3.4}$$

□

4. Relevant USF properties

Basic to the understanding of the USF is a procedure from BLPS [2] that generates the (wired) USF on any transient graph; it is called "Wilson's method rooted at infinity", since it is based on an algorithm from Wilson [15] for picking uniformly a spanning tree in a finite graph. Let $\{v_1, v_2, \ldots\}$ be an

arbitrary ordering of the vertices of a transient graph G. Let X_1 be a simple random walk started from v_1. Let F_1 denote the loop-erasure of X_1, which is obtained by following X_1 and erasing the loops as they are created. Let X_2 be a simple random walk from v_2 which stops if it hits F_1, and let F_2 be the union of F_1 with the loop-erasure of X_2. Inductively, let X_n be a simple random walk from v_n, which is stopped if it hits F_{n-1}, and let F_n be the union of F_{n-1} with the loop-erasure of X_n. Then $F := \bigcup_{n=1}^{\infty} F_n$ has the distribution of the (wired) USF on G. The edges in F inherit the orientation from the loop-erased walks creating them, and hence F may be thought of as an oriented forest. Its distribution does not depend on the ordering chosen for the vertices. See BLPS [2] for details.

We say that a random set A *stochastically dominates* a random set Q if there is a coupling μ of A and Q such that $\mu[A \supset Q] = 1$.

THEOREM 4.1 (Domination). *Let $F, F_0, F_1, F_2, \ldots, F_m$ be independent samples of the wired* USF *in the graph* G. *Let $\mathcal{C}(x, F)$ denote the vertex set of the component of x in F. Fix a distinguished vertex v_0 in* G, *and write $\mathcal{C}_0 = \mathcal{C}(v_0, F)$. For $j \geq 1$, define inductively $\mathcal{C}_j$ to be the union of all vertex components of F that are contained in, or adjacent to, $\mathcal{C}_{j-1}$. Let $Q_0 = \mathcal{C}(v_0, F_0)$. For $j \geq 1$, define inductively Q_j to be the union of all vertex components of F_j that intersect Q_{j-1}. Then $\mathcal{C}_m$ stochastically dominates Q_m.*

Proof. For each $R > 0$ let $B_R := \{v \in \mathbb{Z}^d : |v| < R\}$, where $|v|$ is the graph distance from v_0 to v. Fix $R > 0$. Let B_R^W be the graph obtained from G by collapsing the complement of B_R to a single vertex, denoted v_R^*. Let $F', F_0', \ldots, F_m'$ be independent samples of the uniform spanning tree (UST) in B_R^W. Define F^* to be F' without the edges incident to v_R^*, and define F_i^* for $i = 0, \ldots, m$ analogously.

Write $\mathcal{C}_0^* = \mathcal{C}(v_0, F^*)$. For $j \geq 1$, define inductively $\mathcal{C}_j^*$ to be the union of all vertex components of F^* that are contained in, or adjacent to, $\mathcal{C}_{j-1}^*$.

Let $Q_0^* = \mathcal{C}(v_0, F_0^*)$. For $j \geq 1$, define inductively Q_j^* to be the union of all vertex components of F_j^* that intersect Q_{j-1}^*.

We show by induction on j that $\mathcal{C}_j^*$ stochastically dominates Q_j^*. The induction base where $j = 0$ is obvious. For the inductive step, assume that $0 \leq j < m$ and there is a coupling μ_j of F' and $(F_0', F_1', \ldots, F_j')$ such that $\mathcal{C}_j^* \supset Q_j^*$ holds μ_j-a.s. Let H be a set of vertices in B_R such that $\mathcal{C}_j^* = H$ with positive probability. Let $\widehat{H^c}$ denote the subgraph of B_R^W spanned by the vertices $B_R \cup \{v_R^*\} \setminus H$. Let S_H be $F' \cap \widehat{H^c}$ conditioned on $\mathcal{C}_j^* = H$. Since $\mathcal{C}_j^*$ is a union of components of F^*, it is clear that S_H has the same distribution as a UST on $\widehat{H^c}$. Therefore, by the negative association property of uniform spanning trees (see the discussion following Remark 5.7 in BLPS [2]), S_H stochastically dominates $F_{j+1}' \cap \widehat{H^c}$ (conditional on $\mathcal{C}_j^* = H$). Hence,

we may extend μ_j to a coupling μ_{j+1} of F' and $(F'_0, F'_1, \ldots, F'_{j+1})$ such that $\mathcal{C}^*_j \supset Q^*_j$ and $H = \mathcal{C}^*_j$ satisfies $F^* \cap \widehat{H^c} \supset F^*_{j+1} \cap \widehat{H^c}$ a.s. with respect to μ_{j+1}. In such a coupling, we also have $\mathcal{C}^*_{j+1} \supset Q^*_{j+1}$ a.s.

This completes the induction step, and proves that $\mathcal{C}^*_m$ stochastically dominates Q^*_m. Observe that $\mathcal{C}^*_m$ converges weakly to $\mathcal{C}_m$ and Q^*_m converges weakly to Q_m, as $R \to \infty$. This follows from the fact that F stochastically dominates F^* (see BLPS [2, Cor. 4.3(b)]) and $F^* \to F$ weakly as $R \to \infty$. The theorem follows. □

THEOREM 4.2 (Stochastic dimension of USF). *Let F be the* USF *in $\mathbb{Z}^d$, where $d \geq 5$. Let $\mathcal{U}$ be the relation*

$$\mathcal{U} := \big\{(x, y) \in \mathbb{Z}^d \times \mathbb{Z}^d : \text{ } x \text{ and } y \text{ are in the same component of } F\big\}.$$

Then $\mathcal{U}$ has stochastic dimension 4.

Proof. We use Wilson's method rooted at infinity and a technique from LPS [9]. For $x, z \in \mathbb{Z}^d$, consider independent simple random walk paths $\{X(i)\}_{i\geq 0}$ and $\{Z(j)\}_{j\geq 0}$ starting at x and z respectively. By running Wilson's method rooted at infinity starting with x and then starting another random walk from z, we see that $\mathbf{P}[z\mathcal{U}x]$ is equal to the probability that the walk Z intersects the loop-erasure of the walk X. Given $m \in \mathbb{N}$, denote by $\{L_m(i)\}_{i=0}^{q_m}$ the path obtained from loop-erasing $\{X(i)\}_{i=0}^{m}$. Define

$$\tau_L(m) := \inf\Big\{k \in \{0, 1, \ldots, q_m\} : \text{ } L_m(k) \in \{X(i)\}_{i\geq m}\Big\},$$
$$\tau_L(m, n) := \inf\Big\{k \in \{0, 1, \ldots, q_m\} : \text{ } L_m(k) \in \{Z(j)\}_{j\geq n}\Big\}.$$

Consider the indicator variables $I_{m,n} := \mathbf{1}_{\{X(m)=Z(n)\}}$ and

$$J_{m,n} := I_{m,n}\mathbf{1}_{\{\tau_L(m,n)\leq\tau_L(m)\}}.$$

Observe that on the event $J_{m,n} = 1$, the path $\{Z(j)\}_{j\geq 0}$ intersects the loop-erasure of $\{X(i)\}_{i=0}^{\infty}$ at $L_m\big(\tau_L(m, n)\big)$. Hence $\mathbf{P}[x\mathcal{U}z] = \mathbf{P}[\sum_{m,n} J_{m,n} > 0]$. Given $X(m) = Z(n)$, the law of $\big\{X(m + j)\big\}_{j\geq 0}$ is the same as the law of $\big\{Z(n + j)\big\}_{j\geq 0}$. Therefore,

$$\mathbf{P}[\tau_L(m, n) \leq \tau_L(m) | X(m) = Z(n)] \geq 1/2,$$

so that $\mathbf{E}[I_{m,n}] \geq \mathbf{E}[J_{m,n}] \geq \mathbf{E}[I_{m,n}]/2$. Let

$$\Phi = \Phi_{xz} := \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} I_{m,n};$$
$$\Psi = \Psi_{xz} := \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} J_{m,n}.$$

Then

$$\begin{aligned}\mathbf{E}[\Psi] &\asymp \sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\mathbf{E}[I_{m,n}]\\ &= \sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\sum_{v\in\mathbb{Z}^d}\mathbf{P}[X(m)=v]\,\mathbf{P}[Z(n)=v]\\ &= \sum_{v\in\mathbb{Z}^d}\mathrm{G}(x,v)\mathrm{G}(z,v)\,,\end{aligned}$$

where G is the Green function for a simple random walk in $\mathbb{Z}^d$. Since $\mathrm{G}(x,v)\asymp\langle xv\rangle^{\,2-d}$ (see, e.g., Spitzer [12]), we infer that (see Lemma 2.10)

$$\mathbf{E}[\Psi]\asymp\sum_{v\in\mathbb{Z}^d}\langle xv\rangle^{\,2-d}\langle zv\rangle^{\,2-d}\asymp\langle xz\rangle^{\,4-d}\,. \tag{4.1}$$

The second moment calculation for Φ is classical. Again, the relation $\mathrm{G}(x,v)\asymp\langle xv\rangle^{\,2-d}$ is used:

$$\begin{aligned}\mathbf{E}[\Psi^2]&\le\mathbf{E}[\Phi^2]\\ &\le\sum_{y,w\in\mathbb{Z}^d}\mathrm{G}(x,y)\mathrm{G}(y,w)[\mathrm{G}(z,y)\mathrm{G}(y,w)+\mathrm{G}(z,w)\mathrm{G}(w,y)]\\ &\preccurlyeq\sum_{y\in\mathbb{Z}^d}\mathrm{G}(x,y)\mathrm{G}(z,y)+\sum_{y\in\mathbb{Z}^d}\mathrm{G}(x,y)\sum_{w\in\mathbb{Z}^d}\mathrm{G}(y,w)^2\mathrm{G}(z,w)\\ &\preccurlyeq\sum_{y\in\mathbb{Z}^d}\langle xy\rangle^{\,2-d}\langle zy\rangle^{\,2-d}+\sum_{y\in\mathbb{Z}^d}\langle xy\rangle^{\,2-d}\sum_{w\in\mathbb{Z}^d}\langle yw\rangle^{\,4-2d}\langle zw\rangle^{\,2-d}\,.\end{aligned}$$

The proof of Lemma 2.10 gives

$$\sum_{w\in\mathbb{Z}^d}\langle yw\rangle^{\,4-2d}\langle zw\rangle^{\,2-d}\preccurlyeq\langle yz\rangle^{\,2-d}\,.$$

Hence

$$\mathbf{E}[\Psi^2]\preccurlyeq\langle xz\rangle^{\,4-d}+\sum_{y\in\mathbb{Z}^d}\langle xy\rangle^{\,2-d}\langle yz\rangle^{\,2-d}\preccurlyeq\langle xz\rangle^{\,4-d}\,.$$

Therefore,

$$\mathbf{P}[x\mathcal{U}z]\ge\mathbf{P}[\Psi_{xz}>0]\ge\frac{\mathbf{E}[\Psi_{xz}]^2}{\mathbf{E}[\Psi_{xz}^2]}\succcurlyeq\langle xz\rangle^{\,4-d}\,.$$

This verifies (2.1) for $\mathcal{U}$ with $d-\alpha=4$.

To verify (2.2), we must bound $\mathbf{P}[x\mathcal{U}z,y\mathcal{U}w]$. Let X,Z,Y,W be simple random walks starting from x,z,y,w, respectively. Let $\widehat{X}$ denote the set of vertices visited by X, and similarly for Z,Y and W. Then we may generate the

USF by first using the random walks X, Z, Y, W. On the event $x\mathcal{U}z \wedge y\mathcal{U}w \wedge \neg x\mathcal{U}y$, we must have $\widehat{X} \cap \widehat{Z} \neq \varnothing$ and $\widehat{Y} \cap \widehat{W} \neq \varnothing$. Hence,

$$\begin{aligned}\mathbf{P}[x\mathcal{U}z \wedge y\mathcal{U}w \wedge \neg x\mathcal{U}y] &\le \mathbf{P}[\widehat{X} \cap \widehat{Z} \neq \varnothing]\, \mathbf{P}[\widehat{Y} \cap \widehat{W} \neq \varnothing] \\ &\le \sum_{a\in\mathbb{Z}^d} \mathrm{G}(x,a)\mathrm{G}(z,a) \sum_{b\in\mathbb{Z}^d} \mathrm{G}(y,b)\mathrm{G}(w,b) \\ &\preccurlyeq \langle xz\rangle^{\,4-d}\langle yw\rangle^{\,4-d}.\end{aligned} \tag{4.2}$$

On the other hand, the event $x\mathcal{U}z \wedge y\mathcal{U}w \wedge x\mathcal{U}y$ is the event $U(x,y,z,w)$ that x, y, z, w are all in the same USF tree.

By Theorem 4.3 below, $\mathbf{P}[U(x,y,z,w))] \preccurlyeq \langle xyzw\rangle^{\,4-d}$. This together with (4.2) gives

$$\begin{aligned}\mathbf{P}[x\mathcal{U}z \wedge y\mathcal{U}w] &\le \mathbf{P}[x\mathcal{U}z \wedge y\mathcal{U}w \wedge \neg x\mathcal{U}y] + \mathbf{P}[U(x,y,z,w)] \\ &\preccurlyeq \langle xz\rangle^{\,4-d}\langle yw\rangle^{\,4-d} + \langle xyzw\rangle^{\,4-d},\end{aligned}$$

which verifies (2.2) with $\alpha = d-4$, and completes the proof. □

THEOREM 4.3. *For any finite set $W \subset \mathbb{Z}^d$, denote by $U(W)$ the event that all vertices in W are in the same* USF *component. Then*

$$\mathbf{P}[U(W)] \preccurlyeq \langle W\rangle^{\,4-d}, \tag{4.3}$$

where the implied constant depends only on d and the cardinality of W.

Proof. When $|W| = 2$, say $W = \{x, z\}$, (4.3) follows from (4.1), since $\mathbf{P}[x\mathcal{U}z] = \mathbf{P}[\Psi > 0] \le \mathbf{E}[\Psi]$. We proceed by induction on $|W|$. For the inductive step, suppose that $|W| \ge 3$. For $x, y \in W$ denote by $U(W; x, y)$ the intersection of $U(W)$ with the event that the path connecting x, y in the USF is edge-disjoint from the oriented USF paths connecting the vertices in $V := W \setminus \{x, y\}$ to infinity. By considering all the possibilities for the vertex $z \in \mathbb{Z}^d$ where the oriented USF paths from x and y to ∞ meet, and running Wilson's method rooted at infinity starting with random walks from the vertices in $V \cup \{z\}$ and following by random walks from x and y we obtain

$$\mathbf{P}[U(W; x, y)] \preccurlyeq \sum_{z\in\mathbb{Z}^d} \mathbf{P}\big[U(Vz)\big]\langle zx\rangle^{\,2-d}\langle zy\rangle^{\,2-d}, \tag{4.4}$$

where $Vz := V \cup \{z\}$. By the induction hypothesis and Lemma 2.6,

$$\mathbf{P}[U(Vz)] \preccurlyeq \langle Vz\rangle^{\,4-d} \preccurlyeq \langle V\rangle^{\,4-d} \sum_{v\in V} \langle vz\rangle^{\,4-d}.$$

Inserting this into (4.4) and applying Lemma 2.12 with $\alpha = \beta = d-2$, yields

$$\begin{aligned}\mathbf{P}[U(W; x, y)] &\preccurlyeq \langle V\rangle^{\,4-d} \sum_{v\in V} \sum_{z\in\mathbb{Z}^d} \langle vz\rangle^{\,4-d}\langle zx\rangle^{\,2-d}\langle zy\rangle^{\,2-d} \\ &\preccurlyeq \langle V\rangle^{\,4-d} \sum_{v\in V} \langle vxy\rangle^{\,4-d} \preccurlyeq \langle W\rangle^{\,4-d},\end{aligned}$$

where the last inequality follows from the fact that for every $v \in V$, the union of a tree on $\{v, x, y\}$ and a tree on V is a tree on $W = V \cup \{x, y\}$. Finally, observe that

$$U(W) = \bigcup_{(x,y)} U(W; x, y)\,,$$

where (x, y) runs over all pairs of distinct vertices in W. Indeed, on $U(W)$, denote by $m(W)$ the vertex where all oriented USF paths based in W meet, and pick x, y as the pair of vertices in W such that their oriented USF paths meet farthest (in the intrinsic metric of the tree) from $m(W)$. Consequently,

$$\mathbf{P}[U(W)] \le \sum_{(x,y)} \mathbf{P}[U(W; x, y)] \preccurlyeq \langle W \rangle^{4-d}\,. \qquad \square$$

Remark 4.4. The estimate in Theorem 4.3 is tight, up to constants; i.e., for any finite set $W \subset \mathbb{Z}^d$,

$$\mathbf{P}[U(W)] \succcurlyeq \langle W \rangle^{4-d}\,, \tag{4.5}$$

where the implicit constant may depend only on d and the cardinality of W. Since we will not need this lower bound, we omit the proof.

THEOREM 4.5 (Tail triviality of the USF relation). *The relation $\mathcal{U}$ of Theorem 4.2 has trivial left, right and remote tails.*

In BLPS [2] it was proved that the tail of the (wired or free) USF on every infinite graph is trivial. However, this is not the same as the tail triviality of the relation $\mathcal{U}$. Indeed, if the underlying graph G is a regular tree of degree greater than 2, then the relation $\mathcal{U}$ determined by the (wired) USF in G does not have a trivial left tail.

Proof. The theorem clearly holds when $d \le 4$, for then $\mathcal{U} = \mathbb{Z}^d \times \mathbb{Z}^d$ a.s. Therefore, restrict to the case $d > 4$. Let F be the USF in $\mathbb{Z}^d$. We start with the remote tail. Fix $x \in \mathbb{Z}^d$, $r > 0$, and let $\mathcal{S} = \{K_1 \subset F, K_2 \cap F = \varnothing\}$ be a cylinder event where K_1, K_2 are sets of edges in a ball $B(x, r)$ of radius r and center x. For $R > 0$, let Υ_R be the union of all one-sided-infinite simple paths in F that start at some vertex $v \in \mathbb{Z}^d \setminus B(x, R)$. By BLPS [2, Th. 10.1], almost surely each component of F has one end. (This implies that Υ_R is a.s. the same as the union of all oriented USF paths starting outside of $B(x, R)$.) Therefore, there exists a function $\phi(R)$ with $\phi(R) \to \infty$ as $R \to \infty$ such that

$$\lim_{R\to\infty} \mathbf{P}[\Upsilon_R \cap B(x, \phi(R)) \neq \varnothing] = 0\,. \tag{4.6}$$

Let $\tilde{\Upsilon}_R := \Upsilon_R$ if $\Upsilon_R \cap B(x, \phi(R)) = \varnothing$, and $\tilde{\Upsilon}_R := \varnothing$ otherwise. Note that $\tilde{\Upsilon}_R$ is a.s. determined by $F \setminus B(x, \phi(R))$, or more precisely by the set of edges

in F with both endpoints outside $B(x, \phi(R))$. By tail triviality of F itself, it therefore follows that

$$\lim_{R\to\infty} \mathbf{E}\Big[\big|\mathbf{P}[\mathcal{S}|\tilde{\Upsilon}_R] - \mathbf{P}[\mathcal{S}]\big|\Big] = 0\,. \tag{4.7}$$

For $R > r$, on the event $\Upsilon_R \cap B\big(x, \phi(R)\big) = \varnothing$, we have $\mathbf{P}[\mathcal{S}|\tilde{\Upsilon}_R] = \mathbf{P}[\mathcal{S}|\Upsilon_R]$. Hence, by (4.6),

$$\lim_{R\to\infty} \mathbf{E}\Big[\big|\mathbf{P}[\mathcal{S}|\Upsilon_R] - \mathbf{P}[\mathcal{S}|\tilde{\Upsilon}_R]\big|\Big] = 0\,.$$

With (4.7) this gives

$$\lim_{R\to\infty} \mathbf{E}\Big[\big|\mathbf{P}[\mathcal{S}|\Upsilon_R] - \mathbf{P}[\mathcal{S}]\big|\Big] = 0\,. \tag{4.8}$$

Since the remote tail of $\mathcal{U}$ is determined by Υ_R for every R, the remote tail is independent of $\mathcal{S}$, whence it is trivial.

Now consider the left tail of $\mathcal{U}$. Let A be an event in $\mathfrak{F}_L(x)$. Let $\mathcal{S} = \{K_1 \subset F, K_2 \cap F = \varnothing\}$ as above, where K_1, K_2 are sets of edges in $B(x, r)$. To establish the triviality of $\mathfrak{F}_L(x)$, it is enough to show that A and $\mathcal{S}$ are independent.

Let Υ'_R be the union of all the one-sided-infinite simple paths in F that start at some vertex in the outer boundary of $B(x, R)$. By Wilson's method rooted at infinity we may construct F by first choosing Υ'_R, then the path in F, γ_R say, from x to Υ'_R and after that the paths starting at points outside $\Upsilon'_R \cup \gamma_R$. Let ζ_R be the endpoint of γ_R on Υ'_R. Recall that γ_R is obtained by loop erasure of a simple random walk path starting from x and stopped when it hits Υ'_R. Given Υ'_R, the paths in F to Υ'_R from the vertices in the complement of $\Upsilon'_R \cup B(x, R)$ are conditionally independent of γ_R. Therefore the conditional distribution of ζ_R given Υ_R is just the harmonic measure on Υ'_R for a simple random walk started at x. Given Υ_R, we shall write $\mu_y(z; \Upsilon_R)$ for the probability that a simple random walk started at y first hits Υ_R at z. Similarly, for a set W of vertices we write $\mu_y(W; \Upsilon_R) := \sum_{z\in W} \mu_y(z; \Upsilon_R)$.

It is clear that the pair (Υ_R, ζ_R) determines for which $z \notin B(x, R)$ the relation $x\mathcal{U}z$ holds. Therefore, the indicator function of A is measurable with respect to the pair (Υ_R, ζ_R). Given Υ_R, let A_R denote the set of vertices $z \in \Upsilon_R$ such that A holds if $\zeta_R = z$. Then

$$\mathbf{P}[A] = \mathbf{P}\big[\zeta_R \in A_R\big] = \mathbf{E}\big[\mu_x(A_R; \Upsilon_R)\big]\,. \tag{4.9}$$

A very similar relation holds for $\mathbf{P}[\mathcal{S}, A]$. Instead of choosing γ_R immediately after Υ'_R, we can first determine whether $\mathcal{S}$ occurs after choosing Υ'_R and then choose γ_R. Again when Υ'_R is given, the paths from points outside Υ'_R and outside $B(x, R)$ are not influenced by $\mathcal{S}$, so that $\mathbf{P}[\mathcal{S}|\Upsilon'_R] = \mathbf{P}[\mathcal{S}|\Upsilon_R]$. We further remind the reader of the following fact (see, e.g., BLPS [2]). Suppose that $G = (V, E)$ is a finite graph and $E_1, E_2 \subset E$. Let T' be the (set of edges of the) UST in G. Then T', conditioned on $T' \cap (E_1 \cup E_2) = E_1$ (assuming this

event has positive probability), is the union of E_1 with the set of edges of a UST on the graph obtained from G by contracting the edges in E_1 and deleting the edges in E_2. Let H be the graph obtained from $\mathbb{Z}^d$ by contracting the edges in K_1 and deleting the edges in K_2. By first choosing Υ'_R and continuing with Wilson's algorithm we see that the conditional law of $F \setminus \Upsilon_R$ given Υ_R, is the law of a UST on the finite graph obtained from $\mathbb{Z}^d$ by gluing all vertices in Υ_R to a single vertex. If we have a further condition on the occurrence of $\mathcal{S}$, then we should also contract the edges in K_1 and delete the edges in K_2. Therefore, conditionally on Υ_R and the occurrence of $\mathcal{S}$, the path γ_R has the distribution of the loop-erasure of a simple random walk on H, starting at x and stopping when it hits Υ'_R. If we write $\mu_y^H(z;\Upsilon_R)$ for the probability that a simple random walk on H started at y first hits Υ_R at z, then we obtain, analogously to (4.9), that

$$\mathbf{P}[\mathcal{S}, A] = \mathbf{P}[\mathcal{S}, \zeta_R \in A_R] = \mathbf{E}\Big[\mathbf{P}[\mathcal{S}|\Upsilon_R]\, \mu_x^H(A_R;\Upsilon_R)\Big]\,.$$

In view of (4.8) this gives

$$\Big|\mathbf{P}[\mathcal{S}, A] - \mathbf{P}[\mathcal{S}]\,\mathbf{E}\big[\mu_x^H(A_R;\Upsilon_R)\big]\Big| \le \mathbf{E}\Big[\big|\mathbf{P}[\mathcal{S}|\Upsilon_R] - \mathbf{P}[\mathcal{S}]\big|\Big] \underset{R\to\infty}{\longrightarrow} 0\,. \tag{4.10}$$

Given $\varepsilon > 0$, let $R_1 > r$ be such that a simple random walk started from any vertex outside $B(x, R_1)$ has probability at most ε to visit $B(x, r)$. Since every bounded harmonic function in $\mathbb{Z}^d$ is constant, there exists $R_2 > R_1$ such that any harmonic function u on $B(x, R_2)$ that takes values in $[0, 1]$, satisfies the Harnack inequality

$$\sup_{y,z\in B(x,R_1+1)} \big|u(y) - u(z)\big| < \varepsilon\,. \tag{4.11}$$

(See, e.g., Theorem 1.7.1(a) in Lawler [7].) In particular, when $\Upsilon_R \cap B(x, R_2) = \varnothing$, we can apply this to the harmonic function $y \mapsto \mu_y(A_R;\Upsilon_R)$ to obtain

$$\sup_{y,y'\in B(x,R_1+1)} \big|\mu_y(A_R;\Upsilon_R) - \mu_{y'}(A_R;\Upsilon_R)\big| < \varepsilon \quad \text{provided that } \Upsilon_R \cap B(x, R_2) = \varnothing\,. \tag{4.12}$$

We need to show that $\mu_x(A_R, \Upsilon_R)$ is close to $\mu_x^H(A_R, \Upsilon_R)$. To this end, note that by the definition of R_1, when a simple random walk on H from x first exits $B(x, R_1)$, it has probability at most ε to revisit $B(x, r)$. On the event that it does not, it may be considered as a random walk in $\mathbb{Z}^d$ which does not visit $B(x, r)$. By (4.12) it follows that

$$\big|\mu_x^H(A_R;\Upsilon_R) - \mu_x(A_R;\Upsilon_R)\big| \le 2\varepsilon\,, \tag{4.13}$$

on the event $\Upsilon_R \cap B(x, R_2) = \varnothing$. Finally take $R_3 > R_2$ sufficiently large so that

$$\forall R \ge R_3, \qquad \mathbf{P}[\Upsilon_R \cap B(x, R_2) \ne \varnothing] < \varepsilon\,.$$

From (4.13) it follows that for $R \geq R_3$,

$$\left| \mathbf{E}\big[\mu_x^H(A_R; \Upsilon_R)\big] - \mathbf{E}\big[\mu_x(A_R; \Upsilon_R)\big] \right| < 3\varepsilon\,.$$

Combined with (4.10) and (4.9) this shows that for sufficiently large R

$$\big|\mathbf{P}[\mathcal{S}, A] - \mathbf{P}[\mathcal{S}]\,\mathbf{P}[A]\big| < 4\varepsilon.$$

As $\varepsilon > 0$ was arbitrary, we conclude that A and $\mathcal{S}$ are independent, which proves that $\mathcal{U}$ has trivial left tail. By symmetry, the right tail is trivial too. □

Remark 4.6. The above proof of left-tail triviality for $\mathcal{U}$ is valid for the wired USF in any transient graph G such that there are no nonconstant bounded harmonic functions and a.s. each component of the USF has one end. Only the one-end property is needed for triviality of the remote tail. (For recurrent graphs, the USF is a.s. a tree, whence obviously the USF relation has trivial left, right and remote tails.)

The following lemma will be needed in the proof of the last statement of Theorem 1.1.

LEMMA 4.7. *Let $D \subset \mathbb{Z}^d$ be a finite connected set with a connected complement, and denote by $\widehat{D^c}$ the subgraph of $\mathbb{Z}^d$ spanned by the vertices in $\mathbb{Z}^d \setminus D$. Let F be the* USF *on $\mathbb{Z}^d$, and denote by Γ_D the event that there are no oriented edges in F from D^c to D. Then the distribution of $F \cap \widehat{D^c}$ conditioned on Γ_D, is the same as the distribution of the wired* USF *in the graph $\widehat{D^c}$.*

Proof. We first consider a finite version of this statement. Let $G = (V, E)$ be a finite connected graph, $\rho \in V$ a distinguished vertex, and $D \subset V \setminus \{\rho\}$. Let G_- be the subgraph of G spanned by $V \setminus D$. Denote by $S(G)$ the set of spanning trees in G, and let Γ_D^* be the set of $t \in S(G)$ such that for every $w \in D^c$, the path in t from w to ρ is disjoint from D.

Claim. Let T be a uniform spanning tree (UST) in G. Assume that $\Gamma_D^* \neq \varnothing$. Then conditioned on $T \in \Gamma_D^*$, the edge set $T \cap G_-$ is a UST in G_-.

To prove the claim, fix two trees $t_1, t_2 \in S(G_-)$, and for every $t \in \Gamma_D^*$ that contains t_1, define $t' := (t \setminus t_1) \cup t_2$. For each vertex $v \in V$ there is a path in t' from v to ρ, that uses edges of $(t \setminus t_1)$ until it reaches D^c, and then uses edges of t_2. Note that t cannot contain any edge of $t_2 \setminus t_1$, because t_1 plus any edge of $\widehat{D}^c$ outside t_1 contains a circuit. Thus t and t' have the same number of edges. It follows that $t' \in S(G)$ and moreover, $t' \in \Gamma_D^*$. The map $t \mapsto t'$, is a bijection, because $t' \mapsto (t' \setminus t_2) \cup t_1$ is its inverse. This shows that t_1 and t_2 have the same number of extensions to spanning trees of G that are in Γ_D^*. Moreover any $t \in \Gamma_D^*$ is an extension of some $t_1 \in S(G_-)$. In other words, we have established the claim.

The lemma follows by consideration of the uniform spanning forest on $\mathbb{Z}^d$ as a weak limit of uniform spanning trees in finite subgraphs (with wired complements), where ρ is chosen as the wired vertex. □

Remark 4.8. If a finite connected set $D \subset \mathbb{Z}^d$ has a connected complement, then the lemma above implies that a.s., every component of the wired USF in $\mathbb{Z}^d \setminus D$ has one end. The proof of Theorem 4.5 can then be adapted to show that the wired USF relation in $\mathbb{Z}^d \setminus D$ has trivial remote, right, and left, tail σ-fields. Note that this can also be inferred from Remark 4.6. To verify that every bounded harmonic function on $\mathbb{Z}^d \setminus D$ is constant, we recall that the existence of nonconstant bounded harmonic functions on a graph is equivalent to the existence of two disjoint sets of vertices A, B such that with positive probability the random walk eventually stays in A, and the same holds for B. (If such A and B exist, then define the harmonic function $h(v)$ as the probability that a simple random walk started from v eventually stays in A. If h is a bounded harmonic function with $\sup h = 1$, $\inf h = 0$, say, then we may take $A := h^{-1}([0, 1/4])$ and $B := h^{-1}([3/4, 1])$.) This criterion is clearly unaffected by the removal of D from $\mathbb{Z}^d$.

5. Proofs of main results

Proof of Theorem 1.1. Denote $m = \lfloor \frac{d-1}{4} \rfloor$. By applying Corollary 3.4 to $m + 1$ independent copies of the USF relation, and invoking Theorems 4.2 and 4.5, we infer that a.s., every two vertices in $\mathbb{Z}^d$ are related by the composition of these $m + 1$ copies. Now Theorem 4.1 yields the upper bound $\max\{N(x, y) : x, y \in \mathbb{Z}^d\} \le m$ a.s.

For the final assertion of the theorem, it suffices to show that for every $x, y \in \mathbb{Z}^d$ and every $r > 0$, the event $\Xi(x, y, r, m)$ that there is a path in $\mathbb{Z}^d \setminus B_r$ from $T(x)$ to $T(y)$ with at most m edges outside F, satisfies

$$\mathbf{P}[\Xi(x, y, r, m)] = 1\,. \tag{5.1}$$

Let $\Delta_d(r)$ denote the collection of finite, connected sets $D \subset \mathbb{Z}^d$ such that $B_r \subset D$ and D^c is connected. For $D \in \Delta_d(r)$, denote by Γ_D the event that there are no oriented edges in F from D^c to D. We claim that for every $r > 0$,

$$\mathbf{P}\Big[\bigcup_{D \in \Delta_d(r)} \Gamma_D\Big] = 1\,. \tag{5.2}$$

Indeed, let $\overrightarrow{B}_r$ denote the set of vertices $v \in \mathbb{Z}^d$ such that the oriented path from v to infinity in F enters B_r. Since a.s. each of the USF components has one end, $\overrightarrow{B}_r$ and its complement are connected, and $\overrightarrow{B}_r$ is a.s. finite. Thus $\Gamma_{\overrightarrow{B}_r}$ occurs and (5.2) holds.

Therefore, to prove (5.1), it suffices to show that

$$\forall D \in \Delta_d(r), \qquad \mathbf{P}\big[\Xi(x,y,r,m)|\Gamma_D\big] = 1\,. \tag{5.3}$$

Fix $D \in \Delta_d(r)$, let G_- be the subgraph of $\mathbb{Z}^d$ spanned by the vertices in $\mathbb{Z}^d \setminus D$, and let F_- denote the wired USF on G_-. By Lemma 4.7, conditioned on Γ_D, the law of $F \cap G_-$ is the same as that of F_-. Consequently, it suffices to show that a.s. every pair of vertices in G_- is connected by a path in G_- with at most m edges outside of F_-. The USF relation on G_- has trivial left, right and remote tails, by Remark 4.8. It is also clear that this relation has stochastic dimension 4. Therefore, the above proof for $\mathbb{Z}^d$ applies also to G_-, and shows that every pair of vertices in G_- is connected by a path in G_- with at most m edges outside F_-. Since r is arbitrary and D is an arbitrary set in $\Delta_d(r)$, this proves the last assertion of the theorem, and also completes the proof for $5 \le d \le 8$.

For the case $d > 8$, it remains only to prove the lower bound on $\max N(x,y)$. This is done in the following proposition. □

PROPOSITION 5.1. *The inequality*

$$\mathbf{P}\big[N(x,y) \le k\big] \preccurlyeq \langle xy\rangle^{\,4k+4-d}$$

holds for all $x, y \in \mathbb{Z}^d$ and all $k \in \mathbb{N}$.

Proof. The proposition clearly holds for $k \ge \lfloor (d-1)/4 \rfloor$. So assume that $k < \lfloor (d-1)/4 \rfloor$. Consider a sequence $\{x_j, y_j\}_{j=0}^k$ in $\mathbb{Z}^d$ with $x_0 = x$, $y_k = y$ and $|y_j - x_{j+1}| = 1$ for $j = 0, 1, \dots, k-1$. Let $A = A(x_0, y_0, \dots, x_k, y_k)$ be the event that $y_j \in T(x_j)$ for $j = 0, 1, \dots, k$ and $y_j \notin T(x_i)$ if $j \ne i$. Observe that

$$\big\{N(x,y) = k\big\} \subset \bigcup A(x_0, y_0, \dots, x_k, y_k)\,,$$

where the union is over all such sequences $x_0, y_0, \dots, x_k, y_k$. To estimate the probability of A, we run Wilson's algorithm rooted at infinity, begining with random walks started at $x_0, y_0, \dots, x_k, y_k$. For A to hold, for each $j = 0, \dots, k$, the random walk started at x_j must intersect the path of the random walk started at y_j. Hence, as in (4.1),

$$\mathbf{P}[A] \preccurlyeq \prod_{j=0}^{k} \langle x_j y_j\rangle^{\,4-d}\,. \tag{5.4}$$

Since x_{j+1} and y_j are adjacent, this implies

$$\mathbf{P}\big[N(x,y) = k\big] \preccurlyeq \sum \prod_{j=0}^{k} \langle x_j x_{j+1}\rangle^{\,4-d}\,, \tag{5.5}$$

where the sum extends over all sequences $\{x_j\}_{j=0}^{k+1} \subset \mathbb{Z}^d$ such that $x_0 = x$ and $x_{k+1} = y$.

We use Lemma 2.10 repeatedly in (5.5), first to sum over x_k, then over x_{k-1}, and so on, up to x_1, and obtain $\mathbf{P}\big[N(x,y)=k\big] \preccurlyeq \langle xy\rangle^{\,4k+4-d}$, which completes the proof. □

Proof of Theorem 1.2. The case $d \le 4$ is from Pemantle [11]. The case $d = 5, 6, 7, 8$ follows from Theorem 1.1. For $d \ge 9$ and fixed $x, y \in \mathbb{Z}^d$ and $r > 0$ we infer from (5.4) that

$$\mathbf{P}\big[\exists v \in T(x),\, \exists w \in T(y) : |v| \ge n,\, |w| \ge n,\, |v-w| \le r,\, T(x) \ne T(y)\big] \le \sum_{v,w} \mathbf{P}\big[A(x,v,y,w)\big] \preccurlyeq \sum_{v,w} \langle xv\rangle^{\,4-d} \langle yw\rangle^{\,4-d},$$

where the sums are over v, w satisfying $|v|, |w| \ge n$ and $|v-w| \le r$. The latter sum is bounded by $C(x,y,r)\, n^{8-d}$, by Lemma 2.10. □

6. Further examples and remarks

We now present some examples of relations having a stochastic dimension, and several questions.

Example 6.1. Fix some $\alpha \in (0,d)$. Let $\mathcal{R} \subset \mathbb{Z}^d \times \mathbb{Z}^d$ be the random relation such that $\{x\mathcal{R}y\}_{\{x,y\}\subset\mathbb{Z}^d}$ are independent events, and $\mathbf{P}[x\mathcal{R}y] = \langle xy\rangle^{\,-\alpha}$. Then $\mathcal{R}$ has stochastic dimension $d-\alpha$. This relation corresponds to *long range percolation* in $\mathbb{Z}^d$, where the degree of every vertex is infinite and the infinite cluster contains all vertices of $\mathbb{Z}^d$. Theorem 2.4 implies that the intrinsic diameter of this graph is $\lceil \frac{d}{d-\alpha} \rceil$ almost surely.

Example 6.2. Let $d \ge 3$, and for every $x \in \mathbb{Z}^d$, let S^x be a simple random walk starting from x, with $\{S^x\}_{x\in\mathbb{Z}^d}$ independent. Let $x\mathcal{S}y$ be the relation $\exists n \in \mathbb{N}$ such that $S^x(n) = y$. It is easily seen that $\mathcal{S}$ has stochastic dimension 2. From Theorems 4.2, 3.4 and 4.5, we infer that a simple random walk a.s. intersects each component of an independent USF if and only if $d \le 6$.

Example 6.3. For every $x \in \mathbb{Z}^d$, let S^x be as above. Define a relation $\mathcal{Q}$ where $x\mathcal{Q}y$ if and only if $\exists n \in \mathbb{N}$ such that $S^x(n) = S^y(n)$. Then $\mathcal{Q}$ has stochastic dimension 2 (provided $d \ge 2$). The proof is left to the reader.

Example 6.4. Fix $d \ge 2$. For every $x \in \mathbb{Z}^d$, let $S^x_*(\cdot)$ be a continuous-time simple random walk starting from x, where particles walk independently until they meet; when two particles meet, they coalesce, and the resulting particle carries both labels. Let $x\mathcal{W}y$ be the relation $\exists t > 0$ such that $S^x_*(t) = y$. Then $\mathcal{W}$ has stochastic dimension 2.

Example 6.5 (USF generated by other walks). Let $\{S_n\}_{n\ge0}$ be a symmetric random walk on $\mathbb{Z}^d$; i.e., the increments $S_n - S_{n-1}$ are symmetric

i.i.d. $\mathbb{Z}^d$-valued variables. Say that this random walk has *Greenian index* α if the Green function $G_S(x,y) := \sum_{n=0}^{\infty} \mathbf{P}[S_n = y | S_0 = x]$ satisfies $G_S(x,y) \asymp \langle xy \rangle^{\alpha - d}$ for all $x, y \in \mathbb{Z}^d$. It is well-known that random walks with bounded increments of mean zero have Greenian index 2, provided that $d \geq 3$; for $d = 3$, the boundedness assumption may be relaxed to finite variance (see Spitzer [12]). Williamson [14] shows that for $0 < \alpha < \min\{2, d\}$, if a random walk $\{S_n\}$ in $\mathbb{Z}^d$ with mean zero increments satisfies $\mathbf{P}[S_n - S_{n-1} = x] \sim c|x|^{-d-\alpha}$ (i.e., the ratio of the two expressions here tends to 1 as $|x| \to \infty$), then this random walk has Greenian index α. We can use Wilson's method rooted at infinity to generate the wired spanning forest (WSF) based on a symmetric random walk with Greenian index α. (Note: this will not be a subgraph of the usual graph structure on $\mathbb{Z}^d$.) The arguments of the preceding section extend to show that the relation determined by the components of this WSF will have stochastic dimension 2α.

Example 6.6 (Minimal Spanning Forest). Let the edges of $\mathbb{Z}^d$ have independent and identically distributed weights w_e, which are uniform random variables in $[0,1]$. The *minimal spanning forest* is the subgraph of $\mathbb{Z}^d$ obtained by removing every edge that has the maximal weight in some cycle; see, e.g., Newman and Stein [10].

CONJECTURE 6.7. *Let $x \mathcal{M} y$ if x and y are in the same minimal spanning forest component. Then $\mathcal{M}$ has stochastic dimension 8 in $\mathbb{Z}^d$ if $d \geq 8$. This is a variation on a conjecture of Newman and Stein* [10].

Remark 6.8. It is natural to ask how the notion of stochastic dimension is related to other notions of dimension for subsets of a lattice. In this direction, we state the following. Let $\mathcal{L} \subset \mathbb{Z}^d \times \mathbb{Z}^d$ be a random relation. Denote $\Gamma_o = \Gamma_o(\mathcal{L}) := \{z : o \mathcal{L} z\}$; if $\mathcal{L}$ is an equivalence relation, then $\Gamma_o(\mathcal{L})$ is the equivalence class of the origin o, but we do not assume this.

PROPOSITION 6.9. *Suppose that $\mathcal{L} \subset \mathbb{Z}^d \times \mathbb{Z}^d$ has stochastic dimension $\alpha \in (0, d)$. Denote by η_n the number of vertices in $\Gamma_o(\mathcal{L}) \cap B(o, n)$. Then*

(i) *The laws of η_n / n^α are tight, but do not converge weakly to a point mass at 0.*

(ii) *With positive probability,* $\limsup_n \frac{\log \eta_n}{\log n} = \alpha$.

(iii) *If, moreover, $\mathcal{L}$ has a trivial left tail, then*

$$\limsup_n \frac{\log \eta_n}{\log n} = \alpha \text{ a.s.}$$

Proof. (i) The definition of stochastic dimension and a standard calculation yield that $\mathbf{E}\eta_n \asymp n^\alpha$ and $\mathbf{E}\eta_n^2 \asymp n^{2\alpha}$. These estimates imply the asserted

tightness, and also that

$$\inf_n \mathbf{P}(\eta_n \geq \mathbf{E}\eta_n/2) > 0 \tag{6.1}$$

(see, e.g., Kahane [5, p. 8]).

(ii) The bound $\mathbf{E}\eta_n \preccurlyeq n^\alpha$ implies that $\sum_k \eta_{2^k}/(2^{k\alpha}k^2) < \infty$ a.s. Thus, monotonicity of η_n yields that $\limsup_n \frac{\log \eta_n}{\log n} \leq \alpha$ a.s. The lower bound follows from (6.1).

(iii) This follows from (ii) and the definition of the left tail. □

When $\mathcal{L}$ is the USF relation and $d > 4$, we have $\alpha = 4$, and the limsup in (iii) can be replaced by a limit (we omit the proof). With more work, it can be shown that the "discrete Hausdorff dimension" (in the sense of Barlow and Taylor [1]) of the USF component Γ_o is also 4, almost surely.

A harder task is to prove existence of the *scaling limit* for the USF in dimension $d > 4$, and show that this limiting object has Hausdorff dimension 4 almost surely.

An easily stated problem in this direction, is to show that

$$\mathbf{P}[\Gamma_o \cap B(x,k) \neq \varnothing] \asymp (k/|x|)^{d-4} \tag{6.2}$$

for $k < |x|$. At present, we can only establish such an estimate up to a power of $\log x$. Note that (6.2), if true, depends on properties of the USF beyond its stochastic dimension; the analog of (6.2) fails for the relation considered in Example 6.1.

Example 6.10. As noted in the introduction, for $5 \leq d \leq 8$, the USF defines a translation invariant random partition of $\mathbb{Z}^d$ with infinitely many connected components, where any two components come within distance 1 infinitely often. A partition with these properties exists for any $d \geq 3$: In the hyperplane $H := \{x \in \mathbb{Z}^d : x_d = 0\}$, let L_0 be the lattice consisting of all vertices with all coordinates even, and let L be a random translate of L_0 in H, chosen uniformly among the 2^{d-1} possibilities. Let F_0 be the union of all lines perpendicular to H which contain a vertex of L, and complete F_0 to a spanning forest F using Wilson's algorithm.

Example 6.11. Consider the random relation $\mathcal{P} \subset \mathbb{Z}^2 \times \mathbb{Z}^2$, where $x\mathcal{P}y$ if and only if x and y are in the same connected component of Bernoulli bond percolation at the critical parameter $p = p_c = 1/2$. It is known that a.s. all the components will be finite. If $\mathcal{L}, \mathcal{R} \subset \mathbb{Z}^2 \times \mathbb{Z}^2$ and $\big|\{y : x\mathcal{L}y\}\big| < \infty$ and $\big|\{y : x\mathcal{R}y\}\big| < \infty$ for all $x \in \mathbb{Z}^2$, then $\{y : x\mathcal{L}\mathcal{R}y\}$ is finite too. After inductive application of Theorem 2.4 to independent copies of $\mathcal{P}$, it follows that $\mathcal{P}$ cannot have a positive stochastic dimension. However, the Russo-Seymour-Welsh Theorem implies that $\mathbf{P}[x\mathcal{P}y] \geq \langle xy \rangle^{-\beta}$ for some $\beta < \infty$.

Remark 6.12. Our results on the USF in $\mathbb{Z}^d$ extend readily to other Cayley graphs of polynomial growth. Let G be such a graph, and let o be a vertex of G. By Gromov [3], G satisfies $|B(o,r)| \asymp r^d$ for some integer d (see also Woess [16, Lemma 3.14 and Th. 3.16]). Consequently, we may find an $a > 2$ such that $\left|B(o,a^{n+1}) \setminus B(o,a^n)\right| \asymp a^{nd}$ for $n \in \mathbb{N}$. Thus, in our above proofs, we may work with shells based on powers of a, in place of the shells H_n^N. By Hebisch and Saloff-Coste [4], the Green function estimates which we used in $\mathbb{Z}^d$ apply to G. Moreover, any bounded harmonic function on G is a constant (see, e.g., [6]). This implies, in particular, that the free and the wired USF on G are the same, by Theorem 7.3 of BLPS [2] (so it is clear what we mean by the USF on G). By Theorem 10.1 of BLPS [2] a.s. each component of the USF on G has one end, unless G is quasi-isometric to $\mathbb{Z}$ (in which case G is recurrent, $d = 1$, the USF is a tree a.s. and our results clearly hold). With these basic facts established, the proofs go through just as for $\mathbb{Z}^d$, with the case $d \le 4$ handled by Corollary 9.6 of BLPS [2].

Acknowledgment. We thank Russ Lyons for useful discussions, and Jim Pitman for a reference. H. Kesten thanks the Hebrew University for its hospitality and for appointing him as Schonbrunn Visiting Professor during the Spring of 1997, when this project was started. We also thank the referee for helpful comments.

The Weizmann Institute of Science, Rehovot, Israel
E-mail address: itai@wisdom.weizmann.ac.il
Web page: http://www.widsom.weizmann.ac.il/∼itai/

Cornell University, Ithaca, NY
E-mail address: kesten@math.cornell.edu

University of California, Berkeley, CA
E-mail address: peres@stat.berkeley.edu
Web page: http://www.stat.berkeley.edu/∼peres/

Microsoft Corporation, Redmond, WA
E-mail address: schramm@microsoft.com
Web page: http://research.microsoft.com/∼schramm/

References

[1] M. Barlow and S. J. Taylor, Defining fractal subsets of $\mathbb{Z}^d$, *Proc. London Math. Soc.* **64**, 125–152.

[2] I. Benjamini, R. Lyons, Y. Peres, and O. Schramm, Uniform spanning forests, *Ann. Probab.* **29** (2001), 1–65.

[3] M. Gromov, Groups of polynomial growth and expanding maps, *Inst. Hautes Études Sci. Publ. Math.* **53** (1981), 53–73.

[4] W. Hebisch and L. Saloff-Coste, Gaussian estimates for Markov chains and random walks on groups, *Ann. Probab.* **21** (1993), 673–709.

[5] J.-P. Kahane, *Some Random Series of Functions*, Second edition, Cambridge Univ. Press, Cambridge, U.K. (1985).

[6] V. A. Kaĭmanovich and A. M. Vershik, Random walks on discrete groups: boundary and entropy, *Ann. Probab.* **11** (1983), 457–490.

[7] G. F. Lawler, *Intersections of Random Walks*, Birkhäuser Boston Inc., Boston, MA (1991).

[8] R. Lyons, A bird's-eye view of uniform spanning trees and forests, in *Microsurveys in Discrete Probability* (Princeton, NJ, 1997), *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.* **41**, 135–163, A. M. S., Providence, RI (1998).

[9] R. Lyons, Y. Peres, and O. Schramm, Markov chain intersections and the loop-erased walk, *Ann. Inst. H. Poincaré, Probab. et Stat.* **39** (2003), 779–791.

[10] C. M. Newman and D. Stein, Spin-glass model with dimension-dependent ground state, *Phys. Rev. Lett.* **72** (1994), 2286–2289.

[11] R. Pemantle, Choosing a spanning tree for the integer lattice uniformly, *Ann. Probab.* **19** (1991), 1559–1574.

[12] F. Spitzer, *Principles of Random Walk*, second edition, *Grad. Texts in Math.* **34**, Springer-Verlag, New York (1976).

[13] H. von Weizsäcker, Exchanging the order of taking suprema and countable intersections of σ-algebras, *Ann. Inst. H. Poincaré, Probab. et Stat.* **19** (1983), 91–100.

[14] J. A. Williamson, Random walks and Riesz kernels, *Pacific J. Math.* **25** (1968), 393–415.

[15] D. B. Wilson, Generating random spanning trees more quickly than the cover time, in *Proc. of the Twenty-eighth Annual ACM Symposium on the Theory of Computing* (Philadelphia, PA, 1996), 296–303, ACM, New York (1996).

[16] W. Woess, *Random Walks on Infinite Graphs and Groups*, *Cambridge Tracts in Math.* **138**, Cambridge Univ. Press, Cambridge, U.K. (2000).

(Received July 29, 2001)

The Annals of Probability
2003, Vol. 31, No. 4, 1970–1978
© Institute of Mathematical Statistics, 2003

FIRST PASSAGE PERCOLATION HAS SUBLINEAR DISTANCE VARIANCE

BY ITAI BENJAMINI, GIL KALAI AND ODED SCHRAMM

Weizmann Institute of Science, Hebrew University and Microsoft Research

Let $0 < a < b < \infty$, and for each edge e of $\mathbb{Z}^d$ let $\omega_e = a$ or $\omega_e = b$, each with probability $1/2$, independently. This induces a random metric dist_ω on the vertices of $\mathbb{Z}^d$, called first passage percolation. We prove that for $d > 1$, the distance $\mathrm{dist}_\omega(0, v)$ from the origin to a vertex v, $|v| > 2$, has variance bounded by $C|v|/\log|v|$, where $C = C(a, b, d)$ is a constant which may only depend on a, b and d. Some related variants are also discussed.

1. Introduction. Consider the following model of first passage percolation. Fix some $d = 2, 3, \ldots$ and let $E = E(\mathbb{Z}^d)$ denote the set of edges in $\mathbb{Z}^d$. Also fix numbers $0 < a < b < \infty$. Let $\Omega := \{a, b\}^E$ carry the product measure, where $\mathbf{P}[\omega_e = a] = \mathbf{P}[\omega_e = b] = 1/2$ for each $e \in E$. Given $\omega = (\omega_e : e \in E) \in \Omega$ and vertices $v, u \in \mathbb{Z}^d$, let $\mathrm{dist}_\omega(v, u)$ denote the least distance from v to u in the metric induced by ω; that is, the infimum of $\sum_{e\in\alpha} \omega_e$, where α ranges over all finite paths in $\mathbb{Z}^d$ from u to v. Let $|v| := \|v\|_1$ for vertices $v \in \mathbb{Z}^d$.

THEOREM 1. *There is a constant $C = C(d, a, b)$ such that for every $v \in \mathbb{Z}^d$, $|v| \geq 2$,*

$$\mathrm{var}\big(\mathrm{dist}_\omega(0, v)\big) \leq C\frac{|v|}{\log|v|}.$$

Kesten [12] used martingale inequalities to prove $\mathrm{var}(\mathrm{dist}_\omega(0, v)) \leq C|v|$ and proved some tail estimates. Talagrand [17] used his "convexified" discrete isoperimetric inequality to prove that for all $t > 0$,

$$\mathbf{P}\big[|\mathrm{dist}_\omega(0, v) - M| \geq t\sqrt{|v|}\,\big] \leq C\exp(-t^2/C), \tag{1}$$

where M is the median value of $\mathrm{dist}_\omega(0, v)$. (Both Kesten's and Talagrand's results apply to more general distributions of the edge lengths ω_e.) "Novice readers might expect to hear next of a central limit theorem being proved," writes Durrett [7], describing Kesten's results, "however physicists tell us... that in two dimensions the standard deviation... is of order $|v|^{1/3}$." Recent remarkable work [1, 6, 9] not only supports this prediction, but suggests what the limiting distribution and large deviation behavior is. The case of a certain variant of *oriented* first passage

Received April 2002; revised October 2002.
AMS 2000 subject classifications. Primary 60K35; secondary 60B15, 28A35, 60E15.
Key words and phrases. Hypercontractive, harmonic analysis, discrete harmonic analysis, discrete cube, random metrics, discrete isoperimetric inequalities, influences.

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_26, © Springer Science+Business Media, LLC 2011

percolation is settled by Johansson [9]. For lower bounds on the variance in two-dimensional first passage percolation, see [14, 15].

As in Kesten's and Talagrand's earlier results, the most essential feature about first passage percolation which we use is that the number of edges $e \in E$ such that modifying ω_e increases $\mathrm{dist}_\omega(0, v)$ is bounded by $C|v|$.

Another essential ingredient in the current article is the following extension by Talagrand [16] of an inequality by Kahn, Kalai and Linial [10]. Let J be a finite index set. For $j \in J$ and $\omega \in \{a, b\}^J$, let $\sigma_j\omega$ be the element of $\{a, b\}^J$ which is different from ω only in the jth coordinate. For $f : \{a, b\}^J \to \mathbb{R}$ set $\sigma_j f := f \circ \sigma_j$ and

$$\rho_j f(\omega) := \frac{f(\omega) - \sigma_j f(\omega)}{2}.$$

Talagrand's ([16], Theorem 1.5) inequality is

$$\mathrm{var}(f) \le C \sum_{j \in J} \frac{\|\rho_j f\|_2^2}{1 + \log(\|\rho_j f\|_2 / \|\rho_j f\|_1)}, \tag{2}$$

where C is a universal constant. [In Section 4 we supply a direct proof of (2) from the Bonami–Beckner inequality. A very reasonable upper bound for C can be obtained from this proof. We also explain there how (2) easily implies (1).]

The basic idea in the proof of Theorem 1 is to apply this inequality to $f(\omega) := \mathrm{dist}_\omega(0, v)$. For this, we wanted to show that, roughly, $\mathbf{P}[\rho_e f(\omega) \neq 0]$ is small, except for a small number of edges e. However, since we were not able to prove this, we had to resort to an averaging trick.

Theorem 1 should hold for other models of first passage percolation, where the edge lengths have more general distributions. The relevant result of [10], as well as (2) that we use rely on the Bonami–Beckner [2, 4] inequality, which holds for $\{a, b\}^n$, but fails on some more general product spaces. Bourgain, Kahn, Kalai, Katznelson and Linial [5] extended some of these results to general product spaces; see also [8]. Talagrand's ([16], Theorem 1.5) result applies to product measures on $\{a, b\}^n$, which are not necessarily uniform. Beckner [2] used his inequality to derive a similar result for the Gaussian measure on $\mathbb{R}^n$, and an analog for Talagrand's inequality (2) for the Gaussian measure (pointed out to us by Assaf Naor) can be found in [3]. In the present article, we prefer simplicity to generality, but it will be interesting to extend the theorem to more general distributions. We refer the reader also to the book by Ledoux [13] for a general view of relevant techniques and knowledge.

It seems even more important to put effort into the fundamental task of lowering the upper bound from $C|v|/\log|v|$ to $C|v|^{1-\varepsilon}$, $\varepsilon > 0$.

To first illustrate the basic technique in a simpler setting, we also prove the following theorem about the variance of the first passage percolation diameter in

graphs with symmetries. If $G = (V(G), E(G))$ is a finite connected graph and $\omega : E(G) \to \{a, b\}$, define

$$\operatorname{diam}_\omega(G) := \max_{v,u \in V(G)} \operatorname{dist}_\omega(v, u).$$

(Although we have defined $\operatorname{dist}_\omega$ only for $\mathbb{Z}^d$, the definition obviously extends to arbitrary graphs.) Also, let $\operatorname{diam}(G)$ denote the diameter of G in the graph metric; that is, $\operatorname{diam}(G) = \operatorname{diam}_1(G)$.

THEOREM 2. *Let $G = (V(G), E(G))$ be a finite connected graph and let Γ be the group of automorphisms of G. Set $Y := \min\{|\Gamma e| : e \in E(G)\}$; that is, the cardinality of the smallest orbit of edges under Γ. Let $\omega \in \{a, b\}^{E(G)}$ be random uniform. Then*

$$\operatorname{var}(\operatorname{diam}_\omega G) \le C \frac{(b-a)^2 (b/a) \operatorname{diam} G}{1 + \log \max\{1, (a/b)Y/\operatorname{diam} G\}},$$

where C is a universal constant.

Also consider first passage percolation on an $m \times n$ discrete torus $(\mathbb{Z}/m\mathbb{Z}) \times (\mathbb{Z}/n\mathbb{Z})$. Among all simple closed paths whose projection onto the first coordinate has degree 1 (i.e., every edge in $\mathbb{Z}/m\mathbb{Z}$ is traversed by the projected walk one time more in one direction than in the other), there is at least one which has minimal ω-length. Call this length the circumference of ω. One can apply the proof of Theorem 2 to show that the variance of the circumference is $o(n)$ when $m, n \to \infty$.

2. Proof of Theorem 2. Let $f(\omega) := \operatorname{diam}_\omega(G)$. Let $u_1, u_2 \in V(G)$ be a pair of vertices where $\operatorname{dist}_\omega(u_1, u_2) = \operatorname{diam}_\omega(G)$ and let β be a path from u_1 to u_2 such that $\sum_{e \in \beta} \omega_e = \operatorname{diam}_\omega(G)$. We use some arbitrary but fixed method for choosing between all possible choices for the triple (u_1, u_2, β). Clearly, $\operatorname{diam}_\omega(G) \le b \cdot \operatorname{diam}(G)$. Therefore,

$$|\beta| \le (b/a) \operatorname{diam}(G), \tag{3}$$

where $|\beta|$ denotes the number of edges in β.

Let $e \in E(G)$. Note that if $\rho_e f(\omega) < 0$, then we must have $e \in \beta$. By (3), this gives

$$\sum_{e \in E(G)} \mathbf{P}[\rho_e f(\omega) < 0] \le \mathbf{E}[|\beta|] \le (b/a) \operatorname{diam}(G). \tag{4}$$

Fix $e \in E(G)$. By symmetry, $\mathbf{P}[\rho_e f < 0] = \mathbf{P}[\rho_{e'} f < 0]$ for every $e' \in \Gamma e$. Consequently,

$$|\Gamma e| \mathbf{P}[\rho_e f < 0] \le \sum_{e' \in E(G)} \mathbf{P}[\rho_{e'} f < 0] \le (b/a) \operatorname{diam}(G). \tag{5}$$

Now clearly, $\mathbf{P}[\rho_e f \neq 0] = 2\mathbf{P}[\rho_e f < 0]$ and $\|\rho_e f\|_\infty \leq (b-a)/2$. Therefore,

$$\|\rho_e f\|_2^2 \leq (1/2)(b-a)^2 \mathbf{P}[\rho_e f < 0]. \tag{6}$$

By Cauchy–Schwarz,

$$\|\rho_e f\|_1 \leq \sqrt{\mathbf{P}[\rho_e f \neq 0]}\|\rho_e f\|_2. \tag{7}$$

By (2), we have

$$\operatorname{var}(f) \leq C \frac{\sum_{e\in E(G)} \|\rho_e f\|_2^2}{1 + \min_{e\in E(G)} \log(\|\rho_e f\|_2/\|\rho_e f\|_1)}.$$

To estimate the numerator, we use (6) and (4), and for the denominator, (5) and (7). The theorem easily follows. □

3. Proof of Theorem 1. If $v, u \in \mathbb{Z}^d$, and α is a path from v to u, then α is called an *ω-geodesic* if it minimizes ω-length; that is, $\operatorname{dist}_\omega = \sum_{e\in\alpha} \omega_e$. Given ω, let γ be an ω-geodesic from 0 to v. Although there may be more than one such geodesic, we require that γ depends only on ω. (E.g., we may use an arbitrary deterministic choice among any possible collection of ω-geodesics.)

The general strategy for the proof of Theorem 1 is as for Theorem 2. However, the difficulty is that there is not enough symmetry to get a good bound on $\mathbf{P}[\rho_e \operatorname{dist}_\omega(0, v) < 0]$. It would have been enough to show that $\mathbf{P}[e \in \gamma] < C|v|^{-1/C}$ holds with the exception of at most $C|v|/\log|v|$ edges, for some constant $C > 0$, but we could not prove this. Therefore, we need an averaging argument, for which the following lemma is useful.

LEMMA 3. *There is a constant $c > 0$ such that for every $m \in \mathbb{N}$ there is a function*

$$g = g_m : \{a, b\}^{m^2} \to \{0, 1, 2, \ldots, m\}$$

satisfying

$$\|\sigma_j g - g\|_\infty \leq 1$$

for every $j = 1, 2, \ldots, m^2$ and

$$\max_y \mathbf{P}[g(x) = y] \leq c/m,$$

where x is random-uniform in $\{a, b\}^{m^2}$.

PROOF. Assume $a = 0$ and $b = 1$ for simplicity of notation. Let $k : \mathbb{N} \to \{0, 1, 2, \ldots, m\}$ be the function satisfying $k(0) = 0$, $k(j+1) = k(j) + 1$ when $j \in \bigcup_{s=0}^{\infty}[2sm, 2sm + m - 1]$ and $k(j+1) = k(j) - 1$ for all other $j \in \mathbb{N}$. It is left to the reader to check that $g(x) = k(\|x\|_1)$ has the required properties. □

Fix some $v \in \mathbb{Z}^d$ with $|v|$ large and set $f = f(\omega) := \mathrm{dist}_\omega(0, v)$. Clearly, $f(\omega) \le b|v|$ and, therefore, $|\gamma| \le (b/a)|v|$, where $|\gamma|$ denotes the number of edges in γ. In particular, f depends on only finitely many of the coordinates in ω. Also, $|\gamma| \le (b/a)|v|$ implies

$$\sum_{e \in E} \mathbf{P}[e \in \gamma] = \mathbf{E}[|\gamma|] \le (b/a)|v|. \tag{8}$$

Fix $m := \lfloor |v|^{1/4} \rfloor$ and let $S := \{1, \dots, d\} \times \{1, \dots, m^2\}$. Let $c > 0$ and $g = g_m$ be as in Lemma 3. Given $x = (x_{i,j} : \{i, j\} \in S) \in \{0, 1\}^S$, let

$$z = z(x) := \sum_{i=1}^{d} g(x_{i,1}, \dots, x_{i,m^2})\mathbf{e}_i,$$

where $\{\mathbf{e}_1, \dots, \mathbf{e}_d\}$ is the standard basis for $\mathbb{R}^d$. Define

$$\tilde{f}(x, \omega) := \mathrm{dist}_\omega(z, v + z).$$

We think of $\tilde{f}$ as a function on the space $\tilde{\Omega} := \{a, b\}^{S \cup E}$. Since $|z| \le md$, it follows that $|\tilde{f} - f| \le 2mdb$. In particular,

$$\mathrm{var}(f) \le \mathrm{var}(\tilde{f}) + 4mdb\sqrt{\mathrm{var}(\tilde{f})} + 4m^2d^2b^2. \tag{9}$$

It therefore suffices to find a good estimate for $\mathrm{var}(\tilde{f})$.

Let $e \in E$ be some edge. We want to estimate its influence:

$$I_e(\tilde{f}) := \mathbf{P}[\sigma_e \tilde{f}(x, \omega) \ne \tilde{f}(x, \omega)] = 2\mathbf{P}[\sigma_e \tilde{f}(x, \omega) > \tilde{f}(x, \omega)].$$

Note that if the pair $(x, \omega) \in \tilde{\Omega}$ satisfies $\sigma_e \tilde{f}(x, \omega) > \tilde{f}(x, \omega)$, then e must be on every ω-geodesic from z to $v + z$. Consequently, conditioning on z and translating ω and e by $-z$ gives

$$I_e(\tilde{f}) = 2\mathbf{P}[\sigma_e \tilde{f}(x, \omega) > \tilde{f}(x, \omega)] \le 2\mathbf{P}[e - z \in \gamma]. \tag{10}$$

Let Q be the set of edges $e' \in E(\mathbb{Z}^d)$ such that $\mathbf{P}[e - z = e'] > 0$. The L^1 diameter of Q is $O(m)$. [We allow the constants in the $O(\cdot)$ notation to depend on d, a and b, but not on v.] Hence, the diameter of Q in the dist_ω metric is also $O(m)$ and, therefore, $|\gamma \cap Q| \le O(m)$. However, the lemma gives

$$\max_{z_0} \mathbf{P}[z = z_0] \le (c/m)^d.$$

By conditioning on γ and summing over the edges in $\gamma \cap Q$, we therefore get

$$\mathbf{P}[e \in \gamma + z | \gamma] \le O(1)m^{1-d}.$$

Consequently, (10) and the choice of m give

$$I_e(\tilde{f}) \le O(1)|v|^{-1/4}. \tag{11}$$

Also, (8) implies

$$\sum_{e\in E} \mathbf{P}[e - z \in \gamma | z] \le (b/a)|v|.$$

Combining this with (10) therefore gives

$$\sum_{e\in E} I_e(\tilde{f}) \le 2(b/a)|v|.$$

Applying (11) yields

$$\sum_{e\in E} \frac{I_e(\tilde{f})}{1 + |\log I_e(\tilde{f})|} \le O(1)\frac{|v|}{\log|v|}. \tag{12}$$

On the other hand, $I_s(\tilde{f}) \le b - a$ for $s \in S$. As $|S| = O(1)|v|^{1/2}$ and $\|\rho_q f\|_\infty = O(1)$ for $q \in S \cup E$, we get from (12) and (2),

$$\operatorname{var}(\tilde{f}) \le O(1) \sum_{q\in E\cup S} \frac{I_q(\tilde{f})}{1 + |\log I_q(\tilde{f})|} \le O(1)\frac{|v|}{\log|v|}.$$

Therefore, Theorem 1 now follows from (9).

As alluded to in the Introduction, the proof would have been simpler if we could show that there is a $C > 0$ such that $\mathbf{P}[\rho_e f(\omega) \ne 0] < |v|^{-C}$ holds with the possible exception of $|v|/\log|v|$ edges. It would be interesting to prove the closely related statement that the probability that γ passes within distance 1 from $v/2$ tends to zero as $|v| \to \infty$.

4. A proof of Talagrand's inequality (2). To prove (2), it clearly suffices to take $a = 0$ and $b = 1$. For $f : \{0, 1\}^J \to \mathbb{R}$, consider the Fourier–Walsh expansion of f,

$$f = \sum_{S\subset J} \hat{f}(S) u_S,$$

where $u_S(\omega) = (-1)^{S\cdot\omega}$ and $S \cdot \omega$ is shorthand for $\sum_{s\in S} \omega_s$. For each $p \in \mathbb{R}$ define the operator

$$T_p(f) := \sum_{S\subset J} p^{|S|} \hat{f}(S) u_S,$$

which is of central importance in harmonic analysis. The Bonami–Beckner [2, 4] inequality asserts that

$$\|T_p f\|_2 \le \|f\|_{1+p^2}. \tag{13}$$

Set $f_j := \rho_j f$. Because $\rho_j u_S = u_S$ if $j \in S$ and $\rho_j u_S = 0$ if $j \notin S$, we have

$$\hat{f}_j(S) = \begin{cases} \hat{f}(S), & j \in S, \\ 0, & j \notin S. \end{cases}$$

Since $\|f\|_2^2 = \sum_{S\subset J} \hat{f}(S)^2$, it follows that

$$\mathrm{var}(f) = \sum_{\varnothing\neq S\subset J} \hat{f}(S)^2 = \sum_{S\subset J}\sum_{j\in J} \hat{f}_j(S)^2/|S| = 2\sum_{j\in J}\int_0^1 \|T_p f_j\|_2^2\,dp.$$

Therefore, (13) gives

$$\mathrm{var}(f) \le 2\sum_{j\in J}\int_0^1 \|f_j\|_{1+p^2}^2\,dp. \tag{14}$$

An instance of the Hölder inequality

$$\mathbf{E}[|f_j|^{1+p^2}] \le \mathbf{E}[f_j^2]^{p^2}\mathbf{E}[|f_j|]^{1-p^2}$$

implies

$$\begin{aligned}\int_0^1 \|f_j\|_{1+p^2}^2\,dp &\le \int_0^1 \left(\mathbf{E}[f_j^2]^{p^2}\mathbf{E}[|f_j|]^{1-p^2}\right)^{2/(1+p^2)}dp\\ &= \|f_j\|_2^2\int_0^1 (\|f_j\|_1/\|f_j\|_2)^{2(1-p^2)/(1+p^2)}\,dp\\ &\le 2\|f_j\|_2^2\int_{1/2}^1 (\|f_j\|_1/\|f_j\|_2)^{2(1-p^2)/(1+p^2)}\,dp.\end{aligned}$$

Let $s(p) := 2(1-p^2)/(1+p^2)$. Since $s'(p) \le s'(1) = -2$ when $p \in [1/2, 1]$, the above gives

$$\begin{aligned}\int_0^1 \|f_j\|_{1+p^2}^2\,dp &\le 2\|f_j\|_2^2\int_{s(1/2)}^{s(1)} (\|f_j\|_1/\|f_j\|_2)^s\frac{ds}{s'(p)}\\ &\le \|f_j\|_2^2\int_0^{6/5} (\|f_j\|_1/\|f_j\|_2)^s\,ds\\ &= \|f_j\|_2^2\frac{1-(\|f_j\|_1/\|f_j\|_2)^{6/5}}{\log(\|f_j\|_2/\|f_j\|_1)}.\end{aligned}$$

Now (14) implies

$$\mathrm{var}(f) \le 2\sum_{j\in J}\|f_j\|_2^2\frac{1-(\|f_j\|_1/\|f_j\|_2)^{6/5}}{\log(\|f_j\|_2/\|f_j\|_1)}, \tag{15}$$

from which (2) follows. □

REMARK. It is worth noting that the tail estimate (1) can be derived from the variance inequality (2). Indeed, let $f = \mathrm{dist}_\omega(0, v)$ and $u \in (0, 3/4)$. Set $s = s(u) := \inf\{t \ge 0 : \mathbf{P}[f(\omega) > t] < u\}$, and define $g(\omega) := \max\{f(\omega), s\}$. It follows from (2) that

$$\mathrm{var}(g) \le Cu|v|/\Big(1 + \min_e \log(\|\rho_e g\|_2/\|\rho_e g\|_1)\Big)$$

and, hence, $\operatorname{var}(g) \le Cu|v|/\log(1/u)$. Therefore, there is a constant C' such that

$$\mathbf{P}\big[f(x) > s(u) + C'\sqrt{|v|/\log(1/u)}\,\big] < u/2.$$

That is, $s(u/2) \le s(u) + C'\sqrt{|v|/\log(1/u)}$. Induction gives for $n = 1, 2, \ldots,$

$$s(2^{-n}) \le s(1/2) + O(1)\sqrt{n|v|},$$

which is the "upper tail" bound from (1). The lower tail is treated similarly. (This proof is fairly general. Using the more specific arguments from Section 3, a slight improvement for the tail estimates in certain ranges may be obtained.)

Acknowledgment. We are most grateful to Elchanan Mossel for very useful discussions.

REFERENCES

[1] BAIK, J., DEIFT, P. and JOHANSSON, K. (1999). On the distribution of the length of the longest increasing subsequence of random permutations. *J. Amer. Math. Soc.* **12** 1119–1178.

[2] BECKNER, W. (1975). Inequalities in Fourier analysis. *Ann. of Math.* **102** 159–182.

[3] BOBKOV, S. G. and HOUDRE, C. (1999). A converse Gaussian Poincare-type inequality for convex functions. *Statist. Probab. Lett.* **44** 281–290.

[4] BONAMI, A. (1970). Etude des coefficients Fourier des fonctiones de $L^p(G)$. *Ann. Inst. Fourier (Grenoble)* **20** 335–402.

[5] BOURGAIN, J., KAHN, J., KALAI, G., KATZNELSON, Y. and LINIAL, N. (1992). The influence of variables in product spaces. *Israel J. Math.* **77** 55–64.

[6] DEUSCHEL, J.-D. and ZEITOUNI, O. (1999). On increasing subsequences of I.I.D. samples. *Combin. Probab. Comput.* **8** 247–263.

[7] DURRETT, R. (1999). Perplexing problems in probability. *Progr. Probab.* **44** 1–33.

[8] FRIEDGUT, E. (200**X**). Influences in product spaces, KKL and BKKKL revisited. Preprint.

[9] JOHANSSON, K. (2000). Transversal fluctuations for increasing subsequences on the plane. *Probab. Theory Related Fields* **116** 445–456.

[10] KAHN, J., KALAI, G. and LINIAL, N. (1988). The influence of variables on Boolean functions. In *Proceedings of the 29th Annual Symposium on Foundations of Computer Science* 68–80. IEEE Computer Science Press, Washington, DC.

[11] KESTEN, H. (1986). Aspects of first passage percolation. *École d'Été de Probabilités de Saint-Flour, XIV. Lecture Notes in Math.* **1180** 125–264. Springer, Berlin.

[12] KESTEN, H. (1993). On the speed of convergence in first passage percolation. *Ann. Appl. Probab.* **3** 296–338.

[13] LEDOUX, M. (2001). *The Concentration of Measure Phenomenon*. Amer. Math. Soc., Alexandria, VA.

[14] NEWMAN, C. M. and PIZA, M. (1995). Divergence of shape fluctuations in two dimensions. *Ann. Probab.* **23** 977–1005.

[15] PEMANTLE, R. and PERES, Y. (1994). Planar first-passage percolation times are not tight. In *Probability and Phase Transition* (G. Grimmett, ed.) 261–264. Kluwer, Dordrecht.

[16] TALAGRAND, M. (1994). On Russo's approximate zero–one law. *Ann. Probab.* **22** 1576–1587.
[17] TALAGRAND, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Inst. Hautes Études Sci. Publ. Math.* **81** 73–205.

I. BENJAMINI
WEIZMANN INSTITUTE OF SCIENCE
REHOVOT 76100
ISRAEL
E-MAIL: itai@wisdom.weizmann.ac.il
URL: www.wisdom.weizmann.ac.il/~itai/

G. KALAI
HEBREW UNIVERSITY
GIVAT RAM
JERUSALEM 91904
ISRAEL
E-MAIL: kalai@math.huji.ac.il
URL: www.ma.huji.ac.il/~kalai/

O. SCHRAMM
MICROSOFT RESEARCH
ONE MICROSOFT WAY
REDMOND, WASHINGTON 98052
E-MAIL: schramm@microsoft.com
URL: research.microsoft.com/~schramm/

Part V
Schramm-Loewner Evolution

ISRAEL JOURNAL OF MATHEMATICS 118 (2000), 221–288

SCALING LIMITS OF LOOP-ERASED RANDOM WALKS AND UNIFORM SPANNING TREES

BY

ODED SCHRAMM*

Department of Mathematics, The Weizmann Institute of Science, Rehovot 76100, Israel
e-mail: schramm@wisdom.weizmann.ac.il
http://www.wisdom.weizmann.ac.il/~schramm/

ABSTRACT

The uniform spanning tree (UST) and the loop-erased random walk (LERW) are strongly related probabilistic processes. We consider the limits of these models on a fine grid in the plane, as the mesh goes to zero. Although the existence of scaling limits is still unproven, subsequential scaling limits can be defined in various ways, and do exist. We establish some basic a.s. properties of these subsequential scaling limits in the plane. It is proved that any LERW subsequential scaling limit is a simple path, and that the trunk of any UST subsequential scaling limit is a topological tree, which is dense in the plane.

The scaling limits of these processes are conjectured to be conformally invariant in dimension 2. We make a precise statement of the conformal invariance conjecture for the LERW, and show that this conjecture implies an explicit construction of the scaling limit, as follows. Consider the Löwner differential equation

$$\frac{\partial f}{\partial t} = z\frac{\zeta(t)+z}{\zeta(t)-z}\frac{\partial f}{\partial z},$$

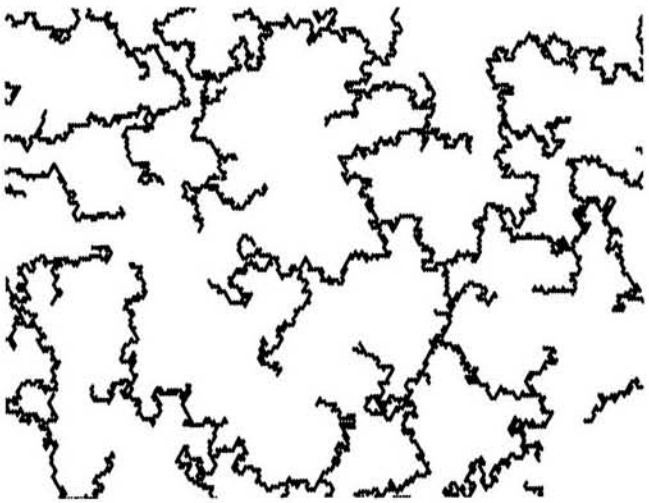
with boundary values $f(z,0)=z$, in the range $z \in \mathbb{U} = \{w \in \mathbb{C}: |w|<1\}$, $t \leqslant 0$. We choose $\zeta(t) := \mathsf{B}(-2t)$, where $\mathsf{B}(t)$ is Brownian motion on $\partial\mathbb{U}$ starting at a random-uniform point in $\partial\mathbb{U}$. Assuming the conformal invariance of the LERW scaling limit in the plane, we prove that the scaling limit of LERW from 0 to $\partial\mathbb{U}$ has the same law as that of the path $f(\zeta(t),t)$ (where $f(z,t)$ is extended continuously to $\partial\mathbb{U}\times(-\infty,0]$). We believe that a variation of this process gives the scaling limit of the boundary of macroscopic critical percolation clusters.

* Research supported by the Sam and Ayala Zacks Professorial Chair.
Received May 24, 1999 and in revised form August 1, 1999

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_27, © Springer Science+Business Media, LLC 2011

TABLE OF CONTENTS

1. Introduction 223
2. Some backround and terminology 234
3. No loops 239
4. First steps in the proof of Theorem 1.3 245
5. Getting uniform convergence 246
6. Recognizing the Löwner parameter as Brownian motion 250
7. The winding number of SLE 252
8. The twisting constant of LERW 256
9. The critical value for the SLE 264
10. Properties of UST subsequential scaling limits in two dimensions265
11. Free and wired trunks and conformal invariance 280
12. Speculations about the Peano curve scaling limit 284
References 286



An approximation of the UST scaling limit trunk.

1. Introduction

GENERAL REMARKS ABOUT SCALING LIMITS. It is often the case that grid-based probabilistic models should be considered as a mere substitute, or simplification, of a continuous process. There are definite advantages for working in the discrete setting, where unpleasant technicalities can frequently be avoided, simulations are possible, and the setup is easier to comprehend. On the other hand, one is often required to pay some price for the simplification. When we adopt the grid-based world, we sacrifice rotational or conformal symmetries which the continuous model may enjoy, and often have to accept some arbitrariness in the formulation of the model. There are also numerous examples where the continuous process is easier to analyze than the discrete process, and in such situations the continuous may be a useful simplification of the discrete.

Understanding the connections between grid-based models and continuous processes is a project of fundamental importance, and so far has only limited success. As mathematicians, we should not content ourselves with the vague notion that the discrete and continuous models behave "essentially the same", but strive to make the relations concrete and precise.

One reasonable way to define a continuous process is by taking a **scaling limit** of a grid process. This means making sense of the limit of a sequence of grid processes on finer and finer grids. Recently, Aizenman [Aiz] has proposed a definition for the scaling limit of percolation, and we shall propose a somewhat different definition [Sch].

Although, in general, the understanding of the connections between grid-based models and continuous models is lacking, there have been some successes. The classical and archetypical example is the relation between simple random walk (SRW) and Brownian motion, which is well studied and quite well understood. See, for example, the discussion of Donsker's Theorem in [Dur91]. We also mention that recently, Tóth and Werner [TW98] have described the scaling limit of a certain self-repelling walk on $\mathbb{Z}$.

The present paper deals with the scaling limits of two very closely related processes, the loop-erased random walk (LERW) and the uniform spanning tree (UST). While these processes are interesting also in dimensions 3 and higher, we restrict attention to two dimensions. In the plane, the scaling limits are conjectured to possess conformal invariance (precise statements appear below), and this can serve as one justification for the special interest in dimension 2. The recent preprint by Aizenman, Burchard, Newman and Wilson [ABNW] discusses scaling limits of random tree processes in two dimensions, including the UST.

The present work answers some of the questions left open in [ABNW].

The most fundamental task in studying a scaling limit process is to set up a conceptual foundation for the scaling limit. This means answering the following two questions: what kind of object is the scaling limit, and what does it mean to be the scaling limit? For the first question, there's often more than one "right" answer. For example, the scaling limit of (two-sided) simple random walk in $\mathbb{Z}$ is usually defined as a probability measure on $C(\mathbb{R})$, the space of continuous, real-valued functions on $\mathbb{R}$, but it could also be defined as a measure on the space of closed subsets of $\mathbb{R} \times \mathbb{R}$, with an appropriately chosen metric. There's also a "wrong" answer here. It is not a good idea to consider the scaling limit of SRW as a probability measure on $\mathbb{R}^{\mathbb{R}}$, though this might seem at first as more natural.

After the conceptual framework is fixed, the next natural question is the existence of the scaling limit. Unfortunately, we cannot report on any progress here. There is every reason to believe that UST and LERW have scaling limits, but a proof is still lacking. However, in the setup we propose below, the existence of subsequential scaling limits is almost a triviality: for every sequence of positive δ_j tending to 0, there is a subsequence δ_{j_n} such that the UST and the LERW on the grids $\delta_{j_n}\mathbb{Z}^2$ do converge to a limit as $n \to \infty$. Such a limit is called a subsequential scaling limit of the model, and is a probability measure on some space.

All the above discussion concerns foundational issues, which are important. But it is not less important to prove properties of the (subsequential) scaling limit.[1] In this paper, we prove several almost sure properties of the UST and LERW subsequential scaling limit. We now describe these models and explain the results.

THE LERW MODEL AND ITS SCALING LIMIT. Consider some set of vertices $K \neq \emptyset$, in a recurrent graph G, and a vertex v_0. The LERW from v_0 to K in G is obtained by running simple random walk (SRW) from v_0, erasing loops as they are created, and stopping when K is hit. Here's a more precise description. Let RW be simple random walk starting at $\mathsf{RW}(0) = v_0$ and stopped at the first time τ such that $\mathsf{RW}(\tau) \in K$. Its loop-erasure, $\mathsf{LE} = \mathsf{LE}_{\mathsf{RW}}$, is defined inductively as follows: $\mathsf{LE}(0) := v_0$, and $\mathsf{LE}(j+1) = \mathsf{RW}(t+1)$ if t is the last time less than τ such that $\mathsf{RW}(t) = \mathsf{LE}(j)$. The walk LE stops when it gets to $\mathsf{RW}(\tau) \in K$. Note

1 Sometimes, this can be done without even defining the scaling limit. For example, the Russo–Seymour–Welsh Theorem [Rus78], [SW78] in percolation theory implies properties of any reasonably defined percolation scaling limit. Similarly, Benjamini's paper [Ben] does not explicitly discuss the scaling limit of UST, but has implication to the UST scaling limit.

that this LERW is a random simple path from v_0 to K.

On a transient graph, it may happen that RW does not hit K. However, one can discuss the loop-erasure of the walk continued indefinitely, since it a.s. visits any vertex only finitely many times.

LERW on $\mathbb{Z}^d$ was studied extensively by Greg Lawler (see the survey paper [Law] and the references therein), who considered LERW as a simpler substitute for the self-avoiding random walk (see the survey [Sla94]), which is harder to analyze. However, we believe that the LERW model is just as interesting mathematically, because of its strong ties with SRW and UST.

For compactness's sake, in the following we consider the plane $\mathbb{C} = \mathbb{R}^2$ as a subset of the two-sphere, $\mathbb{S}^2 = \mathbb{C} \cup \{\infty\}$, which is the one point compactification of the plane, and work with the spherical metric d_{sp} on $\mathbb{S}^2$. Let D be a domain (nonempty open connected set) in the plane $\mathbb{C} = \mathbb{R}^2$. We consider a graph $G = G(D, \delta)$, which is an approximation of the domain D in the square grid $\delta\mathbb{Z}^2$ of mesh δ. The **interior vertices**, $\mathsf{V}_I(G)$, of G are the vertices of $\delta\mathbb{Z}^2$ which are in D, and the **boundary vertices**, $\mathsf{V}_\partial(G)$, are the intersections of edges of $\delta\mathbb{Z}^2$ with ∂D. (The precise definition of G appears in Section 2.) Suppose that each component of ∂D has positive diameter. Let $a \in D$, and let $\mathsf{LE} = \mathsf{LE}_{a,D,\delta}$ be LERW from a vertex $a' \in \delta\mathbb{Z}^2 \cap D$ closest to a to $\mathsf{V}_\partial(G)$ in G.

To make sense of the concept of the **scaling limit** of LERW in D, we think of LE as a random set in $\overline{D}$. Recall that the Hausdorff distance $d_{\mathcal{H}}(X, Y)$ between two closed nonempty sets X and Y in a compact metric space Z is the least $t \geqslant 0$ such that each point $x \in X$ is within distance t from Y and each point $y \in Y$ is within distance t from X; that is,

$$d_{\mathcal{H}}(X, Y) = \inf\left\{t \geqslant 0 \colon X \subset \bigcup_{x \in X} B(x, t),\ Y \subset \bigcup_{y \in Y} B(y, t)\right\}.$$

On the collection $\mathcal{H}(D)$ of closed subsets of D, we use the metric $d_{\mathcal{H}(D)}(X, Y) := d_{\mathcal{H}}(X \cup \partial D, Y \cup \partial D)$, and $\mathcal{H}(D)$ is compact with this metric. Then $\mathsf{LE} \cap D$ is a random element in $\mathcal{H}(D)$, and its distribution $\mu_\delta = \mu_{\delta,D}$ is a probability measure on $\mathcal{H}(D)$. Because the space of Borel probability measures on a compact space is compact in the weak topology,[2] there is a sequence $\delta_j \to 0$ such that the weak limit $\mu_0 := \lim_{j \to \infty} \mu_{\delta_j}$ exists. Such a measure μ_0 will be called a **subsequential scaling limit** measure of LERW from a to ∂D. If $\mu_0 = \lim_{\delta \to 0} \mu_\delta$, then we say that μ_0 is the **scaling limit** measure of LERW from a to ∂D.

Similarly, we may consider the scaling limit of LERW between two distinct

2 We review the notion of weak convergence in Section 2.

points $a, b \in \mathbb{S}^2$, as follows. For $\delta > 0$, we take LE to be the loop-erasure of SRW on $\delta\mathbb{Z}^2$ starting from a vertex of $\delta\mathbb{Z}^2$ within distance 2δ of a and stopped when it first hits a vertex within distance 2δ of b. Since LE is a.s. compact, its distribution is an element of the Hausdorff space $\mathcal{H}(\mathbb{S}^2)$, and there exists a Borel probability measure on $\mathcal{H}(\mathbb{S}^2)$, which is a subsequential scaling limit measure of the law of LE.

THEOREM 1.1: *Let D be a domain in $\mathbb{S}^2$ such that each connected component of ∂D has positive diameter, and let $a \in D$. Then every subsequential scaling limit measure of LERW from a to ∂D is supported on simple paths.*

Similarly, if a, b are distinct points in $\mathbb{S}^2$, then every subsequential scaling limit of the LERW from a to b in $\delta\mathbb{Z}^2$ is supported on simple paths.

Saying that the measure is supported on simple paths means that there's a collection of simple paths whose complement has zero measure.

THE CONFORMAL INVARIANCE CONJECTURE FOR LERW. Consider two domains $D, D' \subset \mathbb{S}^2$. Every homeomorphism $f\colon D \to D'$ induces a homeomorphism $\mathcal{H}(D) \mapsto \mathcal{H}(D')$. Consequently, if μ is a probability measure on $\mathcal{H}(D)$, there is an induced probability measure $f_*\mu$ on $\mathcal{H}(D')$.

CONJECTURE 1.2: *Let $D \subsetneq \mathbb{C}$ be a simply connected domain in $\mathbb{C}$, and let $a \in D$. Then the scaling limit of LERW from a to ∂D exists. Moreover, suppose that $f\colon D \to D'$ is a conformal homeomorphism onto a domain $D' \subset \mathbb{C}$. Then $f_*\mu_{a,D} = \mu_{f(a),D'}$, where $\mu_{a,D}$ is the scaling limit measure of LERW from a to ∂D, and $\mu_{f(a),D'}$ is the scaling limit measure of LERW from $f(a)$ to $\partial D'$.*

Although conformal invariance conjectures have been "floating in the air" in the physics literature for quite some time now, we believe that this precise statement has not yet appeared explicitly. Support for this conjecture comes from simulations which we have performed, and from the work of Rick Kenyon [Ken98a], [Ken98b], [Ken99].

We prove that Conjecture 1.2 implies an explicit description of the LERW scaling limit in terms of solutions of Löwner's differential equation with a Brownian motion parameter. We now give a brief explanation of this.

Let $\mathbb{U} := \{z \in \mathbb{C}\colon |z| < 1\}$, the unit disk. If γ is a compact simple path in $\overline{U} - \{0\}$, such that $\gamma \cap \partial\mathbb{U}$ is an endpoint of γ, then there is a unique conformal homeomorphism $f_\gamma\colon \mathbb{U} \to \mathbb{U} - \gamma$ such that $f_\gamma(0) = 0$ and $f'_\gamma(0) > 0$ (that is, $f'_\gamma(0)$ is real and positive). Moreover, if $\widehat{\gamma}$ is another such path, and $\widehat{\gamma} \supset \gamma$, then $f'_\gamma(0) > f'_{\widehat{\gamma}}(0)$. Now suppose that β is a compact simple path in $\overline{\mathbb{U}}$ such that $\partial\mathbb{U} \cap \beta$ is an endpoint of β and 0 is the other endpoint of β, as in Figure 1.1.

For each point $q \in \beta - \{0\}$, let β_q be the arc of β extending from q to $\partial\mathbb{U}$, and let $h(q) := \log f'_{\beta_q}(0)$. (If q is the endpoint of β on $\partial\mathbb{U}$, then $f_{\beta_q}(z) = z$.) It turns out that h is a homeomorphism from $\beta - \{0\}$ onto $(-\infty, 0]$. We let $q(t)$ denote the inverse map q: $(-\infty, 0] \to \beta$, and set $f(z,t) = f_t(z) := f_{\beta_{q(t)}}(z)$. In this setting, Löwner's Slit Mapping Theorem [Löw23] (see also [Pom66]) states that $f_t(z)$ is the solution of Löwner's equation

$$\frac{\partial}{\partial t} f_t(z) = z f'_t(z) \frac{\zeta(t) + z}{\zeta(t) - z}, \quad \forall z \in \mathbb{U}, \forall t \in (-\infty, 0], \tag{1.1}$$

where ζ: $(-\infty, 0] \to \partial\mathbb{U}$ is some continuous function. In fact, $\zeta(t)$ is defined by the equation

$$f\big(\zeta(t), t\big) = q(t).$$

(The left hand side makes sense, since there is a unique continuous extension of f_t to $\partial\mathbb{U}$.) Note that f also satisfies

$$f(z, 0) = z, \quad \forall z \in \mathbb{U}. \tag{1.2}$$

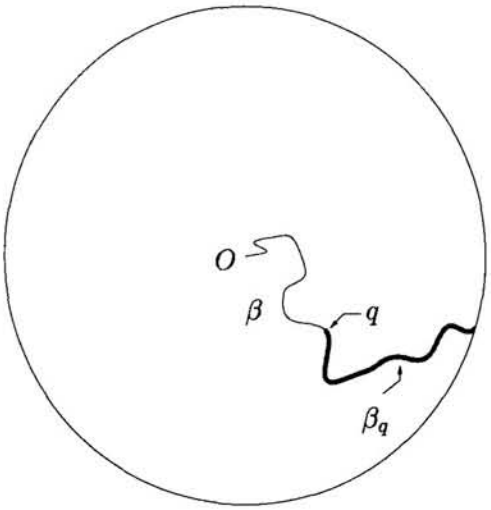


Figure 1.1.

THEOREM 1.3 (The differential equation for the LERW scaling limit): *Assume Conjecture 1.2. Let* $\mathsf{B}(t)$, $t \geqslant 0$ *be Brownian motion on* $\partial\mathbb{U}$ *starting from a uniform-random point on* $\partial\mathbb{U}$. *Let* $f(z,t)$ *be the solution of* (1.1) *and* (1.2), *with* $\zeta(t) := \mathsf{B}(-2t)$. *Set*

$$\sigma(t) = f(\zeta(t), t), \quad t \leqslant 0. \tag{1.3}$$

Then $\{0\} \cup \sigma\big((-\infty, 0)\big)$ *has the same distribution as the scaling limit of LERW from* 0 *to* $\partial\mathbb{U}$.

The Brownian motion $\mathsf{B}(t)$ in the theorem can be defined as $\mathsf{B}(t) := \exp(i\widehat{\mathsf{B}}(t))$, where $\widehat{\mathsf{B}}(t)$ is ordinary Brownian motion on $\mathbb{R}$, starting at a uniform-random point in $[0, 2\pi)$.

Remark 1.4: It is not **a priori** clear that $\sigma(t)$ is well defined and that $\sigma(t)$ is a simple path. For example, we do not know how to prove that f_t extends continuously to $\overline{\mathbb{U}}$ without assuming Conjecture 1.2. The proof of Theorem 1.3 starts by considering the scaling limit of LERW, which is a simple path by Theorem 1.1. It is then shown that the corresponding ζ has the same distribution as $\mathsf{B}(-2t)$.

Remark 1.5: Although (1.1) may look like a PDE, it can in fact be presented as an ODE. Set $\Phi(z, t, s) := f_s^{-1}(f_t(z))$ when $t \leqslant s \leqslant 0$. Then

$$\begin{aligned} \Phi(z, t, 0) &= f_t(z), \\ \Phi(z, t, t) &= z. \end{aligned} \tag{1.4}$$

It is immediate to see that f_t satisfies (1.1) iff $\Phi = \Phi(z, t, s)$ satisfies

$$\frac{\partial \Phi}{\partial s} = -\Phi \frac{\zeta(s) + \Phi}{\zeta(s) - \Phi}. \tag{1.5}$$

Therefore, $f_t(z)$ can be obtained by solving the ODE (1.5) with t fixed and $s \in [t, 0]$.

Note that $(\zeta + \Phi)/(\zeta - \Phi)$ has positive real part when $\Phi \in \mathbb{U}$ and $\zeta \in \partial\mathbb{U}$. Therefore, (1.5) implies that $|\Phi(z, t, s)|$ is monotone decreasing as a function of s. From this it can be deduced that there is a unique solution to the system (1.4) and (1.5) in the interval $s \in [t, 0]$.

Obviously, Theorem 1.3 together with Conjecture 1.2 describe the LERW scaling limit in any simply connected domain $D \subsetneq \mathbb{C}$, since such domains are conformally equivalent to $\mathbb{U}$.

It would be interesting to extract properties of the LERW scaling limit from Theorem 1.3.

At the heart of the proof of Theorem 1.3 lies the following simple combinatorial fact about LERW. Conditioned on a subarc β' of the LERW β from 0 to ∂D, which extends from some point $q \in \beta$ to ∂D, the distribution of $\beta - \beta'$ is the same as that of LERW from 0 to $\partial(D - \beta')$, conditioned to hit q. (See Lemma 4.3.) When we take the scaling limit of this property, and apply the conformal map

from $\partial(D-\beta')$ to $\mathbb{U}$, this translates into the Markov property and stationarity of the associated Löwner parameter ζ.

THE UNIFORM SPANNING TREE AND ITS SCALING LIMIT. Shortly, a definition of the uniform spanning tree (UST) on $\mathbb{Z}^2$ will be given. The UST is a statistical-physics model. It lies in the boundary of the two-parameter family of random-cluster measures, which includes Bernoulli percolation and the Ising model [Häg95]. The UST is very interesting mathematically, partly because it is closely related to the theory of resistor networks, potential theory, random walks, LERW, and in dimension 2, also domino tilings. The paper [BLPS98] gives a comprehensive study of uniform spanning trees (and forests), following earlier pioneering work [Ald90], [Bro89], [Pem91], [BP93], [Häg95]. A survey of current UST theory can be found in [Lyo98].

Let G be a connected graph. A **forest** is a subgraph of G that has no cycles. A **tree** is a connected forest. A subgraph of G is **spanning** if it contains $\mathsf{V}(G)$, the set of vertices of G. We will be concerned with spanning trees. Since a spanning tree is determined by its edges, we often don't make a distinction between the spanning tree T and its set of edges $\mathsf{E}(T)$.

If G is finite, a uniform spanning tree (UST) in G is a random spanning tree $\mathsf{T} \subset G$, selected according to the uniform measure. (That is, $\mathbf{P}[\mathsf{T} = T_1] = \mathbf{P}[\mathsf{T} = T_2]$, whenever T_1 and T_2 are spanning trees of G.)

It turns out that UST's are very closely related to LERW's. If $a, b \in \mathsf{V}(G)$, then the (unique) path in the UST joining a and b has the same law as the LERW from a to b in G. (This, in particular, implies that the LERW from a to b has the same law as the LERW from b to a.) Wilson's algorithm [Wil96], which will be described in Section 2, is a very useful method to build the UST by running LERW's.

R. Lyons proposed (see [Pem91]) to extend the notion of UST to infinite graphs. Let G be an infinite connected graph. Consider a nested sequence of connected finite sugraphs $G_1 \subset G_2 \subset \cdots \subset G$ such that $G = \bigcup_j G_j$. For each j, the uniform spanning tree measure μ_j on G_j may be considered as a measure on $2^{\mathsf{E}(G)}$, the σ-field of subsets of the edges of G, generated by the sets of the form $\{F \subset \mathsf{E}(G)\colon e \in F\}$, $e \in \mathsf{E}(G)$. Using monotonicity properties, it can be shown that the weak limit $\mu := \lim_{j\to\infty} \mu_j$ exists, and does not depend on the sequence $\{G_j\}$. It is called the free uniform spanning forest measure (FSF) on G. (The reason for the word 'free' is that there's another natural kind of limit, the wired uniform spanning forest (WSF). On $\mathbb{Z}^d$, these two measures agree.) R. Pemantle [Pem91] proved that if $d \leqslant 4$, the FSF measure on $\mathbb{Z}^d$ is supported on spanning

trees (that is, if T is random and its law is the FSF measure, then T is a.s. a spanning tree), while if $d \geqslant 5$, the measure is supported on disconnected spanning forests.

Let us now restrict attention to the case $G = \mathbb{Z}^2$. Since it is supported on spanning trees, we call the FSF measure on $\mathbb{Z}^2$ the **uniform spanning tree** (UST) on $\mathbb{Z}^2$. I. Benjamini [Ben] and R. Kenyon [Ken99] studied asymptotic properties of the UST on a rescaled grid $\delta\mathbb{Z}^2$, with δ small, but did not attempt to define the scaling limit. Aizenman, Burchard, Newman and Wilson [ABNW] defined the scaling limit of UST (and other tree processes) in $\mathbb{Z}^2$, and studied some of their properties.

We present a different definition for the scaling limit of the UST. Let $\delta > 0$. Again, we think of $\delta\mathbb{Z}^2$ as a subset of the sphere $\mathbb{S}^2 = \mathbb{R}^2 \cup \{\infty\}$. Let $\widehat{\mathsf{T}}_\delta$ be the UST on $\delta\mathbb{Z}^2$, union with the point at infinity. Then $\widehat{\mathsf{T}}_\delta$ can be thought of as a random compact subset of $\mathbb{S}^2$. However, it is fruitless to consider the weak limit as $\delta \to 0$ of the law of $\widehat{\mathsf{T}}_\delta$ as a measure on the Hausdorff space $\mathcal{H}(\mathbb{S}^2)$, since the limit measure is an atomic measure supported on the single point in $\mathcal{H}(\mathbb{S}^2)$, which is all of $\mathbb{S}^2$.

Given two points $a, b \in \widehat{\mathsf{T}}_\delta$, let $\omega_{a,b}$ be the unique path in $\widehat{\mathsf{T}}_\delta$ with endpoints a and b, allowing for the possibility $\omega_{a,b} = \{a\}$, when $a = b$. (It was proved by R. Pemantle that a.s. the UST T in $\mathbb{Z}^2$ has a single end; that is, there is a unique infinite ray in T starting at 0. This implies that indeed $\omega_{a,b}$ exists and is unique, not only for T, but also for $\mathsf{T} \cup \{\infty\}$.) Let $\mathfrak{T}_\delta = \mathfrak{T}_\delta(\widehat{\mathsf{T}}_\delta)$ be the collection of all triplets, $(a, b, \omega_{a,b})$, where $a, b \in \widehat{\mathsf{T}}_\delta$. Then $\mathfrak{T}_\delta$ is a closed subset of $\mathbb{S}^2 \times \mathbb{S}^2 \times \mathcal{H}(\mathbb{S}^2)$, and the law μ_δ of $\mathfrak{T}_\delta$ is a probability measure on the compact space $\mathcal{H}\big(\mathbb{S}^2 \times \mathbb{S}^2 \times \mathcal{H}(\mathbb{S}^2)\big)$. By compactness, there is a subsequential weak limit μ of μ_δ as $\delta \to 0$, which is a probability measure on $\mathcal{H}\big(\mathbb{S}^2 \times \mathbb{S}^2 \times \mathcal{H}(\mathbb{S}^2)\big)$. We call μ a **subsequential UST scaling limit in** $\mathbb{Z}^2$. If $\mu := \lim_{\delta \to 0} \mu_\delta$, as a weak limit, then μ is the **UST scaling limit**.

We prove

THEOREM 1.6: *Let μ be a subsequential UST scaling limit in $\mathbb{Z}^2$, and let $\mathfrak{T} \in \mathcal{H}\big(\mathbb{S}^2 \times \mathbb{S}^2 \times \mathcal{H}(\mathbb{S}^2)\big)$ be a random variable with law μ. Then the following holds a.s.*

(i) *For every $(a, b) \in \mathbb{S}^2 \times \mathbb{S}^2$, there is some $\omega \in \mathcal{H}(\mathbb{S}^2)$ such that $(a, b, \omega) \in \mathfrak{T}$. For almost every $(a, b) \in \mathbb{S}^2 \times \mathbb{S}^2$, this ω is unique.*

(ii) *For every $(a, b, \omega) \in \mathfrak{T}$, if $a \neq b$, then ω is a simple path, that is, homeomorphic to $[0, 1]$. If $a = b$, then ω is a single point or homeomorphic to a circle. For almost every $a \in \mathbb{S}^2$, the only ω such that $(a, a, \omega) \in \mathfrak{T}$ is $\{a\}$.*

(iii) *The* **trunk**,

$$\mathsf{trunk} := \bigcup_{(a,b,\omega)\in\mathfrak{T}} (\omega - \{a,b\}),$$

is a topological tree (in the sense of Definition 10.1), which is dense in $\mathbb{S}^2$.

(iv) *For each* $x \in \mathsf{trunk}$, *there are at most three connected components of* $\mathsf{trunk} - \{x\}$.

This theorem basically answers all the topological questions about the UST scaling limit on $\mathbb{Z}^2$. It is sharp, in the sense that all the "almost every" clauses cannot be replaced by "every". Benjamini [Ben] proved a result which is closely related to item (iv) of the theorem, and [ABNW] proved (in a different language) that (iv) holds with "three" replaced by some unspecified constant.

The **dual** of a spanning tree $T \subset \mathbb{Z}^2$ is the spanning subgraph of the dual graph $(1/2,1/2)+\mathbb{Z}^2$ containing all edges that do not intersect edges in T. It turns out that duality is measure preserving from the UST on $\mathbb{Z}^2$ to the UST on the dual grid. The key to the proof of Theorem 1.6 is the statement that the trunk is disjoint from the trunk of the dual UST scaling limit.

Let us stress that Theorem 1.6 is not contingent on Conjecture 1.2. The only contingent theorems proved in this paper are Theorem 1.3, and Theorem 11.3, which says that Conjecture 1.2 implies conformal invariance for the scaling limit of the UST on subdomains of $\mathbb{C}$.

Recent work of R. Kenyon [Ken98a], [Ken98b] proves some conformal invariance results for domino tilings of domains in the plane. There is an explicit correspondence between the UST in $\mathbb{Z}^2$ and domino tilings of a finer grid. Based on this correspondence, some properties of the UST can be proved using Kenyon's machinery. For example, Kenyon has shown [Ken98b] that the expected number of edges in a LERW joining two boundary vertices in $(\delta\mathbb{Z}^2)\cap[0,1]^2$ (whose distance from each other is bounded from below) grows like $\delta^{-5/4}$, as $\delta\to 0$. He can also show [Ken99] that the weak limit as $\delta\to 0$ of the distribution of the UST meeting point of three boundary vertices of $D\cap(\delta\mathbb{Z}^2)$ is equivariant with respect to conformal maps. That can be viewed as a partial conformal invariance result for the UST scaling limit. Another example for the applications of Kenyon's work to the UST appears in Section 8. It seems plausible that perhaps soon there would be a proof of Conjecture 1.2.

SLE WITH OTHER PARAMETERS, CRITICAL PERCOLATION, AND THE UST PEANO CURVE. Let $\kappa \geqslant 0$, and take $\zeta(t) := \mathsf{B}(-\kappa t)$, where B is as above, Brownian motion on $\partial\mathbb{U}$, started from a uniform random point. Then there is a solution $f(z,t)$ of (1.1) and (1.2), and for each $t \leqslant 0$, $f_t = f(\cdot,z)$ is a conformal

map from $\mathbb{U}$ into some subdomain $D_t \subset \mathbb{U}$. We call the process $\sigma_t^\kappa := \mathbb{U} - D_t$, $t \leqslant 0$, the **stochastic Löwner evolution** (SLE) with parameter κ. It is not always the case that σ_t^κ is a simple path. Let $\mathfrak{K}$ be the set of all $\kappa \geqslant 0$ such that for all $t < 0$ the set σ_t^κ is a.s. a simple path. We show in Section 9 that $\sup \mathfrak{K} \leqslant 4$, and conjecture that $\mathfrak{K} = [0, 4]$.

In the past, there has been some work on the question of which Löwner parameters ζ produce slitted disk mappings ([Kuf47], [Pom66]), but only limited progress has been made. Partly motivated by the present work, Marshall and Rohde [MR] have looked into this problem again, and have shown that when ζ satisfies a Hölder condition with exponent $1/2$ and Hölder$(1/2)$ norm less than some constant, the maps f_t are onto slitted disks, and this may fail when ζ has finite but large Hölder$(1/2)$ norm.

Given some $\kappa > 0$, even if $\kappa \notin \mathfrak{K}$, the process σ_t^κ, $t \leqslant 0$, is quite interesting. It is a celebrated conjecture that critical Bernoulli percolation on lattices in $\mathbb{R}^2$ exhibits conformal invariance in the scaling limit [LPSA94]. Assuming such a conjecture, we plan to prove in a subsequent work that a process similar to SLE describes the scaling limit of the outer boundary of the union of all critical percolation clusters in a domain D which intersect a fixed arc on the boundary of D. We also plan to prove that this implies Cardy's [Car92] conjectured formula for the limiting crossing probabilities of critical percolation, and higher order generalizations of this formula.

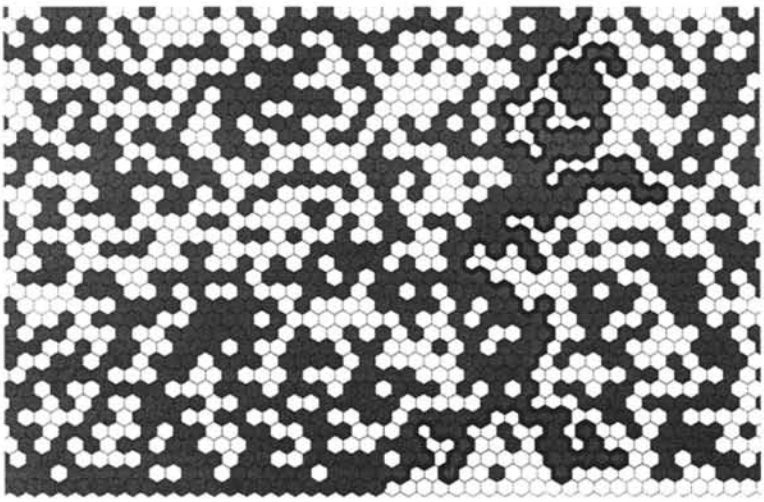


Figure 1.2. The boundary curve for critical percolation with mixed boundary conditions.

Let us now briefly explain this. In Figure 1.2, each of the hexagons is colored black with probability $1/2$, independently, except that the hexagons intersecting the positive real ray are all white, and the hexagons intersecting the negative real ray are all black. Then there is a boundary path β, passing through 0 and separating the black and the white regions adjacent to 0. Note that the percolation in the figure is equivalent to Bernoulli$(1/2)$ percolation on the triangular grid, which is critical. (See [Gri89] for background on percolation.) The intersection of β with the upper half plane, $\mathbb{H} := \{z \in \mathbb{C}: \operatorname{Im} z > 0\}$, which is indicated in the picture, is a random path in $\mathbb{H}$ connecting the boundary points 0 and ∞.

A subsequential scaling limit of $\beta \cap \mathbb{H}$ exists, by compactness, and naturally, we believe that the weak limit exists. Let γ be the scaling limit curve. The physics wisdom (unproven, perhaps not even precisely formulated, but well supported) is that the scaling limit of the "external boundary" of macroscopic critical percolation clusters in two dimensions has dimension $7/4$ and is **not** a simple path [SD87] (see also [ADA]), and we believe that this is true for γ.

In a subsequent paper, we plan to prove (by adapting the proof of Theorem 1.3), under the assumption of a conformal invariance conjecture for the scaling limit of critical percolation, that γ can be described using a Löwner-like differential equation in the upper half plane with Brownian motion parameter, as follows. Consider the differential equation

$$\frac{\partial}{\partial t} f_t(z) = \frac{-2 f_t'(z)}{\zeta(t) - z}, \quad \forall z \in \mathbb{H}, \forall t \in (-\infty, 0], \tag{1.6}$$

where $\zeta(t) = \mathsf{B}(-\kappa t)$, B is Brownian motion on $\mathbb{R}$ starting at $\mathsf{B}(0) = 0$, and $f_0(z) = z$. Then f_t is a conformal mapping from $\mathbb{H}$ onto a subdomain of $\mathbb{H}$, which is normalized by the so-called **hydrodynamic normalization**

$$\lim_{z \to \infty} f_t(z) - z = 0. \tag{1.7}$$

The claim is that for $\kappa = 6$, the image of the path $t \mapsto f_t(\zeta(t))$ has the same distribution as γ. From this, one can derive Cardy's [Car92] conjectured formula for the limiting crossing probabilities of critical percolation, as well as some higher order generalizations. This will be done in subsequent work, but basically depends on the ideas appearing in Section 9 below.

A similar representation applies to the scaling limit of the Peano curve which winds around the UST (this curve was discussed in [DD88] and mentioned in [BLPS98]), but with $\kappa = 8$. Given a domain $D \subset \mathbb{S}^2$, whose boundary is a simple closed path, and given two distinct points $a, b \in \partial D$, there is a naturally

defined (subsequential) scaling limit of the Peano curve of the UST in D, with appropriate boundary conditions, and the scaling limit is an (unparameterized) curve from a to b, whose image covers $\overline{D}$. One can show that Conjecture 1.2 implies a conformal invariance property for the scaling limit. Based on this, it should be possible to adapt the proof of Theorem 1.3 to show that Conjecture 1.2 implies a representation of the form (1.6) for this Peano scaling limit when $D = \mathbb{H}$, $a = 0$, $b = \infty$, and $\kappa = 8$. We give a brief overview of this in Section 12, and hope to give a more thorough treatment in a subsequent paper.

The differential equation (1.6) is very similar to Löwner's equation, and the only essential difference is that a normalization at an interior point for the maps f_t is replaced by the hydrodynamic normalization (1.7) at a boundary point (∞). The interior point normalization is natural for the LERW scaling limit, because the LERW is a path from an interior point to the boundary of the domain. The Peano curve and the boundary of percolation clusters, as discussed above, are paths joining two boundary points, and hence the hydrodynamic normalization is more appropriate for them.

Although Löwner's Slit Mapping Theorem mentioned above applies to domains of the form $\mathbb{U} - \alpha$, where α is a simple path in $\mathbb{U} - \{0\}$ with one endpoint in $\partial\mathbb{U}$, Pommerenke [Pom66] has a generalization, which is valid for some paths α which are not simple paths. It is this generalization (or rather, its version in $\mathbb{H}$ with the hydrodynamic normalization) which will substitute Löwner's Slit Mapping Theorem for the treatment of the percolation boundary or Peano curve scaling limits.

The emerging picture is that different values of κ in the differential equations (1.1) or (1.6) produce paths which are scaling limits of naturally defined processes, and that these paths can be space-filling, or simple paths, or neither, depending on the parameter κ.

ACKNOWLEDGEMENT: I wish to express gratitude to Itai Benjamini, Rick Kenyon and David Wilson for inspiring discussions and helpful information. Lemma 3.1 has been obtained jointly with Itai Benjamini. David Aldous, Mladen Bestvina, Brian Bowditch, Steve Evans, Yakar Kannai, Greg Lawler, Russ Lyons, Yuval Peres, Steffen Rohde, Jeff Steif, Benjamin Weiss and Wendelin Werner have provided very helpful advice.

2. Some background and terminology

This section will introduce some notations which will be used, and discuss some of the necessary background. We begin with a review of uniform spanning trees

and forests. The reader may consult [BLPS98] for a comprehensive treatment of that subject.

THE DOMINATION PRINCIPLE. Suppose that H and H' are two random subsets of some set. We say that H' **stochastically dominates** H if there is a probability measure μ on pairs (A, B) such that A has the same law as H, B has the same law as H', and $\mu\{(A,B)\colon B \supset A\} = 1$. Such a μ is called a **monotone coupling** of H and H'.

Let G be a finite connected graph, and let G_0 be a connected nonempty subgraph. Let M be the set of vertices of G_0 that are incident with some edge in $\mathsf{E}(G) - \mathsf{E}(G_0)$. (We let $\mathsf{E}(G)$ and $\mathsf{V}(G)$ denote the edges and vertices of G, respectively.) Let G_0^W be the graph obtained from G by identifying all the vertices in M to a single vertex, called the **wired vertex**. Then G_0^W is called the **wired graph** associated to the pair (G_0, G). Let T be the UST on G, let T_0^F be the UST on G_0, and let T_0^W be the UST on G_0^W. Then T_0^F is called the **free spanning tree** of the pair (G_0, G), and T_0^W is the **wired spanning tree** of the pair (G_0, G). Sometimes, we call T_0^F [respectively, T_0^W] the UST on G_0 with free [respectively, wired] boundary conditions. The **domination principle** states that $\mathsf{T}_0^F \cap G_0$ stochastically dominates $\mathsf{T} \cap G_0$, and that $\mathsf{T} \cap G_0$, stochastically dominates $\mathsf{T}_0^W \cap G_0$.

Now let G be an infinite connected graph, and let $G_1 \subset G_2 \subset \cdots$ be an infinite sequence of finite connected subgraphs satisfying $\bigcup_j G_j = G$. Let μ_j^W be the law of the wired spanning tree of (G_j, G). Based on the domination principle, it is easy to verify that the weak limit μ^W of μ_j^W exists, and is a probability measure on spanning forests of G. It is called the **wired spanning forest of** G (WSF).

The domination principle, when appropriatly interpreted, carries over to infinite and to disconnected graphs as well. If G is a disconnected graph, we take the FSF [respectively, WSF] on G to be the spanning forest of G whose intersection with every component of G is the FSF [respectively, WSF] of that component, and with the restriction to the different components being independent. The more general formulation of the domination principle states that when G_0 is a subgraph of G, then $\mathsf{T}_0^F \cap G_0$ stochastically dominates $\mathsf{T} \cap G_0$ and $\mathsf{T} \cap G_0$ stochastically dominates $\mathsf{T}_0^W \cap G_0$, where $\mathsf{T}, \mathsf{T}_0^F$ and T_0^W are the FSF on G, G_0 and G_0^W, respectively, and the same statement holds when FSF is replaced by WSF.

On all recurrent connected graphs, the WSF is equal to the FSF, and both are trees. Therefore, on recurrent graphs we shall refer to this measure as the uniform spanning tree (UST).

GRID APPROXIMATIONS OF DOMAINS. Let $D \subset \mathbb{C}$ be a domain, that is, an

open, connected set. Given $\delta > 0$, we define a graph $G = G(D, \delta)$, which is a discrete approximation of the domain D in the grid $\delta\mathbb{Z}^2$, as follows. The **interior vertices** $\mathsf{V}_I(G) = \mathsf{V}_I(D, \delta)$ of G are the vertices of $\delta\mathbb{Z}^2$ which are in D. The **boundary vertices** $\mathsf{V}_\partial(G) = \mathsf{V}_\partial(D, \delta)$ are the points of intersection of the edges of the grid $\delta\mathbb{Z}^2$ with ∂D, the boundary of D. The vertices of G are $\mathsf{V}(G) = \mathsf{V}_\partial(G) \cup \mathsf{V}_I(G)$. If $a, b \in \mathsf{V}(G)$ are distinct, then $[a, b]$ is an edge of G iff there is an edge $e \in \mathsf{E}(\delta\mathbb{Z}^2)$ such that the open segment $\{ta + (1-t)b\colon t \in (0,1)\}$ is contained in $D \cap e$.

We will often be considering random walks on $G(D, \delta)$ starting at 0, when $0 \in D$. It will be useful to denote by V_∂^0 the set of vertices $v \in \mathsf{V}_\partial(D, \delta)$ such that there is a path from 0 in $G(D, \delta)$ whose intersection with V_∂ is v.

The wired graph, $G^W(D, \delta)$, associated with D is $G(D, \delta)$ with all the vertices $\mathsf{V}_\partial(D, \delta)$ collapsed to a single vertex, which we simply denote ∂D.

We shall often not distinguish between a graph and its planar embedding, if it has an obvious planar embedding. For example, the UST on G will also be interpreted as a random set in the plane.

WILSON'S ALGORITHM. Let G be a finite graph. Wilson's algorithm [Wil96] for generating a UST in G proceeds as follows. Let $v_0 \in \mathsf{V}(G)$ be an arbitrary vertex (which we call the root), and set $T_0 := \{v_0\}$. Inductively, assume that a tree $T_j \subset G$ has been constructed. If $\mathsf{V}(T_j) \neq \mathsf{V}(G)$, choose a vertex $v_{j+1} \in \mathsf{V}(G) - \mathsf{V}(T_j)$, let W_{j+1} be LERW from v to T_j in G, and set $T_{j+1} := T_j \cup W_{j+1}$. Otherwise $\mathsf{V}(T_j) = \mathsf{V}(G)$, and the algorithm stops and outputs T_j. It is somewhat surprising, but true, that no matter how the choices of the vertices v_j are made, the output of the algorithm is a tree chosen according to the uniform measure.

If G is infinite, connected and recurrent, Wilson's algorithm also "works". When the subtree T_j generated by the algorithm includes all the vertices in a certain finite set $K \subset \mathsf{V}(G)$, the subtree of T_j spanned by K (that is, the minimal connected subgraph of T_j that contains K) has the same law as the subtree of the UST of G spanned by K. (There is also a version of Wilson's algorithm which is useful for generating the WSF of a transient graph, but we shall not need this.)

HARMONIC MEASURE ESTIMATES. Because of Wilson's algorithm, many questions about the UST can be reduced to questions about simple random walks (SRW's). It is therefore hardly surprising that we often need to obtain a **harmonic measure estimate**, that is, an estimate on the probability that SRW starting from a vertex v will hit a certain set of vertices K_1 before hitting another

set K_0. As a function of v, this probability is harmonic[3] away from $K_0 \cup K_1$.

Almost all the harmonic measure estimates which we will use are entirely elementary, and follow from the following easy fact. Consider an annulus $A = A(p, r, R)$, with center p, inner radius r and outer radius $R > r$. Suppose that δ is sufficiently small so that there is a path in $\delta\mathbb{Z}^2 \cap A$ which separates the boundary components of A. Let $q \in \delta\mathbb{Z}^2 \cap A$ be some vertex such that the distance from q to the boundary ∂A is at least r/c, where $c > 0$ is some constant. Let X be the image of SRW starting from q, which is stopped when it first leaves A. Then the probability that X contains a path separating the boundary components of A is bounded below by some positive function of c. (This can be proved directly using only the Markov property and the invariance of SRW under the automorphisms of $\delta\mathbb{Z}^2$.) One consequence of this fact and the Markov property, which we will often use, is as follows.

LEMMA 2.1: *Suppose that K is a connected subgraph of $\delta\mathbb{Z}^2$ of diameter at least R, and $v \in \delta\mathbb{Z}^2$. Then the probability that SRW starting from v will exit the ball $B(v, R)$ before hitting K is at most $C_0\big(\mathrm{dist}(v, K)/R\big)^{C_1}$, where $C - 0, C_1 > 0$ are absolute constants.*

This lemma holds for the spherical as well as Euclidean metric.

LAPLACIAN RANDOM WALK. Although this will not be needed in the paper, we have to mention another interpretation of LERW. Let G be a finite connected graph, let $K \subset \mathsf{V}(G)$ be a set of vertices, and let $a \in \mathsf{V}(G) - K$. The LERW from a to K can also be inductively constructed, as follows. Suppose that the first n vertices $a = \mathsf{LE}(1), \mathsf{LE}(2), \ldots, \mathsf{LE}(n)$ have been determined and $\mathsf{LE}(n) \notin K$. Let $h_n\colon \mathsf{V}(G) \to [0, 1]$ be the function which is 1 on K, 0 on $\{\mathsf{LE}(1), \ldots, \mathsf{LE}(n)\}$ and harmonic on $\mathsf{V}(G) - \big(K \cup \{\mathsf{LE}(1), \ldots, \mathsf{LE}(n)\}\big)$. Then $\mathsf{LE}(n+1)$ is chosen among the neighbors w of $\mathsf{LE}(n)$, with probability proportional to $h_n(w)$.

This formulation of the LERW may serve as a heuristic for Conjecture 1.2, since discrete harmonic functions are good approximations for continuous harmonic functions, and continuous harmonic functions in 2D have conformal invariance properties. However, this heuristic is quite weak, since near a non-smooth boundary of a domain, the approximation is not good.

WEAK CONVERGENCE OF MEASURES. We now recall several facts and definitions regarding weak convergence. The reader may consult [EK86, Chap. 3] for proofs and further references. Let (X, d) be a compact metric space, and let $\{\mu_j\}$

3 A function h is **harmonic** at v, if $h(v)$ is the average of the value of h on the neighbors of v.

be a sequence of Borel probability measures on X. Let μ be a Borel probability measure on X. Saying that the sequence μ_j converges weakly to μ means that $\lim_j \int f\,d\mu_j = \int f\,d\mu$ for all continuous $f\colon X \to \mathbb{R}$. The Prohorov metric on the space of Borel probability measures on X is defined by

$$\widehat{d}(\mu,\mu') := \inf\{\epsilon > 0\colon \mu(K) \leqslant \mu'(N_\epsilon(K)) + \epsilon \text{ for all closed } K \subset X\},$$

where $N_\epsilon(K) := \bigcup_{x\in K} B(x,\epsilon)$ is the ϵ-neighborhood of K. The space of Borel probability measures on X is compact with respect to the Prohorov metric, and weak convergence is equivalent to convergence in the Prohorov metric.

Let $\mathcal{M}(\mu,\mu')$ be the collection of all Borel measures ν on $X \times X$ such that $\nu(A \times X) = \mu(A)$ and $\nu(X \times A) = \mu'(A)$ for all measurable $A \subset X$. Such a ν is called a **coupling** of μ and μ'. The Prohorov metric satisfies

$$\widehat{d}(\mu,\mu') = \inf_{\nu\in\mathcal{M}(\mu,\mu')} \inf\Big\{\epsilon > 0\colon \nu\{(x,y)\colon d(x,y) \geqslant \epsilon\} \leqslant \epsilon\Big\}. \tag{2.1}$$

In other words, $\widehat{d}(\mu,\mu') < \epsilon$ means that one can find a probability space $(\Omega,\mathbf{P})$ and two X valued random variables $x,y\colon \Omega \to X$, such that $\mathbf{P}[d(x,y) \geqslant \epsilon] \leqslant \epsilon$ and such that x has law μ and y has law μ'. This is obtained by taking an appropriate $\mathbf{P} = \nu \in \mathcal{M}(\mu,\mu')$, $\Omega := X \times X$, and letting x and y be the projections on the first and second factors, respectively.

CONFORMAL MAPS. We review some elementary facts about conformal mappings, as may be found in [Dur83], for example. Let $D \subset \mathbb{C}$ be some domain. A continuous map $f\colon D \to \mathbb{C}$, which is injective and complex-differentiable, is **conformal**. If f is conformal, then $f^{-1}\colon f(D) \to \mathbb{C}$ is also conformal. (Some authors use the word **univalent** instead of conformal.)

Let $D \subsetneq \mathbb{C}$ be simply connected. Then Riemann's Mapping Theorem states that there is a conformal homeomorphism $f = f_D$ from $\mathbb{U}$ onto D. Suppose also that $0 \in D$; then f can be chosen to satisfy the normalizations $f(0) = 0$ and $f'(0) > 0$, which render f unique. In this case, the number $f'(0)$ is called the **conformal radius of** D (with respect to 0). The Schwarz Lemma implies that

$$f'(0) \geqslant \inf\{|z|\colon z \notin f(\mathbb{U})\}, \tag{2.2}$$

while on the other hand, the Koebe 1/4 Theorem gives

$$f'(0) \leqslant 4\inf\{|z|\colon z \notin f(\mathbb{U})\}. \tag{2.3}$$

Hence, up to a factor of 4, the conformal radius can be determined from the in-radius.

If $0 < a < b < \infty$, then the set of all conformal maps $f\colon \mathbb{U} \to \mathbb{C}$ satisfying $f(0) = 0$ and $a \leqslant |f'(0)| \leqslant b$ is compact, in the topology of uniform convergence on compact subsets of $\mathbb{U}$. If $f_j\colon \mathbb{U} \to \mathbb{C}$ are conformal, $f_j(0) = 0$, and $f_j \to f$ locally uniformly, then the image $f(\mathbb{U})$ can be described in terms of the images $D_j := f_j(\mathbb{U})$. Let D be the maximal open connected set containing 0 and contained in $\bigcup_{n=1}^{\infty} \bigcap_{j=n}^{\infty} D_j$. If $D \neq \emptyset$, then $f(\mathbb{U}) = D$; otherwise, $f(\mathbb{U}) = \{0\}$. This is called Carathéodory's Kernel Convergence Theorem.

If $f\colon \mathbb{U} \to D$ is a conformal homeomorphism onto D, and ∂D is a simple closed curve, then f extends continuously to $\partial\mathbb{U}$. The same is true if $D = \mathbb{U} - \beta$, where β is a simple path.

3. No loops

In this section, we prove that any LERW subsequential scaling limit is supported on the set of simple paths. That is, we prove Theorem 1.1.

Let o_1 and o_2 be distinct points in $\mathbb{R}^2$. For each $\delta > 0$, let o_1^* and o_2^* be vertices of $\delta\mathbb{Z}^2$ closest to o_1 and o_2, respectively. Note that in $\delta\mathbb{Z}^2$ the combinatorial distance between two vertices $v, v' \in \delta\mathbb{Z}^2$ is $\delta^{-1}\|v - v'\|_1$. However, all metric notions we use will refer to the Euclidean or spherical distance. In this section, we will mainly use the Euclidean metric.

Let RW be a random walk on $\delta\mathbb{Z}^2$ starting at o_2^* and stopped when o_1^* is reached for the first time. Let $\omega = \omega_\delta$ denote the loop-erasure of RW, and let $\mathbf{P}_\delta$ denote the law of ω_δ. The following lemma will show that the diameter of ω_δ is "tight".

LEMMA 3.1:

$$\mathbf{P}_\delta\big[\operatorname{diam}\omega > s \operatorname{dist}(o_1^*, o_2^*)\big] \leqslant C_0 s^{-C_1},$$

where $C_0, C_1 > 0$ are absolute constants.

The proof is based on Wilson's algorithm and an elementary harmonic measure estimate. A more precise estimate can be obtained by using the discrete Beurling Projection Theorem (see [Kes87] or [Law93]).

Proof: Set $r := \operatorname{dist}(o_1^*, o_2^*)$, $R := s\operatorname{dist}(o_1^*, o_2^*)/4$, and let $z \in \delta\mathbb{Z}^2$ be some vertex such that $\operatorname{dist}(o_1^*, z) \geqslant 10R$. For $a, b \in \delta\mathbb{Z}^2$, let $\omega_{a,b}$ denote the path in the UST of $\delta\mathbb{Z}^2$ joining a and b. Let m be the meeting point of o_1^*, o_2^*, z in the UST, that is, the vertex in $\omega_{o_1^*,o_2^*} \cap \omega_{o_1^*,z} \cap \omega_{o_2^*,z}$. By Wilson's algorithm, the distribution of ω is identical with the distribution of $\omega_{o_1^*,m} \cup \omega_{o_2^*,m}$.

We now estimate $\mathbf{P}\big[\operatorname{diam}(\omega_{o_1^*,m}) > s\operatorname{dist}(o_1^*, o_2^*)\big]$. Using Wilson's algorithm, we may generate $\omega_{o_1^*,m}$ by letting $\omega_{o_2^*,z}$ be LERW from o_2^* to z, and letting $\omega_{o_1^*,m}$ be LERW from o_1^* to $\omega_{o_2^*,z}$. Condition on $\omega_{o_2^*,z}$. By Lemma 2.1, the

probability that SRW starting at o_1^* will exit $B(o_1^*, R)$ before hitting $\omega_{o_2^*,z}$ is at most $C_2(r/R)^{C_3}$, for some constants $C_2, C_3 > 0$. Therefore,

$$\mathbf{P}[\operatorname{diam}\omega_{o_1^*,m} > 2R] \leqslant C_2(r/R)^{C_3}.$$

Moreover, the same estimate holds for $\operatorname{diam}\omega_{o_2^*,m}$. Since $\omega_{o_1^*,o_2^*} = \omega_{o_1^*,m} \cup \omega_{o_2^*,m}$, we get $\mathbf{P}[\operatorname{diam}\omega_{o_1^*,o_2^*} > 4R] \leqslant 2C_2(r/R)^{C_3}$. This completes the proof of the lemma. ∎

Remark 3.2: The proof of the lemma can be easily adapted to show that if $a, b \in \delta\mathbb{Z}^2$ and $K \subset \delta\mathbb{Z}^2$, then the probability that LERW from a to K will intersect $B(b, r)$ is at most $C_4\big(r/\operatorname{dist}(a, b)\big)^{C_5}$, provided $r \geqslant \delta$, where $C_4, C_5 > 0$ are absolute constants.

Definition 3.3: Let $z_0 \in \mathbb{R}^2$, $r, \epsilon > 0$. An (z_0, r, ϵ)-quasi-loop in a path ω is a pair $a, b \in \omega$ with $a, b \in B(z_0, r)$, $\operatorname{dist}(a, b) \leqslant \epsilon$, such that the subarc of ω with endpoints a, b is not contained in $B(z_0, 2r)$. Let $\mathcal{A}(z_0, r, \epsilon)$ denote the set of simple paths in $\mathbb{R}^2$ that have a (z_0, r, ϵ)-quasi-loop.

LEMMMA 3.4: *Let c be the distance from o_1^* to o_2^*, let $r \in (0, c/4)$, $\epsilon > 0$ and $z_0 \in \mathbb{R}^2$. Then $\lim_{\epsilon \to 0} \mathbf{P}_\delta[\mathcal{A}(z_0, r, \epsilon)] = 0$, uniformly in δ.*

Proof: Let $B_1 = B(z_0, r)$ and $B_2 = B(z_0, 2r)$. The distance from B_2 to at least one of the points o_1^*, o_2^* is at least $c/4$. By symmetry, we assume with no loss of generality that $\operatorname{dist}\big(o_2^*, B_2\big) \geqslant c/4$. Let $\epsilon_1 \in (0, c/8)$, and let q be a vertex in $\delta\mathbb{Z}^2$ such that $\operatorname{dist}\big(q, o_2^*\big) \in [\epsilon_1 - \delta, \epsilon_1]$.

Let ω^q be a LERW from q to o_2^* in $\delta\mathbb{Z}^2$. Let RW be an independent simple random walk from o_1^*. Let RW′ be the part of the walk RW until ω^q is first hit. Then, by Wilson's algorithm, ω_δ has the same distribution as the arc connecting o_1^* to o_2^* in $\mathsf{LE}(\mathsf{RW}') \cup \omega^q$.

Let $\mathcal{A}'$ be the event that $\mathsf{LE}(\mathsf{RW}')$ has a (z_0, r, ϵ)-quasi-loop, and let $\mathcal{C}$ be the event that ω^q intersects B_2. Observe that

$$\mathcal{A}(z_0, r, \epsilon) \subset \mathcal{A}' \cup \mathcal{C}. \tag{3.1}$$

Since $\operatorname{dist}\big(q, o_2^*\big) \leqslant \epsilon_1$ and $\operatorname{dist}\big(B_2, o_2^*\big) \geqslant c/4$, from Lemma 3.1 we get an estimate of the form

$$\mathbf{P}[\mathcal{C}] \leqslant C_6(\epsilon_1/c)^{C_7}. \tag{3.2}$$

We now find an upper bound for $\mathbf{P}[\mathcal{A}' - \mathcal{C}]$. Let s_1 be the first time $s \geqslant 0$ such that $\mathsf{RW}(s) \in B_1$. Let t_1 be the first time $t \geqslant s_1$ such that $\mathsf{RW}(t_1) \notin B_2$.

Inductively, define s_j to be the first time $s \geqslant t_{j-1}$ such that $\mathsf{RW}(s) \in B_1$ and t_j to be the first time $t \geqslant s_j$ such that $\mathsf{RW}(t) \notin B_2$. Let τ be the first time $t \geqslant 0$ such that $\mathsf{RW}(t) \in \omega^q$. Finally, for each $s \geqslant 0$ let RW^s be the restriction of RW to the interval $t \in [0, s]$.

For each $j = 1, 2, \ldots$, we consider several events depending on RW^{t_j} and ω^q. Let $\mathcal{Y}_j$ be the event that $\mathsf{LE}(\mathsf{RW}^{t_j})$ has a (z_0, r, ϵ)-quasi-loop. Let $\mathcal{T}_j$ be the event that $\tau \geqslant t_j$. Observe that $\mathsf{LE}(\mathsf{RW}^t) \cap B_1 \subset \mathsf{LE}(\mathsf{RW}^{t_j})$ if $t \in [t_j, s_{j+1})$. It follows that $\mathcal{A}' - \mathcal{C} \subset \mathcal{Y}_{j'}$ if j' is the largest j with $t_j \leqslant \tau$. Therefore,

$$\mathcal{A}' - \mathcal{C} \subset \bigcup_{j=1}^{\infty} (\mathcal{Y}_j \cap \mathcal{T}_j).$$

Since $\mathcal{T}_j \supset \mathcal{T}_{j+1}$ for each j, this implies

$$\mathcal{A}' - \mathcal{C} \subset \mathcal{T}_{m+1} \cup \bigcup_{j=1}^{m} \mathcal{Y}_j \tag{3.3}$$

for every m.

We first estimate $\mathbf{P}[\mathcal{T}_{j+1} \mid \mathsf{RW}^{t_j}, \omega^q]$. Conditioned on any ω^q, the probability that a SRW starting at any vertex outside of B_2 will hit ω^q before hitting B_1 is at least

$$C_8(\epsilon_1)^{C_9},$$

where $C_8 > 0$ depends only on c and r, and $C_9 > 0$ is an absolute constant. This is based on the fact that ω^q is connected, contains o_2^*, and has diameter at least $\epsilon_1 - \delta$. Applying this to the walk RW from time t_j on, we therefore get

$$\mathbf{P}[\mathcal{T}_{j+1} \mid \mathsf{RW}^{t_j}, \omega^q] \leqslant 1 - C_8(\epsilon_1)^{C_9}.$$

By induction, we therefore find that

$$\mathbf{P}[\mathcal{T}_m \mid \omega^q] \leqslant \left(1 - C_8(\epsilon_1)^{C)9}\right)^{m-1}. \tag{3.4}$$

We now estimate $\mathbf{P}[\mathcal{Y}_{j+1} \mid \neg\mathcal{Y}_j, \mathsf{RW}^{t_j}]$. Let Q_j be the set of components of $\mathsf{LE}(\mathsf{RW}^{s_{j+1}}) \cap B_2$ that do not contain $\mathsf{RW}(s_{j+1})$. Observe that for $\mathcal{Y}_{j+1}$ to occur, there must be a $K \in Q_j$ such that the random walk RW comes at some time $t \in [s_{j+1}, t_{j+1}]$ within distance ϵ of $K \cap B_1$ but $\mathsf{RW}(t) \notin K$ for all $t \in [s_{j+1}, t_{j+1}]$. But if $\mathsf{RW}(t)$ is close to K, $t \in [s_{j+1}, t_{j+1}]$, then Lemma 2.1 can be applied, to estimate the probability that RW will not hit K before time t_{j+1}. That is, conditioned on $\mathsf{RW}^{s_{j+1}}$, for each given $K \in Q_j$, the probability that

$\mathsf{RW}([s_{j+1}, t_{j+1}])$ gets to within distance ϵ of K but does not hit K is at most $C_1 0(\epsilon/r)^{C_1 1}$, where $C_1 0, C_1 1 > 0$ are absolute constants. Consequently, we get

$$\mathbf{P}[\mathcal{Y}_{j+1} \mid \neg\mathcal{Y}_j, \mathsf{RW}^{t_j}] \leqslant C_1 0 |Q_j| (\epsilon/r)^{C_1 1}.$$

Observe that $|Q_j|$, the cardinality of Q_j, is at most j. Therefore,

$$\mathbf{P}[\mathcal{Y}_{j+1} \mid \neg\mathcal{Y}_j] \leqslant C_1 0 j (\epsilon/r)^{C_1 1}.$$

This gives

$$\begin{aligned} \mathbf{P}\Big[\bigcup_{j=1}^{m} \mathcal{Y}_j\Big] &\leqslant \sum_{j=1}^{m-1} \mathbf{P}[\mathcal{Y}_{j+1} \cap \neg\mathcal{Y}_j] \leqslant \sum_{j=1}^{m-1} \mathbf{P}[\mathcal{Y}_{j+1} \mid \neg\mathcal{Y}_j] \\ &\leqslant \sum_{j=1}^{m-1} j C_1 0 (\epsilon/r)^{C_1 1} \leqslant C_1 2 m^2 (\epsilon/r)^{C_1 1}. \end{aligned} \tag{3.5}$$

Combining this with (3.1), (3.2), (3.3) and (3.4), we find that

$$\mathbf{P}_\delta\big[\mathcal{A}(z_0, r, \epsilon)\big] \leqslant C_6(\epsilon_1/c)^{C_7} + C_1 2 m^2 (\epsilon/r)^{C_1 1} + \Big(1 - C_8 \epsilon_1^{C_9}\Big)^{m-1}.$$

The lemma follows by taking $m := \lfloor \epsilon^{-C_1 1/3} \rfloor$ and $\epsilon_1 := -1/\log \epsilon$, say. ∎

THEOREM 3.5: *Let (X, d) be a compact metric space, let $o_1, o_2 \in X$, let f: $(0, \infty) \to (0, \infty)$ be monotone increasing and continuous, and let $\Gamma = \Gamma(f)$ be the set of all compact simple paths $\gamma \subset X$ with endpoints o_1 and o_2 which satisfy the following property. Whenever x, y are points in γ and $D(x, y)$ is the diameter of the arc of γ joining x and y we have*

$$d(x, y) \geqslant f(D(x, y)). \tag{3.6}$$

Then Γ is compact in the Hausdorff metric.

For this we will need Janiszewski's [Jan12] topological characterization of $[0, 1]$ (see [New92, IV.5]):

LEMMA 3.6 (Topological Characterization of Arcs): *Let K be a compact, connected metric space, and let $o_1, o_2 \in K$. Suppose that for every $x \in K - \{o_1, o_2\}$ the set $K - \{x\}$ is disconnected. Then K is homeomorphic to $[0, 1]$.* ∎

Proof of Theorem 3.5: Let $\mathcal{H} = \mathcal{H}(X)$ denote the space of compact nonempty subsets of X with the Hausdorff metric $d_{\mathcal{H}}$. Let γ be in the closure of Γ in $\mathcal{H}$.

Then γ is connected, compact, and $\gamma \supset \{o_1, o_2\}$. We now use Lemma 3.6 to show that γ is a simple path. Indeed, suppose that $x \in \gamma - \{o_1, o_2\}$.

We show that o_1 and o_2 are in distinct components of $\gamma - \{x\}$. Let $\{\gamma_n\}$ be a sequence in Γ such that $d_{\mathcal{H}}(\gamma_n, \gamma) < 1/n$, and let $\{x_n\}$ be a sequence with $x_n \in \gamma_n$ and $d(x_n, x) < 1/n$. For each n let $\gamma_n^{o_1}$ be the closed arc of γ_n with endpoints o_1 and x_n, and let $\gamma_n^{o_2}$ be the closed arc of γ_n with endpoints x_n and o_2. By passing to a subsequence, if necessary, assume with no loss of generality that the Hausdorff limits $\gamma^{o_1} = \lim \gamma_n^{o_1}$ and $\gamma^{o_2} = \lim \gamma_n^{o_2}$ exist. If $p_n \in \gamma_n^{o_1}$, $q_n \in \gamma_n^{o_2}$, then $d(p_n, q_n) \geqslant f\big(d(p_n, x_n)\big)$, since $\gamma_n \in \Gamma(f)$. By taking limits we find that if $p \in \gamma^{o_1}$ and $q \in \gamma^{o_2}$, then $d(p, q) \geqslant f\big(d(p, x)\big)$. Consequently, $\gamma^{o_1} - \{x\}$ and $\gamma^{o_2} - \{x\}$ are disjoint. Because $\gamma^{o_1} \cup \gamma^{o_2} = \gamma$ and $\gamma^{o_1}, \gamma^{o_2}$ are compact, the set $\gamma - \{x\}$ is not connected. Hence, by Lemma 3.6, γ is a simple path.

It remains to prove that $\gamma \in \Gamma(f)$. For any simple path $\beta \subset X$ with endpoints o_1, o_2, let $R(\beta)$ be the set of all $(x, y) \in \beta \times \beta$ such that x belongs to the subarc of β with endpoints o_1 and y. With arguments as above, it is not hard to show that $\lim R(\gamma_n) = R(\gamma)$, in the Hausdorff metric on $X \times X$. Since f is continuous, it then easily follows that $\gamma \in \Gamma(f)$. The details are left to the reader. (Actually, one can see that this statement is not essential for the proof of Theorem 1.1. There, we only need the fact that the Hausdorff closure of $\Gamma(f)$ is contained in the set of simple paths.) ∎

Proof of Theorem 1.1: We start with the proof of the second statement, and first assume that $a, b \neq \infty$. Let o_1^* and o_2^* be vertices of $\delta\mathbb{Z}^2$ closest to a and b, respectively. Let ω_δ be LERW from o_1^* to o_2^* in $\delta\mathbb{Z}^2$, and let ω_δ' be a path from a to b, obtained by taking the line segment joining a to a closest point a' on ω_δ, taking the line segment joining b to a closest point b' on ω_δ, and taking the path in ω_δ joining a' and b'. Then the Hausdorff distance from ω_δ to ω_δ' is less than 2δ. Let $\mathbf{P}_\delta'$ be the law of ω_δ'. Let μ be some subsequential weak limit of $\mathbf{P}_\delta$ as $\delta \to 0$. Then it is also a subsequential scaling limit of $\mathbf{P}_\delta'$. Let m be large. By Lemma 3.1, there is an R_m such that with probability at least $1 - 2^{-m-1}$ we have $\omega_\delta \subset B(o_1, R_m)$. For each $j \in \mathbb{N}$, let $z_1^j, \ldots, z_{k_j}^j$ be a finite set of points in $\mathbb{R}^2$ such that the open balls of radius 2^{-j-2} about these points cover $\overline{B}(o_1, R_m)$. For each $j \in \mathbb{N}$, let $\epsilon_j^m \in (0, 1)$ be sufficiently small so that

$$\mathbf{P}_\delta\big[\mathcal{A}(z_i^j, 2^{-j}, \epsilon_j^m)\big] < 2^{-m-2-j}/k_j, \tag{3.7}$$

for all $i = 1, \ldots, k_j$, and for all $\delta > 0$. Such ϵ_j^m exist, by Lemma 3.4. Finally, let $f_m\colon (0, \infty) \to (0, \infty)$ be a continuous monotone increasing function satis-

fying $f_m(2^{2-j}) \leqslant 2^{-3}\min(\epsilon_j^m, 2^{-j})$ for each $j = 1, 2, \ldots$ and $\sup_{s>0} f_m(s) \leqslant 2^{-3}\min(\epsilon_1^m, 1/2)$.

Let $\mathcal{X}_m$ be the space of all compact nonempty subsets of $\overline{B}(o_1, R_m)$, and set

$$\mathcal{Q}_m := \mathcal{X}_m - \bigcup_{j=1}^{\infty}\bigcup_{i=1}^{k_j} \mathcal{A}(z_i^j, 2^{-j-1}, \epsilon_j^m).$$

Note that

$$\mathbf{P}'_\delta(\neg\mathcal{Q}_m) \leqslant 2^{-m}, \tag{3.8}$$

for all δ and m. Also note that if we set $X := \overline{B}(o_1, R_m)$ and $f := f_m$ in Theorem 3.5, then

$$\mathcal{Q}_m \subset \Gamma_m := \Gamma(f_m). \tag{3.9}$$

Indeed, suppose that $\gamma \in \mathcal{X}_m$ is a path in X joining a and b and $\gamma \notin \Gamma_m$. Then there are $x, y \in \gamma$ and w contained in the arc of γ joining x and y such that $\operatorname{dist}(x, y) < 2f(\operatorname{dist}(x, w))$. Since $\sup f \leqslant 2^{-4}$, we have $\operatorname{dist}(x, y) < 2^{-3}$. Let j be such that $2^{-j+1} \leqslant \operatorname{dist}(x, w) \leqslant 2^{-j+2}$. Then

$$\operatorname{dist}(x, y) < 2f(\operatorname{dist}(x, w)) \leqslant 2f(2^{-j+2}) \leqslant 2^{-j-2}, \quad \operatorname{dist}(x, y) < 2^{-3}\epsilon_j^m.$$

Let $i \in \{1, \ldots, k_j\}$ be such that $\operatorname{dist}(x, z_i^j) \leqslant 2^{-j-2}$. Then $\operatorname{dist}(z_i^j, w) \geqslant 2^{-j}$ and $x, y \in B(z_i^j, 2^{-j-1})$. Consequently, since $d(x, y) < \epsilon_j^m$, we have

$$\gamma \in \mathcal{A}(z_i^j, 2^{-j-1}, \epsilon_j^m).$$

This proves (3.9).

By (3.8) and (3.9), we get $\mathbf{P}'_\delta(\neg\Gamma_m) \leqslant 2^{-m}$. Theorem 3.5 tells us that Γ_m is compact. Therefore, we also have $\mu(\neg\Gamma_m) \leqslant 2^{-m}$, and so

$$\mu\left(\neg\bigcup_m \Gamma_m\right) = 0.$$

This completes the proof for the case $a, b \neq \infty$, because each element of $\bigcup_m \Gamma_m$ is a simple path.

The proof when a or b is ∞ is similar. One only needs to note that Lemma 3.4 is valid when $z_0 = \infty$ and the distances are measured in the spherical metric. Indeed, the basic harmonic measure estimate Lemma 2.1 is also valid in the context of the spherical metric.

The proof of the first statement of Theorem 1.1 is also similar. The details are left to the reader. ∎

Note that the first statement of Theorem 1.1 implies that for almost every subsequential scaling limit path γ from a to ∂D, the closure of $\gamma \cap D$ intersects ∂D in a single point. This fact is easy to deduce directly, since it is also true for the image of SRW starting near a and stopped when ∂D is hit.

4. First steps in the proof of Theorem 1.3

Throughout this section we assume Conjecture 1.2. Let σ be random, with the law of the scaling limit of LERW from 0 to $\partial\mathbb{U}$. From Theorem 1.1 we know that σ is a.s. a simple path.

Recall that if $D \subset D' \subsetneq \mathbb{C}$ are simply connected domains with $0 \in D$, then the conformal radius of D' is at least as large as the conformal radius of D. This follows from the Schwarz Lemma applied to the map $f_{D'}^{-1} \circ f_D \colon \mathbb{U} \to \mathbb{U}$.

For each $t \in (-\infty, 0]$, let σ_t be the subarc of σ with one endpoint in $\partial\mathbb{U}$ such that the conformal radius of $\mathbb{U} - \sigma_t$ is $\exp t$. It is clear that σ_t varies continuously in t. Let $f_t \colon \mathbb{U} \to \mathbb{U} - \sigma_t$ be the conformal map satisfying $f_t(0) = 0$ and $f_t'(0) = \exp t$. By Löwner's slit mapping theorem [Löw23], there is a unique continuous $\zeta = \zeta_\sigma \colon (-\infty, 0] \to \partial\mathbb{U}$ such that the differential equation (1.1) holds. Let $\widehat{\zeta} = \widehat{\zeta}_\sigma \colon (-\infty, 0] \to \mathbb{R}$ be the continuous function satisfying $\zeta(t) = \exp\bigl(i\widehat{\zeta}(t)\bigr)$ and $\widehat{\zeta}(0) \in [0, 2\pi)$. Our goal is to prove

PROPOSITION 4.1: *The law of $\widehat{\zeta}$ is stationary, and $\widehat{\zeta}$ has independent increments.*

This means that for each $s < 0$ the law of the map $t \mapsto \widehat{\zeta}(s+t)$ restricted to $(-\infty, 0]$ is the same as the law of $\widehat{\zeta}$, and that for every $n \in \mathbb{N}$ and $t_0 \leqslant t_1 \leqslant \cdots \leqslant t_n \leqslant 0$, the increments $\widehat{\zeta}(t_1) - \widehat{\zeta}(t_0), \widehat{\zeta}(t_2) - \widehat{\zeta}(t_1), \ldots, \widehat{\zeta}(t_n) - \widehat{\zeta}(t_{n-1})$ are independent. The proof of this proposition, as well as the next, will be completed in later sections.

Note that Conjecture 1.2 implies that the distribution of σ is invariant under rotations of $\mathbb{U}$ about 0. Let σ^1 be random with the law of σ conditioned to hit ∂U at 1. If λ denotes the (random) point in $\sigma \cap \partial\mathbb{U}$, then σ^1 has the same law as $\lambda^{-1}\sigma$. It turns out that Proposition 4.1 will follow quite easily from

PROPOSITION 4.2: *Assume Conjecture 1.2. Fix some $t < 0$. Take σ^1 and σ to be independent. As above, let σ_t be the compact arc of σ that has one endpoint on $\partial\mathbb{U}$ and such that the conformal radius of $\mathbb{U} - \sigma_t$ is $\exp(t)$. Let $q(t)$ be the endpoint of σ_t that is in $\mathbb{U}$. Let ϕ be the conformal map from $\mathbb{U}$ onto $\mathbb{U} - \sigma_t$ satisfying $\phi(0) = 0$ and $\phi(1) = q(t)$. Then $\sigma_t \cup \phi(\sigma^1)$ has the same law as σ.*

Now comes an easy lemma about LERW, and Proposition 4.2 will be obtained from this lemma by passing to the scaling limit. The passage to the scaling limit is quite delicate. Recall the definition of the graph $G(D,\delta)$ approximating a domain D, from Section 2.

LEMMA 4.3: *Let $\delta > 0$ and $t < 0$ be fixed, and let $D \subsetneq \mathbb{C}$ be a simply connected domain with $0 \in D$. Let β be LERW from 0 to ∂D in $G(D,\delta)$. Let β_t be the compact arc in β such that $\beta_t \cap \partial D$ is an endpoint of β_t and such that the conformal radius of $D - \beta_t$ is $\exp(t)$. Let $q_1(t)$ be the endpoint of β_t that is not on $\partial\mathbb{U}$. Set $D_t = D - \beta_t$. Then the law of $\beta - \beta_t$ conditioned on β_t is equal to the law of LERW from 0 to ∂D_t, conditioned to hit $q_1(t)$.*

Proof: There are several different ways to prove this lemma. We prove it using the relation between LERW and the UST. Suppose that α is a path such that $\beta_t = \alpha$ has positive probability. We assume for now that the endpoint q of α which is in D lies in the relative interior of an edge e of $\delta\mathbb{Z}^2$ (this must be true except for at most a countable possible choices of t), and set $\tilde{\alpha} := \alpha - e$. Let $\tilde{q}$ be the endpoint of $\tilde{\alpha}$ in D. Let T be the UST on $G^W(D,\delta)$, the wired graph of D. Then β may be taken as the path in T from 0 to ∂D. We may generate T using Wilson's algorithm with root ∂D, and starting with vertices $v_1 = \tilde{q}$ and $v_2 = 0$. Conditioning on β_t being equal to α is the same as conditioning on $\alpha \subset \beta$, which is the same as conditioning on the LERW LE_1 from v_1 to ∂D to be $\tilde{\alpha}$ and that the LERW from v_2 to LE_1 hits $\tilde{q}$ through the edge e. This completes the proof in the case where q is not a vertex of $\delta\mathbb{Z}^2$. The case where $q \in \mathsf{V}(\delta\mathbb{Z}^2)$ is treated similarly. ∎

5. Getting uniform convergence

The principle goal of this section is to state and prove Proposition 5.5 below. The main point there is that Conjecture 1.2 implies that the weak convergence of loop erased random walk in a domain $D \subset \mathbb{U}$ is uniform in D.

The first lemma shows that it is unlikely for a simple random walk to get far away from the boundary of a domain D after getting close to ∂D but before hitting ∂D, even if the walk is conditioned to hit ∂D at a certain set of vertices Q.

LEMMA 5.1: *Let K be a compact connected set in $\overline{\mathbb{U}}$ that contains $\partial\mathbb{U}$ but with $0 \notin K$. Let F be a compact subset of $\mathbb{U} - K$, and let $\epsilon > 0$. Then there is a $\delta_1 > 0$ with the following property. Let K' be a compact connected set in $\overline{\mathbb{U}}$ with*

$K' \supset \partial\mathbb{U}$ and $d_{\mathcal{H}(\mathbb{U})}(K, K') < \delta_1/5$. Set $D = \mathbb{U} - K'$. Let $\delta \in (0, \delta_1/5)$, and let $Q \subset \mathsf{V}^0_\partial(D, \delta)$ be nonempty. Let RW^Q be the random walk on $G(D, \delta)$ starting at 0 that stops when it hits $\mathsf{V}_\partial(D, \delta)$, conditioned to hit Q. Then the probability that RW^Q will reach F after visiting some vertex within distance δ_1 of K is less than ϵ.

Proof: We need to recall some basic facts relating the conditioned random walk RW^Q to the unconditioned random walk RW (that stops when hitting $\mathsf{V}_\partial(D, \delta)$). First recall that RW^Q is a Markov chain. (This is easy to prove directly. See also the discussion of Doob's h-transform in [Dur84, §3.1].) Let v_0 be some vertex, $t_0 \in \mathbb{N}$, and W a set of vertices. Let $\tau = \tau_W$ be the least $t \geqslant t_0$ such that $\mathsf{RW}^Q(t) \in W$, if such exists, and otherwise set $\tau = \infty$. For $w \in W$ let

$$a_W(v_0, w) := \mathbf{P}[\tau < \infty, \mathsf{RW}^Q(\tau) = w \mid \mathsf{RW}^Q(t_0) = v_0],$$

and let $\tilde{a}_W(v_0, w)$ be the corresponding quantity for RW. Then

$$a_W(v_0, w) = \frac{\tilde{a}_W(v_0, w) h(w)}{h(v_0)}, \tag{5.1}$$

where $h(v)$ is the probability that RW hits Q when it starts at v. This formula is easy to verify.

Let W_1 be the set of vertices v of $G(D, \delta)$ such that $h(v) < \epsilon h(0)/5$. By (5.1),

$$\sum_{w \in W_1} a_{W_1}(0, w) \leqslant (\epsilon/5) \sum_{w \in W_1} \tilde{a}_{W_1}(0, w) \leqslant \epsilon/5.$$

Consequently, the probability that RW^Q visits W_1 is at most $\epsilon/5$.

Let ρ be the distance from F to K, and assume that $100\delta_1 < \rho$. Then an easy discrete Harnack inequality shows that $h(v)/h(0) < C_0$ for all vertices v in F, where C_0 is some constant which does not depend on K' or δ, but may depend on F. There is a first vertex, say $\widehat{v}$, visited by RW^Q such that the distance from $\widehat{v}$ to K is at most δ_1. Let W_2 be the set of vertices of $G(D, \delta)$ in F. If $\widehat{v} \notin W_1$, then $h(\widehat{v}) \geqslant \epsilon h(0)/5$ and hence

$$\sum_{w \in W_2} a_{W_2}(\widehat{v}, w) = \sum_{w \in W_2} \tilde{a}_{W_2}(\widehat{v}, w) \frac{h(w)}{h(0)} \frac{h(0)}{h(\widehat{v})} \leqslant 5C_0\epsilon^{-1} \sum_{w \in W_2} \tilde{a}_{W_2}(\widehat{v}, w). \tag{5.2}$$

But since K is connected, Lemma 2.1 shows that the probability that a simple random walk starting at $\widehat{v}$ will get to distance $\rho/2$ from $\widehat{v}$ without hitting $V_\partial(D, \delta)$ is bounded by $C_1(\delta_1/\rho)^{C_2}$, where $C_1, C_2 > 0$ are absolute constants. By (5.2),

$$\sum_{w \in W_2} a_{W_2}(\widehat{v}, w) \leqslant 5C_0\epsilon^{-1} C_1 (\delta_1/\rho)^{C_2}.$$

Consequently, if δ_1 is chosen sufficiently small, the probability that RW^Q starting at $\widehat{v}$ will hit W_2 is less than $\epsilon/5$, provided $\widehat{v} \notin W_1$. But $\mathbf{P}[\widehat{v} \in W_1] \leqslant \epsilon/5$, since the probability that W_1 is visited is at most $\epsilon/5$. The lemma follows. ∎

In the following, we let $\mathsf{RW}^{D,\delta}$ denote SRW on $G(D,\delta)$ that stops when it hits $\mathsf{V}_\partial(D,\delta)$, and let $\mathsf{RW}^{D,\delta,Q}$ denote $\mathsf{RW}^{D,\delta}$ conditioned to hit Q, if $Q \subset \mathsf{V}^0_\partial(D,\delta)$. Suppose that ν is a probability measure on $\mathsf{V}^0_\partial(D,\delta)$ and p is random with law ν; then $\mathsf{RW}^{D,\delta,\nu}$ will denote $\mathsf{RW}^{D,\delta}$ conditioned to hit p given p. In other words, the law of $\mathsf{RW}^{D,\delta,\nu}$ is the convex combination of the laws of the walks $\mathsf{RW}^{D,\delta,\{p\}}$, with coefficients $\nu(\{p\})$.

LEMMA 5.2: *Assume Conjecture 1.2. Let $D \subset \mathbb{U}$ be a Jordan domain with $0 \in D$. Let ϕ: $D \to \mathbb{U}$ be the conformal homeomorphism from D to $\mathbb{U}$ satisfying $\phi(0) = 0$, $\phi'(0) > 0$. Then as $\delta \to 0$ the law of the pair*

$$\left(\phi(\mathsf{LE}(\mathsf{RW}^{D,\delta})), \partial\mathbb{U} \cap \phi(\mathsf{RW}^{D,\delta})\right)$$

tends weakly to the law of the pair $(\sigma, \partial\mathbb{U} \cap \sigma)$.

This lemma is easily proved using arguments as in the proof of Lemma 5.1, and is therefore left to the reader.

If X and Y are random closed subsets of $\mathbb{U}$, we let $\widehat{d}_{\mathbb{U}}(X,Y)$ denote the Prohorov distance between the law of $X \cup \partial\mathbb{U}$ and the law of $Y \cup \partial\mathbb{U}$ (see Section 2, towards the end), where the metric $d_{\mathcal{H}(\overline{\mathbb{U}})}$ is used on $\mathcal{H}(\overline{\mathbb{U}})$. If F is a subset of $\overline{\mathbb{U}}$, we set $\widehat{d}_F(X,Y) := \widehat{d}_{\mathbb{U}}\left(X \cup \overline{\mathbb{U} - F}, Y \cup \overline{\mathbb{U} - F}\right)$. This is a measure of how much X and Y differ inside F.

The next lemma shows that up to a specified accuracy the loop-erased random walk on a domain D can be approximated by the loop-erased random walk on another domain D', where D' is selected from a finite collection of smooth Jordan domains.

LEMMA 5.3: *Let $\epsilon > 0$. Then there is a $\delta_0 > 0$ and a finite collection of smooth Jordan domains $D_1, D_2, \ldots, D_n \subset \mathbb{U}$ with $0 \in D_j$ for all j, and with the following property. Let $D \subset \mathbb{U}$ be a simply connected domain with $B(0,\epsilon) \subset D$, let $\delta \in (0, \delta_0)$ and let $Q \subset \mathsf{V}^0_\partial(D,\delta)$ be nonempty. Let F_ϵ be the component of 0 in the set of points in D that have distance at least ϵ to ∂D. Then there is a $D' \in \{D_1, \ldots, D_n\}$ and a probability measure ν on $\mathsf{V}^0_\partial(D',\delta)$ such that*

$$\widehat{d}_{F_\epsilon}(\mathsf{LE}(\mathsf{RW}^{D,\delta,Q}), \mathsf{LE}(\mathsf{RW}^{D',\delta,\nu})) < \epsilon. \tag{5.3}$$

Moreover, we may require that the Hausdorff distance from ∂D to $\partial D'$ is at most ϵ.

Proof: We first prove the lemma for some specific D and Q as above. By Lemma 5.1, there is a $\delta_1 > 0$ such that with probability $\geqslant 1 - \epsilon/5$ the walk $\mathsf{RW}^{D,\delta,Q}$ does not reach F_ϵ after exiting F_{δ_1}, provided $\delta \in (0, \delta_1/5)$. Also, there is a $\delta_1' < \delta_1$ such that with probability $\geqslant 1 - \epsilon/5$ this walk does not reach F_{δ_1} after exiting $F_{\delta_1'}$. Let $D' \subset D$ be a smooth Jordan domain with $D' \supset F_{\delta_1'}$, and such that the Hausdorff distance from $\partial D'$ to ∂D is less than ϵ. For every $\delta < \delta_1'/5$, let $\nu = \nu_\delta$ be the hitting measure of $\mathsf{RW}^{D,\delta,Q}$ on $\mathsf{V}_\partial^0(D', \delta)$. Observe that we may think of $\mathsf{RW}^{D',\delta,\nu}$ as equal to $\mathsf{RW}^{D,\delta,Q}$ stopped when $V_\partial(D', \delta)$ is hit.

Let $\mathcal{A}_1$ be the event that $\mathsf{RW}^{D,\delta,Q}$ does not visit F_ϵ after exiting F_{δ_1}, and let $\mathcal{A}_2$ be the event that $\mathsf{RW}^{D,\delta,Q}$ does not visit F_{δ_1} after exiting $F_{\delta_1'}$. Note that on the event $\mathcal{A}_1 \cap \mathcal{A}_2$, after exiting $F_{\delta_1'}$ the random walk does not visit any vertex v which was already visited prior to the last visit to F_ϵ. Consequently, the intersection of F_ϵ with the loop erasure of the walk does not change after the first exit of $F_{\delta_1'}$. Since we may couple $\mathsf{RW}^{D',\delta,\nu}$ to equal $\mathsf{RW}^{D,\delta,Q}$ stopped on $V_\partial(D', \delta)$, this means that we may obtain a coupling giving $F_\epsilon \cap \mathsf{LE}(\mathsf{RW}^{D,\delta,Q}) = F_\epsilon \cap \mathsf{LE}(\mathsf{RW}^{D',\delta,\nu})$ on $\mathcal{A}_1 \cap A_2$. Since $\mathbf{P}[\mathcal{A}_1 \cap \mathcal{A}_2] \geqslant 1 - \epsilon/2$, this proves the lemma for a single D. However, the same pair (D', δ_1) would work for every D'' with $\partial D''$ sufficiently close to ∂D in the Hausdorff metric. Hence, the compactness of the Hausdorff space of compact, connected subsets of $\overline{\mathbb{U}} - B(0, \epsilon)$ completes the proof. ∎

LEMMA 5.4: *Let $D \subset \mathbb{U}$ be a smooth Jordan domain with $0 \in D$, let $\delta > 0$, let $q \in \mathsf{V}_\partial^0(D, \delta)$, and let $\mathsf{RW}^q := \mathsf{RW}^{D,\delta,\{q\}}$. Then the law of $\mathsf{LE}(\mathsf{RW}^q)$ is uniformly continuous in q. That is, for every $\epsilon > 0$ there is a $\delta_1 > 0$ such that*

$$\widehat{d}_D(\mathsf{LE}(\mathsf{RW}^q), \mathsf{LE}(\mathsf{RW}^{q'})) < \epsilon$$

provided $\delta \in (0, \delta_1)$, and $|q - q'| < \delta_1$.

Proof: Let $\delta_0 > 0$ be very small. It is easy to see that when $|q - q'|$ is small we may couple RW^q and $\mathsf{RW}^{q'}$ so that with probability at least $1 - \epsilon/5$ they are equal until they both come within distance δ_0 of ∂D. Hence, the lemma follows by using an argument similar to the one used in the proof of Lemma 5.3. ∎

We now come to the principle goal of this section, proving that Conjecture 1.2 implies that the weak convergence of loop erased random walk in a domain $D \subset \mathbb{U}$ is uniform in D.

PROPOSITION 5.5: *Assume Conjecture 1.2, and let $\epsilon > 0$. Then there is a $\delta_1 > 0$ with the following property. Let $D \subset \mathbb{U}$ be a simply connected domain with $B(0, \epsilon) \subset D$, and let F_ϵ be the connected component of 0 in the set of all points*

z with $d(z,\partial D) \geqslant \epsilon$. Let $\delta \in (0,\delta_1)$, and let $Q \subset \mathsf{V}^0_\partial(D,\delta)$ be nonempty. Let ϕ be the conformal homeomorphism from $\mathbb{U}$ to D that satisfies $\phi(0) = 0$ and $\phi'(0) > 0$. Then there is a random $\lambda \in \partial\mathbb{U}$ independent of σ^1 such that

$$\widehat{d}_{F_\epsilon}\left(\phi(\lambda\sigma^1), \mathsf{LE}(\mathsf{RW}^{D,\delta,Q})\right) < \epsilon.$$

The main point here is that δ_1 does not depend on D or on Q.

Proof: Let $\epsilon_1 > 0$ be much smaller than ϵ. Suppose that D' is a domain in the list appearing in Lemma 5.3 that satisfies the requirements there with ϵ_1 in place of ϵ, and let ν be as in that lemma. Fix some small δ_1. For each $q \in \mathsf{V}^0_\partial(D',\delta)$ let ν_q be restriction of the hitting measure of $\mathsf{RW}^{D',\delta}$ to $B(q,\delta_1) \cap V_\partial(D,\delta)$, normalized to be a probability measure. Lemma 5.4 implies that we may replace ν by a probability measure ν', which is a convex combination of such ν_q, while having

$$\widehat{d}_{F_{\epsilon_1}}(\mathsf{LE}(\mathsf{RW}^{D',\delta,\nu}), \mathsf{LE}(\mathsf{RW}^{D',\delta,\nu'})) < \epsilon_1, \tag{5.4}$$

provided δ_1 is sufficiently small. Moreover, by Conjecture 1.2, provided δ_1 is sufficiently small and $\delta \in (0,\delta_1)$, we have

$$\widehat{d}_{F_{\epsilon_1}}(\mathsf{LE}(\mathsf{RW}^{D',\delta,\nu_q}), \psi(\lambda_q\sigma^1)) \leqslant \epsilon_1, \tag{5.5}$$

where $\lambda_q \in \partial\mathbb{U}$ is random and independent of σ^1, and ψ is the conformal homeomorphism from $\mathbb{U}$ to D' satisfying $\psi(0) = 0$ and $\psi'(0) > 0$. Since the list $D_1,\ldots,D_n$ in Lemma 5.3 is finite, we may take δ_1 to be independent of D. Consequently, there is a random $\lambda \in \partial\mathbb{U}$ independent of σ^1 with

$$\widehat{d}_{F_{\epsilon_1}}(\mathsf{LE}(\mathsf{RW}^{D',\delta,\nu'}), \psi(\lambda\sigma^1)) \leqslant \epsilon_1. \tag{5.6}$$

Provided we have chosen ϵ_1 sufficiently small, we have that $|\psi(\phi^{-1}(z)) - z| < \epsilon_1$ for $z \in F_{\epsilon/2}$. Proposition 5.5 now follows from (5.3) with ϵ_1 in place of ϵ and from (5.4), (5.5) and (5.6). ∎

6. Recognizing the Löwner parameter as Brownian motion

Proof of Proposition 4.2: Let $\beta := \mathsf{LE}(\mathsf{RW}^{\mathbb{U},\delta})$ and let β_t, D_t and $q_1(t)$ be defined as in Lemma 4.3, where $D := \mathbb{U}$. Set $\gamma := \mathsf{LE}(\mathsf{RW}^{D_t,\delta,q_1(t)})$, where $\mathsf{RW}^{D_t,\delta,q_1(t)}$ is taken to be independent of β conditioned on β_t. Using Conjecture 1.2, $\widehat{d}_{\mathbb{U}}(\beta,\sigma) \to 0$ as $\delta \to 0$. By (2.1), this means that we may couple β and σ

(that is, make them defined on the same probability space, where they are not necessarily independent) such that $\beta \xrightarrow{\mathbf{P}} \sigma$ in $\mathcal{H}(\mathbb{U})$, where $\xrightarrow{\mathbf{P}}$ denotes convergence in probability as $\delta \to 0$. Since σ is a.s. a simple path, this also implies that $\beta_t \xrightarrow{\mathbf{P}} \sigma_t$.

Let $\widehat{\phi}$ be the conformal map from $\mathbb{U}$ onto $\mathbb{U} - \sigma_t$ that satisfies $\widehat{\phi}(0) = 0$ and $\widehat{\phi}'(0) > 0$, and let $\widehat{\psi}$ be the similarly normalized conformal map from $\mathbb{U}$ onto $\mathbb{U} - \beta_t$. Because $\beta_t \xrightarrow{\mathbf{P}} \sigma_t$, it follows that $\widehat{\psi} \xrightarrow{\mathbf{P}} \widehat{\phi}$, in the topology of uniform convergence on compact subsets of $\mathbb{U}$.

Set $\widehat{\psi}_\lambda(z) := \widehat{\psi}(\lambda z)$ and $\widehat{\phi}_\lambda(z) := \widehat{\phi}(\lambda z)$ for $\lambda \in \partial\mathbb{U}$. Given every $\epsilon > 0$ and a closed set $A \subset \mathbb{U}$ let $F_\epsilon(A)$ be the connected component of 0 in the set of points with distance at least ϵ from A (or the empty set, if $d(0, A) < \epsilon$), and let $W_\epsilon(A) := \overline{\mathbb{U} - F_\epsilon(A)}$.

By Proposition 5.5, for every $\epsilon > 0$ there is a random $\lambda \in \partial\mathbb{U}$ independent of σ^1 (but not of β_t) such that $\widehat{d}_{F_\epsilon(\beta_t)}(\widehat{\psi}_\lambda(\sigma^1), \gamma) \to 0$. (The law of λ may depend on δ and ϵ.) Observe that $\mathbf{P}[F_{2\epsilon}(\sigma_t) \subset F_\epsilon(\beta_t)] \to 1$ as $\delta \to 0$, because $\beta_t \xrightarrow{\mathbf{P}} \sigma_t$. Therefore, we may conclude that $\widehat{d}_{F_{2\epsilon}(\sigma_t)}(\widehat{\psi}_\lambda(\sigma^1), \gamma) \to 0$. Since this is true for every $\epsilon > 0$, it follows that we may choose $\lambda = \lambda_\delta$ so that $\widehat{d}_{\mathbb{U}-\sigma_t}(\widehat{\psi}_\lambda(\sigma^1), \gamma) \to 0$, as $\delta \to 0$. Because $\widehat{\psi}_\lambda \xrightarrow{\mathbf{P}} \widehat{\phi}_\lambda$, we therefore also have $\widehat{d}_{\mathbb{U}-\sigma_t}(\widehat{\phi}_\lambda(\sigma^1), \gamma) \to 0$, that is, $\widehat{d}_{\mathbb{U}}(\sigma_t \cup \widehat{\phi}_\lambda(\sigma^1), \sigma_t \cup \gamma) \to 0$. Since $\beta_t \cup \gamma$ has the same law as β (by Lemma 4.3), and since $\beta_t \xrightarrow{\mathbf{P}} \sigma_t$, this gives,

$$\widehat{d}_{\mathbb{U}}(\sigma_t \cup \widehat{\phi}_\lambda(\sigma^1), \beta) = \widehat{d}_{\mathbb{U}}(\sigma_t \cup \widehat{\phi}_\lambda(\sigma^1), \beta_t \cup \gamma) \to 0. \tag{6.1}$$

Let λ^* be random in $\partial\mathbb{U}$ with a law that is some weak (subsequential) limit of the law of λ as $\delta \to 0$. It follows from (6.1) that $\sigma_t \cup \widehat{\phi}_{\lambda^*}(\sigma^1)$ has the same law as σ. In particular, it is a simple path. The only possibility is therefore that $\widehat{\phi}_{\lambda^*} = \phi$ a.s., which completes the proof of Proposition 4.2. ∎

Proof of Proposition 4.1: Recall that $\zeta = \zeta_\sigma$: $(-\infty, 0] \to \partial\mathbb{U}$ is the Löwner parameter associated to the LERW scaling limit $\sigma \subset \mathbb{U}$. Let $\tilde{\zeta}$ be the Löwner parameter associated with the path σ^1, and let $\tilde{f}_t$ be the associated solution of the Löwner system. Note that $\tilde{\zeta}(0) = 1$, since $\sigma^1 \cap \partial\mathbb{U} = \{1\}$. Fix some $t_0 < 0$. Using Proposition 4.2 and its notations (with t_0 replacing t), we know that the path $\check{\sigma} := \sigma_{t_0} \cup \phi(\sigma^1)$ has the same law as σ. Let $\check{f}_t$ be the solution of the Löwner system associated with the path $\check{\sigma}$, and let $\check{\zeta}$ be the associated Löwner parameter. Then $\check{\zeta}$ has the same law as ζ, by Proposition 4.2. Set $\lambda := |\phi'(0)|/\phi'(0)$. When $t < t_0$, we have

$$\check{f}_t(z) = \phi \circ \tilde{f}_{t-t_0}(\lambda z),$$

because the right hand side is a suitably normalized conformal map from $\mathbb{U}$ onto $\mathbb{U} - \check{\sigma}_t$. We differentiate with respect to t, and use (1.1), to get

$$\frac{\partial}{\partial t}\check{f}_t(z) = \phi'(\tilde{f}_{t-t_0}(\lambda z))\frac{\partial}{\partial t}\tilde{f}_{t-t_0}(\lambda z)$$
$$= \phi'(\tilde{f}_{t-t_0}(\lambda z))\lambda z \tilde{f}'_{t-t_0}(\lambda z)\frac{\tilde{\zeta}(t-t_0)+\lambda z}{\tilde{\zeta}(t-t_0)-\lambda z} = z\check{f}'_t(z)\frac{\lambda^{-1}\tilde{\zeta}(t-t_0)+z}{\lambda^{-1}\tilde{\zeta}(t-t_0)-z}.$$

Consequently, it follows that $\check{\zeta}(t) = \lambda^{-1}\tilde{\zeta}(t-t_0)$ for $t < t_0$. It is clear that $\check{\zeta} = \zeta$ for $t \in [t_0, 0]$. Continuity of $\check{\zeta}$ gives $\lambda^{-1} = \zeta(t_0)/\tilde{\zeta}(0) = \zeta(t_0)$. Since ζ and $\tilde{\zeta}$ are independent, and $\tilde{\zeta}$ has the same law as ζ conditioned on $\zeta(0) = 1$, Proposition 4.1 follows. ∎

We shall need the following

THEOREM 6.1: *Let $a(t)$, $t \geqslant 0$, be a real valued process (that is, a random function a: $[0,\infty) \to \mathbb{R}$). Suppose that a is continuous a.s. and for every $n \in \mathbb{N}$ and every $(n+1)$-tuple $0 = t_0 \leqslant t_1 \leqslant t_2 \leqslant \cdots \leqslant t_n$, the increments $a(t_j) - a(t_{j-1})$, $j = 1, \ldots, n$, are independent. Then for every fixed $s_0 \in (0,\infty)$, the random variable $a(s_0)$ is Gaussian.*

This theorem follows from the general theory of Lévy processes. An entirely elementary proof can be found in Section 4.2 of [Itô61].

COROLLARY 6.2: *There is a constant $c > 0$ such that the process $\zeta(t)$ has the same law as $\mathsf{B}(-ct)$, where $\mathsf{B}(t)$ is Brownian motion on $\partial\mathbb{U}$ started at a uniform random point.*

Proof: That $\zeta(t)$ has the same law as $\mathsf{B}(-ct)$ for some $c \geqslant 0$ follows immediately from Proposition 4.1 and Theorem 6.1. The fact that $c > 0$ is clear, since the LERW scaling limit is not equal a.s. to a line segment. ∎

7. The winding number of SLE

Let $\kappa \geqslant 0$, let $\mathsf{B}(t)$ be Brownian motion on $\partial\mathbb{U}$ started at a uniform random point on $\partial\mathbb{U}$, and set

$$\zeta = \zeta_\kappa := \mathsf{B}(-\kappa t). \tag{7.1}$$

Definition 7.1: Let $\mathfrak{K}$ denote the set of all $\kappa \geqslant 0$ such that the Löwner evolution f_t defined by (7.1), (1.1) and (1.2) is a.s. for every $t < 0$ a Riemann map to a slitted disk. For $\kappa \in \mathfrak{K}$, let ξ_κ denote the (random) path defined by $\xi_\kappa(t) =$

$f_t(\zeta(t))$. That is, ξ_κ is the path in $\overline{\mathbb{U}}$ such that f_t is the nomalized Riemann map to $\mathbb{U} - \xi_\kappa\big([t,0]\big)$.

The random process $\xi_\kappa\big([t,0]\big)$, $t \leqslant 0$, will be called **stochastic Löwner evolution** (SLE) with constant κ.

As before, we let $\widehat{\mathsf{B}}$: $[0,\infty) \to \mathbb{R}$ be the continuous map satisfying $\mathsf{B} = \exp i\widehat{\mathsf{B}}$ and $\widehat{\mathsf{B}}(0) \in [0, 2\pi)$.

THEOREM 7.2: *Let $\kappa \in \mathfrak{K}$. Let $T \leqslant 0$, and let $\theta_\kappa(T)$ be the winding number of the path $\xi_\kappa\big([T,0]\big)$ around 0, that is $\theta_\kappa(T) = \arg(\xi_\kappa(0)) - \arg(\xi_\kappa(T))$, with* arg *chosen continuous along ξ_κ. Then for all $s > 0$,*

$$\mathbf{P}\Big[\big|T - \log|\xi_\kappa(T)|\big| > s\Big] \leqslant C_0 \exp(-C_1 s), \tag{7.2}$$

and

$$\mathbf{P}\Big[\big|\theta_\kappa(T) - \widehat{\mathsf{B}}(0) + \widehat{\mathsf{B}}(-\kappa T)\big| > s\Big] \leqslant C_0 \exp(-C_1 s), \tag{7.3}$$

where $C_0, C_1 > 0$ are constants, which depend only on κ.

Loosely speaking, the theorem says that $t + i\widehat{\mathsf{B}}(-\kappa t)$ is a good approximation of the path $\log \xi_\kappa(t)$. A consequence of the theorem is that $\theta_\kappa(t)/\sqrt{\kappa t}$ converges to a gaussian of unit variance as $T \to -\infty$.

Before we prove the theorem, it is worthwhile to have another look at the maps Φ defined by $\Phi(z,t,s) := f_s^{-1}(f_t(z))$, $t \leqslant s \leqslant 0$, which were discussed in Remark 1.5. For every $t < s \leqslant 0$ the set $\gamma_t^s := \mathbb{U} - \Phi(\mathbb{U}, t, s)$ is a path in $\mathbb{U}$. If $r \in (t,s)$, then $\gamma_t^s = \gamma_r^s \cup \Phi\big(\gamma_t^r, r, s\big)$. Therefore, when s is held fixed and t is decreasing, the path γ_t^s is growing from its endpoint inside $\mathbb{U}$, and if t is held fixed and s is increasing, the path γ_t^s deforms conformally and a new arc is added to it to reconnect it to $\partial\mathbb{U}$.

Proof: Let f_t be defined by (7.1), (1.1) and (1.2). Set $\xi := \xi_\kappa$. Let $w(t,z) := f_t^{-1}\big(f_T(z)\big) = \Phi(z,T,t)$ $(t \in [T,0])$, and let $y = y(t,z) := \arg w(t,z)$, where $\arg w$ is chosen to be continuous in t.

By Remark 1.5, w satisfies the differential equation

$$\partial_t w = -w\frac{\mathsf{B}(-\kappa t) + w}{\mathsf{B}(-\kappa t) - w}, \tag{7.4}$$

where ∂_t denotes differentiation with respect to t. Set $x = x(t,z) := \log|w(t,z)|$. Then $w = \exp(x + iy)$, and (7.4) can be rewritten

$$\partial_t x + i\partial_t y = \frac{\sinh x + i\sin\big(\widehat{\mathsf{B}}(-\kappa t) - y\big)}{\cosh x - \cos\big(\widehat{\mathsf{B}}(-\kappa t) - y\big)}. \tag{7.5}$$

Let z_1 be a random point on $\partial\mathbb{U}$, chosen uniformly, and independent of the Brownian motion B. Then $w(0,z_1) = f_T(z_1)$ is some point on the boundary of $D_T := f_T(\mathbb{U})$. Note that ∂D_T is a connected set that contains $\partial\mathbb{U}$ and intersects the circle $\partial B\big(0,\exp(T)\big)$, by (2.2). Set $A_s := \big\{z \in \partial D_T\colon |z| > \exp(T+s)\big\}$. It follows from the continuous version of Lemma 2.1 for Brownian motion that the harmonic measure of A_s in D_T at 0 is bounded by $O(1)\exp(-C_2 s)$, for some constant $C_2 > 0$ and every $s \in \mathbb{R}$; that is, at zero, the bounded harmonic function on D_T that has boundary values 1 on A_s and has boundary values 0 on $\partial D_T - \overline{A}_s$ is bounded from above by $O(1)\exp(-C_2 s)$. Since harmonic measure is invariant under conformal maps, we conclude that the measure of $f_T^{-1}(A_s)$ is at most $O(1)\exp(-C_2 s)$. This means that

$$\mathbf{P}\big[\log|w(0,z_1)| - T > s\big] = \mathbf{P}\big[\log|f_T(z_1)| - T > s\big] \leqslant O(1)\exp(-C_2 s). \tag{7.6}$$

Now set $z_0 = \mathsf{B}(-\kappa T)$. Then $\xi(T) = w(0,z_0)$, and so we need to relate $|w(0,z_0)|$ and $|w(0,z_1)|$. Let τ be the least $t \in [T,0]$ such that $w(t,z_1) = \mathsf{B}(-\kappa t)$, if such a t exists, and set $\tau = 0$ if not. Note that $|w(t,z_1)| = 1$ while $t < \tau$, and $|w(t,z_1)| < 1$ for $t \in (\tau,0]$. Also observe that conditioned on $\tau < 0$, the law of the process $\big(w(t,z_1)\colon t \in [\tau,0]\big)$ is the same as the law of the process $\big(w(t+T-\tau,z_0)\colon t \in [\tau,0]\big)$. Consequently, the random variable $w(\tau - T, z_1)$ (where B is taken as two-sided Brownian motion and (7.4) is extended to the range $t > 0$), conditioned on $\tau < 0$, has the same distribution as the random variable $w(0,z_0)$. By (7.5), $\partial_t x \leqslant 0$, and therefore $|w(\tau - T, z_1)| \leqslant |w(0,z_1)|$ on the event $\tau < 0$. Thus, for every $s \in \mathbb{R}$ we have

$$\mathbf{P}\big[|w(0,z_1)| > s \mid \tau < 0\big] \geqslant \mathbf{P}\big[|w(0,z_0)| > s\big].$$

Because $|w(0,z_1)| = 1$ when $\tau = 0$, we may drop the conditioning on $\tau < 0$. Now (7.6) gives

$$\mathbf{P}\big[\log|\xi(T)| - T > s\big] = \mathbf{P}\big[\log|w(0,z_0)| - T > s\big] \leqslant O(1)\exp(-C_2 s).$$

On the other hand, the Koebe 1/4 Theorem (2.3) gives

$$\exp T = |f_T'(0)| \leqslant 4\inf\big\{|z|\colon z \notin f_T(\mathbb{U})\big\} \leqslant 4|\xi(T)|,$$

and so $\log|\xi(T)| + \log 4 \geqslant T$ always. This completes the proof of (7.2).

Now let τ_1 be the least $t \in [T,0]$ such that $x(t) = \log|w(t,z_0)| \leqslant -1$, and set $\tau_1 := 0$ if such a t does not exist. Since $x(t)$ is monotone decreasing, we may write $y(t)$ as a function of x: $y = g(x)$. By (7.5),

$$g'\big(x(t)\big) = \frac{\sin\big(\widehat{\mathsf{B}}(-\kappa t) - y(t)\big)}{\sinh x(t)}$$

and hence

$$\left|g'\big(x(t)\big)\right| \leqslant \left|\sinh x(t)\right|^{-1}.$$

And so we get

$$|y(0) - y(\tau_1)| = \int_{x(0)}^{x(\tau_1)} |g'(x)|\,dx \leqslant \int_{-\infty}^{-1} \left|\sinh x\right|^{-1} dx < \infty. \tag{7.7}$$

Let $\phi(s,t) := f_t^{-1}\big(\xi(s)\big)$ for $T \leqslant s \leqslant t \leqslant 0$. Then ϕ is continuous and its image does not contain 0. Hence, it may be considered as a homotopy in $\mathbb{C} - \{0\}$ from the path $\phi(s,0) = \xi(s)$, $s \in [T,0]$, to the concatenation of the inverse of the path $\phi(T,t) = f_t^{-1}\big(\xi(T)\big) = w(t,z_0)$, $t \in [T,0]$, with the path $\phi(t,t) = \mathsf{B}(-\kappa t)$, $t \in [T,0]$. Therefore, its winding number is the sum of the corresponding winding numbers. This means that

$$\theta_\kappa(T) = \widehat{\mathsf{B}}(0) - \widehat{\mathsf{B}}(-\kappa T) + y(T) - y(0).$$

By (7.7), it therefore suffices to prove the appropriate bound on the tail of $|y(\tau_1) - y(T)|$.

Let $|y|_{2\pi} := \min\{|y - 2\pi n|\colon n \in \mathbb{Z}\}$. Set $t_0 = T$, and inductively, let t_j be the first $t \in [t_{j-1}, 0]$ such that $\pi/2 = \left|\widehat{\mathsf{B}}(-\kappa t) - \widehat{\mathsf{B}}(-\kappa t_{j-1})\right|_{2\pi}$, and set $t_j = 0$ if no such t exists. Equation (7.5) shows that $\partial_t y(t)$ has the same sign as $\sin\big(\widehat{\mathsf{B}}(-\kappa t) - y(t)\big)$, and hence

$$\partial_t \left|y(t) - \widehat{\mathsf{B}}(-\kappa s)\right|_{2\pi} \leqslant 0, \quad \text{at } t = s,$$

for every $s \in (T,0)$. In other words, $y(t)$ is always moving in the direction which would decrease $\left|\exp\big(iy(t)\big) - \mathsf{B}(-\kappa t)\right|$. Consequently, for every $j \in \mathbb{N}$, if there is an $s \in [t_j, t_{j+1}]$ such that $\left|y(s) - \widehat{\mathsf{B}}(-\kappa t_j)\right|_{2\pi} < \pi/2$, then this is satisfied also for all $s' \in [s, t_{j+1})$, because $y(t)$ cannot get out of the set $\big\{p \in \mathbb{R}\colon \left|p - \widehat{\mathsf{B}}(-\kappa t_j)\right|_{2\pi} < \pi/2\big\}$ while $\widehat{\mathsf{B}}(-\kappa t)$ is in it. This implies that $|y(t_{j+1}) - y(t_j)| < 2\pi$. Hence $|y(\tau_1) - y(T)| \leqslant 2\pi \min\{j \in \mathbb{N}\colon t_j > \tau_1\}$. Therefore, for every $a > 0$ and $n \in \mathbb{N}$,

$$\mathbf{P}\big[|y(\tau_1) - y(T)| \geqslant 2\pi n\big] \leqslant \mathbf{P}[\tau_1 - T \geqslant a] + \mathbf{P}[t_n \leqslant T + a]. \tag{7.8}$$

The first summand on the right hand side is bounded by $O(1)\exp(-C_3\,a)$, for some constant $C_3 > 0$, by (7.2). To estimate the second summand, observe that conditioned on $t_n \leqslant T + a$, we have probability at least 2^{-n-1} for the event

$$B\big(-\kappa\,(T+a)\big) - \widehat{\mathsf{B}}(-\kappa T) \geqslant n\pi/2, \tag{7.9}$$

because when $t_n \leqslant T + a$ and $B\big(-\kappa(T+a)\big) \geqslant B\big(-\kappa t_n\big) \geqslant \cdots \geqslant B\big(-\kappa t_0\big)$, we have (7.9). However, (7.9) has probability

$$(2\pi a\kappa)^{-1/2} \int_{n\pi/2}^{\infty} \exp\big(-s^2/(2\kappa a)\big)\, ds \leqslant O(1)\exp\big(-n^2/C_4 a\big),$$

and hence

$$\mathbf{P}[t_n \leqslant T + a] \leqslant O(1)2^n \exp\big(-n^2/C_4 a\big).$$

We choose a to be n times a very small constant. Then our above estimates, together with (7.8), give

$$\mathbf{P}\big[|y(\tau_1) - y(T)| \geqslant 2\pi n\big] \leqslant O(1)\exp(-C_5 n),$$

with the constants depending only on κ. This completes the proof of the theorem. ∎

8. The twisting constant of LERW

Consider some scaling limit measure $\mathbf{P}$ of LERW from 0 to $\partial\mathbb{U}$, and let γ be random with law $\mathbf{P}$. Assuming Conjecture 1.2, we have established that SLE with some constant κ_o has law $\mathbf{P}$. In this section we show that $\kappa_o = 2$, and thereby complete the proof of Thereom 1.3.

Let $\epsilon \in (0,1)$, let γ_ϵ be the connected component of $\gamma - B(0,\epsilon)$ which has a point in ∂U, and let $W(\gamma_\epsilon, 0)$ be the winding number of γ_ϵ around 0, in radians. That is, $W(\gamma_\epsilon, 0)$ is the imaginary part of $\int_{\gamma_\epsilon} z^{-1} dz$. By symmetry, it is clear that $\mathbf{E}[W(\gamma_\epsilon, 0)] = 0$. We shall show that

$$\mathbf{E}[W(\gamma_\epsilon, 0)^2] = 2\log(1/\epsilon) + O(1)\sqrt{\log(1/\epsilon)}. \tag{8.1}$$

Based on this and the results of Section 7, it will follow that $\kappa_o = 2$.

The proof of (8.1) will use Kenyon's work [Ken98a]. The overall idea of the proof is very simple, and based on the relations between UST and domino tilings. We now briefly review the relations between the UST on $\mathbb{Z}^2$ and domino tilings, and the height function for domino tilings. For a more thorough discussion, the reader should consult [Ken98a].

A **domino tiling** of the grid $\mathbb{Z}^2$ is a tiling of $\mathbb{R}^2$ by tiles of the forms $[k, k+1] \times [j, j+2]$ and $[k, k+2] \times [j, j+1]$, where $k, j \in \mathbb{Z}$. A domino tiling of $\mathbb{Z}^2$ may also be thought of as a **perfect matching** of the dual grid $\big((1/2) + \mathbb{Z}\big)^2$. (A perfect matching of a graph G is a set of edges $M \subset \mathsf{E}(G)$ such that every vertex is incident with precisely one edge in M.)

Let us start with finite graphs. Let D be a simply connected domain in $\mathbb{R}^2$ whose boundary is a simple closed curve in the grid $\mathbb{Z}^2$, and let $G := \mathbb{Z}^2 \cap \overline{D}$. Let ρ_0 be some vertex in $\partial D \cap \mathbb{Z}^2$, which we call the **root**. Let $\check{G}$ be the graph $((1/2)\mathbb{Z}^2) \cap \overline{D}$ with ρ_0 and its incident edges removed. Then there is a bijection, discovered by Temperley, between the set of perfect matchings on $\check{G}$ and spanning trees of G.

Temperley's bijection (see Figure 8.1) works as follows. For every edge $[v, u]$ in the matching M such that $v \in \mathbb{Z}^2$, we put in the tree the edge e_u whose center is u. This gives the set of edges in the tree T. If $[v, u] \in M$ is as above, we may orient the edge e_u away from v, and then the tree T will be oriented towards the root ρ_0.

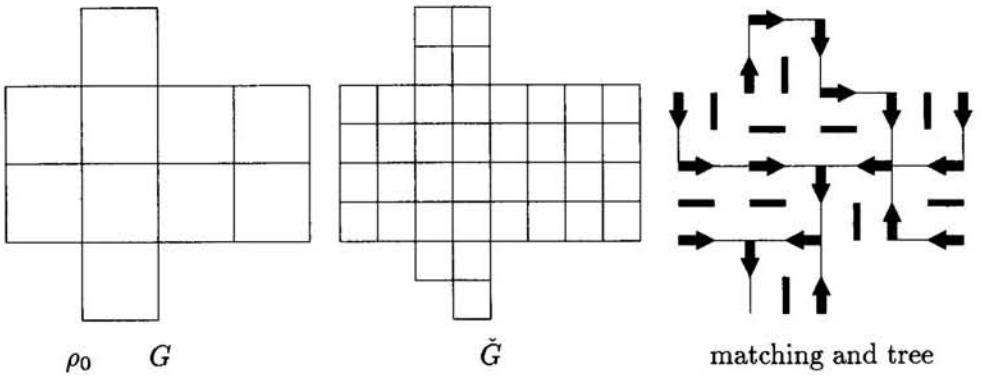


Figure 8.1. Temperley's bijection. On the right, the arrows are edges in the matching that contains vertices of $\mathbb{Z}^2$, the solid segments are other edges in the matching, and the thin lines are edges in the tree.

Temperley's bijection works also in more general situations. There is a simple modification to make it work for the wired graph associated to the domain D (recall the notion of the wired graph from Section 2). Also, given a perfect matching on all of $(1/2)\mathbb{Z}^2$, there is an associated (oriented) spanning forest of $\mathbb{Z}^2$. The collection of all domino tilings of $\mathbb{Z}^2$ has a natural stationary probability measure (of maximal entropy), and for a.e. domino tiling the corresponding spanning forest is a spanning tree. Temperley's map from perfect matchings on $(1/2)\mathbb{Z}^2$ to spanning forests of $\mathbb{Z}^2$ maps the cannonical probability measure on the set of domino tilings to the law of the UST of $\mathbb{Z}^2$.

Let G and $\check{G}$ be as above, and let $\widehat{G}$ be the graph of the domino tiling, that is, the union of the squares of edge length 1/2 with centers at the vertices of

$\check{G}$, thought of as a subgraph of the grid $(1/4,1/4)+(1/2)\mathbb{Z}^2$. Associated to a domino tiling of $\widehat{G}$ is a **height function** h defined on the vertices of $\widehat{G}$. Here is the definition of h. Pick some vertex $v_0 \in \mathsf{V}(\widehat{G})$ and some $a_0 \in \mathbb{R}$, and set $h(v_0) = a_0$. Color a square face of the grid $(1/4,1/4)+(1/2)\mathbb{Z}^2$ **white** if its center is a vertex of $\mathbb{Z}^2$ or if it is contained in a face of $\mathbb{Z}^2$, and **black** otherwise. If $[u,v] \in \mathsf{E}(\widehat{G})$ is on the boundary of a domino tile in the tiling, then we require that $h(v)-h(u)=1$ if the square to the right of the directed edge $[u,v]$ is white and $h(v)-h(u)=-1$ if the square to the right of $[u,v]$ is black. These constraints uniquely specify the height function h (except that the choices of v_0 and a_0 are arbitrary).

We will work in the upper half plane $\mathbb{H} := \{z \in \mathbb{C}\colon \operatorname{Im} z > 0\}$. Let $G_\delta(\mathbb{H}) := G^W(\mathbb{H},\delta)$, the wired graph of mesh δ associated with the domain $\mathbb{H}$, and $\widehat{G}_\delta(\mathbb{H}) := \mathbb{H} \cap \big((\delta/4,\delta/4)+(\delta/2)\mathbb{Z}^2\big)$. The discussion above carries through for the grid $\delta\mathbb{Z}^2$, in place of $\mathbb{Z}^2$. (Although the distance between adjacent vertices in the graph $\widehat{G}$ is $\delta/2$ when $G \subset \delta\mathbb{Z}^2$, we still work with the height function where the height difference along an edge on the boundary of a tile is ± 1.) Temperley's bijection induces a measure preserving transformation between domino tilings of the grid $\widehat{G}_\delta(\mathbb{H})$ and the UST of $G_\delta(\mathbb{H})$. (If we keep the orientation, then the UST is directed towards $\partial\mathbb{H}$.)

We normalize the height function associated to a domino tiling of $\widehat{G}_\delta(\mathbb{H})$ by requiring that $h\big((\delta/4,\delta/4)\big) = 1/2$. Then $h\big(((2k+1)\delta/4,\delta/4)\big) = (-1)^k/2$, for $k \in \mathbb{Z}$. If v is some vertex in $G_\delta(\mathbb{H})$, let $h(v)$ be the average of the value of h on the vertices of $\widehat{G}_\delta(\mathbb{H})$ closest to v.

LEMMA 8.1: *Let T be a spanning tree of $G_\delta(\mathbb{H})$, and let h be the associated height function. Let $v \in \mathsf{V}(G_\delta(\mathbb{H}))$ be a vertex different from the wired vertex $\partial\mathbb{H}$, and let a be the real part of v. Let Q'_v be the path from v to $\partial\mathbb{H}$ in T, considered as a path in the plane, and let Q_v be the union of Q'_v with the line segment joining the intersection $Q'_v \cap \partial\mathbb{H}$ to a. Then $-\pi h(v)/2 = W(Q_v,v)$, the winding number of Q'_v around v.*

This lemma is a special case of a more general observation made by Kenyon. (Since Q_v is a path with v as an endpoint, we define

$$W(Q_v,v) := \lim_{r\to 0} W\big(Q_v - B(v,r),v\big),$$

which is the same as $W\big(Q_v - B(v,\delta/2),v\big)$.)

Proof: Use induction on the length of the path Q'_v. ∎

Symmetry implies that $\mathbf{E}[h(v)] = 0$ for all $v \in \mathsf{V}(G_\delta(\mathbb{H}))$. Kenyon has shown [Ken98a] that

$$\mathbf{E}[h(v)^2] = 8\pi^{-2}\log(1/\delta) + O(1) \tag{8.2}$$

(provided that v stays in a compact subset of $\mathbb{H}$). Hence $\mathbf{E}[W(Q_v, v)^2] = 2\log(1/\delta) + O(1)$, which seems very close to a proof of (8.1). However, to make it into a proof of (8.1) requires some effort (it seems).

The advantage of the height function over the winding number is that the height difference between two vertices can be computed along any path joining them. On the other hand, to compute $W(Q_v, v)$, one might think that it is necessary to follow Q_v, which is a random path. It is immediate that $h(v)$ is $\sum_j \lambda_j \chi_j$, where χ_j is the indicator of the event that a certain domino tile is present in the tiling, and λ_j are some explicit easy to compute (non-random) weights. This means that to calculate $\mathbf{E}[h(v)^2]$ one needs to have a good estimate for the behavior of the correlations $\mathbf{E}[\chi_i \chi_j]$ for small δ. That's how Kenyon proves (8.2).

Recall that $A(p, r_1, r_2)$ denotes an annulus with center p, inner radius r_1, and outer radius r_2. The following result is an immediate consequence of [ABNW].

LEMMA 8.2: *Let $D \subset \mathbb{C}$ be a domain, and let $v_0 \in \delta\mathbb{Z}^2 \cap D$ be some vertex. Consider T, the UST on $\delta\mathbb{Z}^2 \cap D$, with free or wired boundary. Let $r_2 > 2r_1 > 2\delta$, and suppose that r_2 is smaller than the distance from v_0 to ∂D. Let A be the annulus $A := A(v_0, r_1, r_2)$, and let $k(A)$ be the maximum number of disjoint paths in T each of which intersects both boundary components of A. Then for each $k \in \mathbb{N}$*

$$\mathbf{P}[k(A) \geqslant k] \leqslant C_0 (r_2/r_1)^{C_1(k-1)},$$

where $C_0, C_1 > 0$ are universal constants. ∎

LEMMA 8.3: *Let $D \subset \mathbb{C}$ be some simply connected domain, and consider T, the UST in $G^W(D, \delta)$ (wired boundary). Let $p_0 \in D \cap \delta\mathbb{Z}^2$. Given a set $K \subset D$, let $X(K)$ denote the maximum winding number around p_0 of a path in T with endpoints in K. Let $r \in \big(0, d(p_0, \partial D)/2\big)$. Then for each $h > 0$ and $s \geqslant 2$,*

$$\mathbf{P}\Big[X\big(A(p_0, r/s, r)\big) > h\Big] \leqslant C_2 \exp\big(-C_3 h/\log s\big) \log s,$$

where C_2 and C_3 are absolute constants.

Proof: Set $A := A(p_0, r/s, r)$, $X := X(A)$. To begin, assume that $s = 2$ and that $r > 10\delta$. Let $B_1, B_2, \ldots, B_N$ be a covering of A with balls of radius $r/10$, where

$N \leqslant C_4$, with C_4 some universal constant. Let $v_0 \in A \cap \delta\mathbb{Z}^2$ be some vertex. For any vertex u, let $W(u)$ be the signed winding number around p_0 of the path in T from v_0 to u. Consider neighbors v, w in A, and let γ be the UST path joining v and w in T. If $\gamma \neq [v, w]$, then $\gamma \cup [v, w]$ is a simple closed path, and hence has winding number at most 2π around p_0. This implies that $|W(v) - W(u)|$ is at most 2π plus the absolute value of the winding number of the edge $[v, u]$ around p_0, and therefore $|W(v) - W(u)| \leqslant 3\pi$. Since any two vertices in A can be connected by a path in A, the set $\{W(v)\colon v \in A\}$ intersects every interval of length 3π which lies between its maximum and minimum. From this it follows that we may find vertices $v_1, v_2, \ldots, v_n$ in A such that $|W(v_j) - W(v_k)| \geqslant 3\pi$ for $j \neq k$ and $n \geqslant X/6\pi$. Since $N \leqslant C_4$, we may find some ball B_m from the above collection, satisfying $|B_m \cap \{v_1, \ldots, v_n\}| \geqslant X/6\pi C_4$. However, if $v, u \in B_m$ and $|W(v) - W(u)| \geqslant 2\pi$, then the path in T joining v and u must go around p_0. In particular, it must cross twice the annulus $A_m := A(c_m, r/10, r/5)$, where c_m is the center of B_m. It follows that the number of disjoint crossings of A_m in T is at least $|B_m \cap \{v_1, \ldots, v_n\}|$. Hence, by Lemma 8.2, for any fixed m the probability that $|B_m \cap \{v_1, \ldots, v_n\}| > b$ is at most $O(1)\exp(-C_5 b)$, for some constant $C_5 > 0$. Consequently, $\mathbf{P}[X > h] \leqslant O(1)C_4 \exp(-C_5 h/6\pi C_4)$, which completes the proof in the case $s = 2$ and $r > 10\delta$.

The case $r \leqslant 10\delta$ is easy to verify, for then the combinatorial distance between any two vertices in A is bounded, which implies that X is bounded (because $|W(v) - W(u)| \leqslant 3\pi$ for neighbors v, u).

If $s > 2$, then we may cover the annulus A with at most $2\log s + 1$ disjoint concentric annuli with radii ratio 2. In order that $X(A)$ be at least h, there must be one of these smaller annuli A' with $X(A') \geqslant h/(2\log s + 1)$. The lemma follows. ∎

LEMMA 8.4 ([Ken98a]): *Let $p, q \in \mathbb{H}$ be any two points. Consider a uniform domino tiling of the grid $\widehat{G}_\delta(\mathbb{H}) \subset \mathbb{H}$ of mesh δ, and let $p', q' \in \mathsf{V}(\widehat{G}_\delta(\mathbb{H}))$ be vertices closest to p and q, respectively. Then*

$$\lim_{\delta \to 0} \mathbf{E}[h(p')h(q')] = 8\pi^{-2} \log \left| \frac{\overline{p} - q}{p - q} \right|. \qquad \blacksquare$$

It may be noted that the right hand side is invariant under conformal automorphisms of $\mathbb{H}$, since it is the log of the square root of a cross ratio of $p, \overline{p}, q, \overline{q}$.

Let $\delta \in (0, 1/4)$, let v_0 be a vertex of the grid $\delta\mathbb{Z}^2$ which is closest to $i = \sqrt{-1}$. Fix some $r \in (0, 1/4)$. Let Q_v be as in Lemma 8.1, let γ_r^δ be the connected

component of $Q_{v_0} - B(v_0, r)$ that intersects $\partial\mathbb{H}$, and let $W_\delta(r)$ be the winding number of γ_r^δ around v_0.

PROPOSITION 8.5: *Assuming* $r < 1/4$,

$$\limsup_{\delta\to 0}\Big|\mathbf{E}\big[W_\delta(r)^2\big] - 2\log(1/r)\Big| \leqslant C_6\sqrt{\log(1/r)}, \tag{8.3}$$

where C_6 *is an absolute constant.*

Proof: Let v_1 be a vertex in $\delta\mathbb{Z}^2$ such that $v_1 - v_0 - r/3 \in [0, \delta)$. Let V_0 be the set of vertices of $\delta\mathbb{Z}^2$ whose Euclidean distance to $\{v_1, v_0\}$ is in the range $(r/9, 2r)$. Then, assuming that $\delta < r/9$, V_0 separates v_0 from v_1 and $\{v_0, v_1\}$ from $\partial\mathbb{H}$ in the grid $\delta\mathbb{Z}^2$. Given a vertex v, let Q_v^0 be the union of Q_v' with the line segment joining the intersection $Q_v' \cap \partial\mathbb{H}$ to 0. Set $Q := \bigcup\{Q_v^0 \colon v \in V_0\}$.

Set $X_j := W(Q_{v_j}, v_j)$, $j = 0, 1$. By Lemma 8.1, we have $X_j = -\pi h(v_j)/2$, and consequently, Lemma 8.4 gives

$$\lim_{\delta\to 0}\mathbf{E}[X_0X_1] = 2\log(1/r) + O(1). \tag{8.4}$$

Since V_0 separates v_0 from v_1, when conditioning on Q, X_0 becomes independent of X_1. Therefore,

$$\mathbf{E}[X_0X_1] = \mathbf{E}\big[\mathbf{E}[X_0X_1 \mid Q]\big] = \mathbf{E}\Big[\mathbf{E}[X_0 \mid Q]\cdot\mathbf{E}[X_1 \mid Q]\Big]. \tag{8.5}$$

We shall show that for small $\delta > 0$,

$$\mathbf{E}\Big[\big(W_\delta(r) - \mathbf{E}[X_j \mid Q]\big)^2\Big] = O(1), \quad j = 0, 1. \tag{8.6}$$

Using (8.5), this implies

$$\mathbf{E}[W_\delta(r)^2] - \mathbf{E}[X_0X_1] = O(1)\sqrt{\mathbf{E}[W_\delta(r)^2]} + O(1).$$

Consequently, by (8.4),

$$\limsup_{\delta\to 0}\Big|\mathbf{E}[W_\delta(r)^2] - 2\log(1/r) + O(1)\sqrt{\mathbf{E}[W_\delta(r)^2]}\Big| \leqslant O(1),$$

which implies (8.3). It therefore suffices to prove (8.6).

Let $s := \min\{|v - v_0| \colon v \in Q\}$. Given Q, let T' be a UST of $G^W(B(v_0, s/2), \delta)$, and let T'' be the UST of the graph obtained from $G_\delta(\mathbb{H})$ by identifying the vertices of Q. Note that (by Wilson's algorithm, say) T, the UST on $G_\delta(\mathbb{H})$, has the same law as $T'' \cup Q$ (as a set of edges). By the domination principle, given

Q, we may couple T'' and T' so that $\mathsf{E}(T'') \supset \mathsf{E}(T')$. Given Q, we couple T and T' so that $T \supset T'$.

Let α be the path in T' from v_0 to $\partial B(v_0, s)$, and let a be the point where α hits $\partial B(v_0, s)$. Then $\mathbf{E}[W(\alpha, v_0) \mid Q] = 0$, by symmetry, because given Q, T' is just ordinary UST on $G^W(B(v_0, s/2), \delta)$. But α is also a path in T. Let β be the path in T from a to the endpoint of γ_r^δ near $\partial B(v_0, r)$. Then $X_0 = W(\alpha, v_0) + W(\beta, v_0) + W_\delta(r)$, and therefore,

$$\begin{aligned} W_\delta(r) - \mathbf{E}[X_0 \mid Q] &= -\mathbf{E}[W(\alpha, v_0) \mid Q] - \mathbf{E}[W(\beta, v_0) \mid Q] \\ &= -\mathbf{E}[W(\beta, v_0) \mid Q]. \end{aligned} \tag{8.7}$$

Let $t > \delta$. Note that if $s < t$, then there are at least two disjoint crossings in T of the annulus $A(v_0, t, r/10)$ (both on the path Q_v^0, $v \in V_0$, which has minimum distance to v_0). Therefore, Lemma 8.2 gives

$$\mathbf{P}[s < t] \leqslant O(1)(t/r)^{C_7}. \tag{8.8}$$

Fix some $y \geqslant 2$. By Lemma 8.3, we have

$$\mathbf{P}\Big[|W(\beta, v_0)| > t,\ s \geqslant r/y\Big] \leqslant O(1)\exp\big(-C_3 t/\log y\big)\log y.$$

Hence, using (8.8),

$$\begin{aligned} \mathbf{P}\Big[|W(\beta, v_0)| > t\Big] &\leqslant \mathbf{P}\Big[s < r/y\Big] + O(1)\exp\big(-C_3 t/\log y\big)\log y \\ &\leqslant O(1)y^{-C_7} + O(1)\exp\big(-C_3 t/\log y\big)\log y. \end{aligned}$$

Assuming that $t \geqslant 1$, we may choose $y = \exp\sqrt{t}$, and then get

$$\mathbf{P}\Big[|W(\beta, v_0)| > t\Big] \leqslant O(1)\sqrt{t}\exp\big(-C_8\sqrt{t}\big),$$

for some constant $C_8 > 0$. This gives $\mathbf{E}\big[W(\beta, v_0)^2\big] = O(1)$. But for every random variable Y, we have $\mathbf{E}[Y^2] \geqslant \mathbf{E}\big[\mathbf{E}[Y \mid Q]^2\big]$. Therefore, (8.7) implies (8.6) for $j = 0$. The proof of (8.6) for $j = 1$ is entirely the same. This completes the proof of the proposition. ∎

PROPOSITION 8.6: *Assuming Conjecture 1.2, $\kappa_o = 2$, where κ_o is the constant such that SLE with parameter κ_o is the scaling limit of LERW.*

Proof: Recall the definition of $W_\delta(r)$, which appears above Proposition 8.5. Set $r_0 = 1/2$, and let $r_1 > 0$ be very small. Let ψ: $\mathbb{H} \to \mathbb{U}$ be the conformal

map satisfying $\psi(i) = 0$ and $\psi'(i) > 0$. Let C_j be the circle of radius r_j about i, $j = 0, 1$, and set $C_j' := \psi(C_j)$. Note that $Z_\delta(r_1) = W_\delta(r_1) - W_\delta(r_0)$ is the winding number around i of some arc on γ^δ, the LERW from a vertex near i to $\partial\mathbb{H}$ in $\delta\mathbb{Z}^2 \cap \mathbb{H}$, and the arc has one endpoint near the circle $\partial B(i, r_0)$ and the other endpoint near the circle $\partial B(i, r_1)$. It follows that $Z_\delta(r_1)$ converges weakly to a winding number $Z(r_1)$ of an arc β of $\psi^{-1}(\xi_{\kappa_o})$ with endpoints on C_0 and C_1, as $\delta \to 0$ along some sequence, where ξ_{κ_o} is the SLE curve with parameter κ_o. Moreover, since we have good tail estimates on $Z_\delta(r_1)$ (Lemma 8.3), from the dominated convergence theorem it follows that

$$\mathbf{E}[Z(r_1)^2] = \lim_{\delta \to 0} \mathbf{E}[Z_\delta(r_1)^2].$$

Hence, Proposition 8.5 gives $\mathbf{E}[Z(r_1)^2] = 2\log(1/r_1) + O(1)\sqrt{\log(1/r_1)}$. Observe that for any path α in $\mathbb{H} - \{i\}$, the winding number of α around i minus the winding number of $\psi(\alpha)$ around 0 is bounded by some constant. Consequently, the winding number W' of $\beta' := \psi(\beta)$ around 0 also satisfies

$$\mathbf{E}[W'^2] = 2\log(1/r_1) + O(1)\sqrt{\log(1/r_1)}. \tag{8.9}$$

Set $t_j = \log r_j$, $j = 0, 1$, let $\tilde{\beta}$ be the arc $\xi_{\kappa_o}(t)$: $[t_1, t_0] \to \mathbb{U}$, and let $\tilde{W}$ be the winding number of $\tilde{\beta}$ around 0. By Theorem 7.2, with high probability, the log of the absolute value of the endpoints of $\tilde{\beta}$ is not far from the log of the absolute value of the endpoints of β'. Therefore, it is easy to conclude with the help of Lemma 8.3 that

$$\mathbf{E}[(\tilde{W} - W')^2] = O(1). \tag{8.10}$$

We know from Theorem 7.2 again that

$$\mathbf{E}[\tilde{W}^2] = \kappa_o |t_1| + O(1)\sqrt{|t_1|}.$$

Combining this with (8.9) and (8.10) gives

$$(2 - \kappa_o)\,|t_1| = O(1)\sqrt{|t_1|}.$$

Letting $t_1 \to -\infty$ now completes the proof. ∎

Proof of Theorem 1.3: Immediate from Corollary 6.2 and Proposition 8.7. ∎

9. The critical value for the SLE

THEOREM 9.1: $\sup \mathfrak{K} \leqslant 4$, *where* $\mathfrak{K}$ *is as in Definition 7.1.*

Proof: Fix some $\kappa \in \mathfrak{K}$, and let f_t be the solution of the Löwner equation with parameter $\zeta(t) = \mathsf{B}(-\kappa t)$, where B: $[0,\infty) \to \partial\mathbb{U}$ is Brownian motion starting from a uniform point in $\partial\mathbb{U}$. Note the for every $t < 0$ the map f_t^{-1} is well defined and injective on $\partial\mathbb{U} - \{\zeta(0)\}$, since f_t is a Riemann map onto a slit domain, and the slit hits $\partial\mathbb{U}$ at $\zeta(0)$. Set $b(t) := -i \log f_t^{-1}(1)$, with $b(0) - \widehat{\mathsf{B}}(0) \in [0, 2\pi)$ and $b(t)$ continuous in t. (A.s. $\zeta(0) \neq 1$, and on this event $b(t)$ is well defined for all $t < 0$.) Then $b(t)$ is real. As in (7.5), we have

$$b'(t) = \frac{\sin\bigl(\widehat{\mathsf{B}}(-\kappa t) - b(t)\bigr)}{1 - \cos\bigl(\widehat{\mathsf{B}}(-\kappa t) - b(t)\bigr)} = \cot\frac{1}{2}\bigl(\widehat{\mathsf{B}}(-\kappa t) - b(t)\bigr). \tag{9.1}$$

Let $p(s) = b(-s) - \widehat{\mathsf{B}}(\kappa s)$, and let $\epsilon > 0$. Set $\tau_\epsilon = \inf\{s \geqslant 0\colon p(s) = \epsilon\}$ and $\tau_\pi = \inf\{s \geqslant 0\colon p(s) = \pi\}$. For $x \in [\epsilon, \pi]$, let $g_\epsilon(x)$ be the probability that $\tau_\pi < \tau_\epsilon$, conditioned on $p(0) = x$. Also set $g_\epsilon(x) = 0$ for $x < \epsilon$ and $g_\epsilon(x) = 1$ for $x > \pi$. We now show that g_ϵ satisfies

$$\frac{\kappa}{2} g_\epsilon''(x) + g_\epsilon'(x)\cot(x/2) = 0 \tag{9.2}$$

inside (ϵ, π), using Itô's formula. (The reader unfamiliar with stochastic calculus can have a look at [Dur84], for example, or try to derive (9.2) directly. The latter is a bit tricky, but can be done.) Observe that $g_\epsilon\bigl(p(s^*)\bigr)$ is a martingale, where $s^* = \min\{s, \tau_\epsilon, \tau_\pi\}$. By (9.1), we have

$$dp(s) = -b'(-s)ds - d\widehat{\mathsf{B}}(\kappa s) = \cot\bigl(p(s)/2\bigr)ds - d\widehat{\mathsf{B}}(\kappa s),$$

and therefore, by Itô's Formula (assuming, for the moment, that g_ϵ is C^2),

$$\begin{aligned} dg_\epsilon\bigl(p(s)\bigr) &= g_\epsilon'\bigl(p(s)\bigr)\Bigl(\cot\bigl(p(s)/2\bigr)\,ds - d\widehat{\mathsf{B}}(\kappa s)\Bigr) + (1/2)g_\epsilon''\bigl(p(s)\bigr)\,d\Bigl\langle\widehat{\mathsf{B}}(\kappa s)\Bigr\rangle \\ &= \Bigl(\cot\bigl(p(s)/2\bigr)g_\epsilon'\bigl(p(s)\bigr) + (\kappa/2)g_\epsilon''\bigl(p(s)\bigr)\Bigr)\,ds - g_\epsilon'\bigl(p(s)\bigr)\,d\widehat{\mathsf{B}}(\kappa s) \end{aligned}$$

for $s < \min\{\tau_\epsilon, \tau_\pi\}$. Since $g_\epsilon\bigl(p(s^*)\bigr)$ is a martingale, the ds term must vanish, and so (9.2) holds inside (ϵ, π). Consequently, in that range,

$$g_\epsilon'(x) = c_\epsilon\bigl(\sin(x/2)\bigr)^{-4/\kappa},$$

where c_ϵ is some constant depending on ϵ. Since $g_\epsilon(\epsilon) = 0$ and $g_\epsilon(\pi) = 1$, we have $\int_\epsilon^\pi g_\epsilon'(x)\,dx = 1$, which gives

$$c_\epsilon^{-1} = \int_\epsilon^\pi \bigl(\sin(x/2)\bigr)^{-4/\kappa}\,dx.$$

We know that a.s. $p(s) \neq 0$ for all s, which is equivalent to $\lim_{\epsilon \to 0} g_\epsilon(x) = 1$ on $(0, \pi)$. This gives $\lim_{\epsilon \to 0} g'_\epsilon(x) = 0$, that is, $\lim_{\epsilon \to 0} c_\epsilon = 0$. Therefore, $\kappa \leqslant 4$.

This completes the proof, except that we have not shown that g_ϵ is C^2 (there should be a reference implying this, but we have not located one). To deal with this, the above procedure is reversed. **Define** g_ϵ as the solution of (9.2) satisfying $g_\epsilon(\epsilon) = 0$ and $g_\epsilon(1) = 1$. Then the above application of Itô's Formula shows that $g_\epsilon\big(p(s^*)\big)$ is a martingale. By the Optional Sampling Theorem, this implies that $g_\epsilon(x)$ is the probability that $\tau_\pi < \tau_\epsilon$, conditioned on $p(0) = x$, and completes the proof. ∎

CONJECTURE 9.2: $\mathfrak{K} = [0, 4]$.

10. Properties of UST subsequential scaling limits in two dimensions

Before we go into the study of the UST scaling limit, let us remark that the definition we have adopted for the scaling limit is by no means the only reasonable one. There are several other reasonable variations, and choosing one is partly a matter of convenience and taste.

We now recall some definitions. Again, we think of $\delta\mathbb{Z}^2$ as a subset of the sphere $\mathbb{S}^2 = \mathbb{R}^2 \cup \{\infty\}$. Recall that $\widehat{\mathsf{T}}_\delta$ denotes the UST on $\delta\mathbb{Z}^2$, with the point ∞ added, to make it compact. Given two points $a, b \in \widehat{\mathsf{T}}_\delta$, $a \neq b$, $\omega_{a,b} = \omega^\delta_{a,b}$ denotes the unique path in $\widehat{\mathsf{T}}_\delta$ with endpoints a and b. For the case $a = b$, we set $\omega_{a,a} = \{a\}$. Let $\mathfrak{T}_\delta$ be the collection of all triplets, $(a, b, \omega_{a,b})$, where $a, b \in \widehat{\mathsf{T}}_\delta$. $\mathfrak{T}_\delta$ will be called the **paths ensemble** of $\widehat{\mathsf{T}}_\delta$. Let $\mathfrak{T}$ denote a random variable in $\mathcal{H}\big(\mathbb{S}^2 \times \mathbb{S}^2 \times \mathcal{H}(\mathbb{S}^2)\big)$ whose law is a weak subsequential limit of the law of $\mathfrak{T}_\delta$ as $\delta \to 0$. The trunk is defined by

$$\mathsf{trunk} = \mathsf{trunk}(\mathfrak{T}) := \bigcup_{(a,b,\omega) \in \mathfrak{T}} (\omega - \{a, b\}). \tag{10.1}$$

Let $\epsilon \in (0, 1]$ and $a, b \in \widehat{\mathsf{T}}_\delta$. We define $\omega_{a,b}(\epsilon)$ as follows. Let a' be the first point along the path $\omega_{a,b}$ (which is oriented from a to b) where $d(a, a') = \epsilon$, and let b' be the last point along the path where $d(b, b') = \epsilon$, provided that such points exist. If a' and b' exist, and a' appears on the path before b', then let $\omega_{a,b}(\epsilon)$ be the (closed) subarc of $\omega_{a,b}$ from a' to b'; and otherwise set $\omega_{a,b}(\epsilon) = \emptyset$. Let $\mathfrak{T}_\delta(\epsilon)$ denote the set of all triplets $\big(a, b, \omega_{a,b}(\epsilon)\big)$ such that $a, b \in \widehat{\mathsf{T}}_\delta$ and $\omega_{a,b}(\epsilon) \neq \emptyset$. Note that if $d(a, b) > 2\epsilon$, then $\omega_{a,b}(\epsilon) \neq \emptyset$. We define

$$\mathsf{trunk}_\delta(\epsilon) := \bigcup \Big\{ \omega \colon (a, b, \omega) \in \mathfrak{T}_\delta(\epsilon) \Big\}.$$

Then $\mathsf{trunk}_\delta(\epsilon)$ is a compact subset of $\widehat{\mathsf{T}}_\delta$, which we call the **$\epsilon$-trunk** of $\widehat{\mathsf{T}}_\delta$. By compactness, for every $\epsilon \in (0,1]$ there is a subsequential scaling limit of the law of $\mathsf{trunk}_\delta(\epsilon)$. By passing to a subsequence, if necessary, we assume that for all $n \in \mathbb{N}_+ := \{1,2,\ldots\}$ the weak limit $\mathsf{trunk}_0(1/n)$ of $\mathsf{trunk}_\delta(1/n)$, as $\delta \to 0$, exists.

Recall that the **dual** $T^\dagger$ of a spanning tree $T \subset \delta\mathbb{Z}^2$ is the spanning subgraph of the dual graph $(\delta\mathbb{Z}^2)^\dagger := (\delta/2,\delta/2) + \delta\mathbb{Z}^2$ containing all edges that do not intersect edges in T. If T is the UST on $\delta\mathbb{Z}^2$, then $T^\dagger$ has the law of the UST on $(\delta\mathbb{Z}^2)^\dagger$. (See, e.g., [BLPS98].) Let $\widehat{\mathsf{T}}^\dagger_\delta$ be $T^\dagger \cup \{\infty\}$, where T is the UST on $\delta\mathbb{Z}^2$. $\mathsf{trunk}^\dagger$ and $\mathsf{trunk}^\dagger_0(\epsilon)$ are defined for $\widehat{\mathsf{T}}^\dagger_\delta$ as trunk and $\mathsf{trunk}_0(\epsilon)$ were defined for $\widehat{\mathsf{T}}_\delta$.

We may think of the random variables trunk, $\mathsf{trunk}^\dagger$, $\mathsf{trunk}_0(1/n)$, and $\mathsf{trunk}^\dagger_0(1/n)$ $(n \in \mathbb{N}_+)$ as defined on the same probability space, by taking a subsequential limit of the joint distribution of $\mathfrak{T}_\delta$, $\mathfrak{T}^\dagger_\delta$, $\langle \mathsf{trunk}_\delta(1/n)\colon n \in \mathbb{N}_+\rangle$, and $\langle \mathsf{trunk}^\dagger_\delta(1/n)\colon n \in \mathbb{N}_+\rangle$. It is immediate to verify that a.s.

$$\mathsf{trunk} = \bigcup_{n\in\mathbb{N}_+} \mathsf{trunk}_0(1/n),$$

and $\mathsf{trunk}_0\big(1/(n+1)\big) \supset \mathsf{trunk}_0(1/n)$ for $n \in \mathbb{N}_+$.

We shall prove that trunk is a.s. a topological tree, in the sense of the following definition.

Definition 10.1: Trees. An **arc** joining two points x, y in a metric space X is a set $J \subset X$ such that there is a homeomorphism $\phi\colon [0,1] \to J$ with $\phi(0) = x$ and $\phi(1) = y$. A metric space X will be called a **topological tree** if it is uniquely arcwise connected (that is, given $x \neq y$ in X there is a unique arc in X joining x and y) and locally arcwise connected (that is, whenever $x \in U$ and U is an open subset of X there is an open $W \subset U$ with $x \in W$ and W is arcwise connected). A **finite topological tree** is a topological space which is homeomorphic to a finite, connected, simply connected, 1-dimensional simplicial complex.

Note that a connected subset of a topological tree is a topological tree [Bow].

Although we shall not need this fact, it is instructive to note that a metric space which is a topological tree is homeomorphic to an $\mathbb{R}$-tree[4] [MO90] (see also [MMOT92], for a slightly less general but simpler proof).

The next theorem establishes a finiteness property of the ϵ-trunks, which is the first step in the proof of Theorem 1.6.

4 An $\mathbb{R}$-tree is a metric space (T,d) such that for every two distinct points $x, y \in T$ there is a unique isometry ϕ from $[0, d(x,y)]$ onto a subset of T satisfying $\phi(0) = x$ and $\phi(d(x,y)) = y$.

THEOREM 10.2 (Finiteness): *For every $\epsilon > 0$ there is a $\widehat{\delta} > 0$ with the following property. Suppose that $0 < \delta < \widehat{\delta}$. Let $\widehat{V}$ be a set of vertices of $\delta\mathbb{Z}^2$ such that every point in $\mathbb{S}^2$ is within distance $\widehat{\delta}$ of some vertex in $\widehat{V}$. Let $Q := Q_\delta(\widehat{V})$ be the subtree of $\widehat{\mathsf{T}}_\delta$ that is spanned by $\widehat{V}$, that is, the minimal connected subset of $\widehat{\mathsf{T}}_\delta$ containing $\widehat{V}$. Then with probability at least $1 - \epsilon$ we have $Q \supset \mathsf{trunk}_\delta(\epsilon)$.*

Proof: Fix some small $\widehat{\delta} > 0$, and suppose that $\delta \in (0, \widehat{\delta})$. Let $V_0 := \widehat{V}$, and for each $j \in \mathbb{N}_+$ let V_j be a set of vertices containing V_{j-1} such that every vertex of $\delta\mathbb{Z}^2$ is within spherical distance $\delta_j := 2^{-j}\widehat{\delta}$ of some vertex in V_j, and V_j is a minimal set satisfying these properties. Note that the number of vertices in $V_j - V_{j-1}$ is bounded by $O(1)\delta_j^{-2}$. Let Q_j be the subtree of $\widehat{\mathsf{T}}_\delta$ spanned by V_j.

We now estimate the probability that there is some component of $Q_{j+1} - Q_j$ whose diameter is large. Let v be some vertex in $\delta\mathbb{Z}^2$, let $Q(v, j)$ be the arc of $\widehat{\mathsf{T}}_\delta$ that connects v to Q_j, and let $\mathcal{D}(v, j, a)$ be the event the diameter of $Q(v, j)$ is at least $a\delta_j$. By Wilson's algorithm, we may obtain $Q(v, j)$ by conditioning on Q_j and loop-erasing a simple random walk from v that stops when Q_j is hit. Every vertex $w \in \delta\mathbb{Z}^2$ is within distance δ_j from a vertex in Q_j. Since Q_j is connected and has diameter at least 1, Lemma 2.1 shows that there is a universal constant $C_0 > 0$ so that the probability that a random walk from w gets to distance $C_0\delta_j$ from w before hitting Q_j is at most $1/2$. Consequently, $\mathcal{D}(v, j, a)$ has probability at most $O(1)\exp(-C_1 a)$, where $C_1 > 0$ is an absolute constant. We choose $a_j := j^2(\log\widehat{\delta})^2/C_1$. Since there are at most $O(1)\delta_j^{-2}$ vertices in $V_{j+1} - V_j$, we find that the probability of

$$\mathcal{D} := \bigcup_{j=1}^{\infty} \bigcup_{v \in V_{j+1}} \mathcal{D}(v, j, a_j)$$

is bounded by

$$O(1)\widehat{\delta}^{-2} \sum_{j=1}^{\infty} 2^{2j} \exp\bigl(-j^2(\log\widehat{\delta})^2\bigr),$$

which goes to zero as $\widehat{\delta} \to 0$.

Let $v \in \delta\mathbb{Z}^2$. There is a sequence $v_1, v_2, \ldots, v_n$ with $v_j \in V_j$ such that $Q(v, 1) \subset \bigcup_{j=2}^{n} Q(v_j, j-1)$, and the latter union is connected. If we are in the complement of $\mathcal{D}$, it follows that the diameter of $Q(v, 1)$ is at most $s := \sum_{j=1}^{\infty} a_j\delta_j$. Since $s \to 0$ as $\delta_0 \to 0$, this establishes the theorem. ∎

Several corollaries follow from this theorem.

COROLLARY 10.3: *For each $n \in \mathbb{N}_+$, a.s. $\mathsf{trunk}_0(1/n)$ is a finite topological tree.*

Proof: Let $W \subset \mathbb{R}^2$ be finite. For each $w \in W$ and $\delta > 0$, let $w_\delta \in \delta\mathbb{Z}^2$ be closest to w, with ties broken arbitrarily, and set $W_\delta = \{w_\delta : w \in W\}$. Let $Q_\delta(W)$ be the subtree of $\widehat{\mathsf{T}}_\delta$ spanned by W_δ. The theorem shows that we may choose a finite $W \subset \mathbb{R}^2$ such that $\mathsf{trunk}_\delta(1/n) \subset Q_\delta(W)$ with probability at least $1-\epsilon$, for every sufficiently small $\delta > 0$. Consequently, we may couple a subsequential scaling limit $Q(W)$ of $Q_\delta(W)$ as $\delta \to 0$ so that $\mathsf{trunk}_0(1/n) \subset Q(W)$ with probability at least $1-\epsilon$. Because $\mathsf{trunk}_0(1/n)$ is connected and ϵ is an arbitrary positive number, it suffices to prove that $Q(W)$ is a.s. a finite tree. The latter is easily proved by induction on $|W|$ using Theorem 1.1, Wilson's algorithm, and the following easy fact: the tree spanned by a subset of the points in W is unlikely to pass close by to the other points. (See Remark 3.2.) ∎

COROLLARY 10.4: *The Hausdorff dimension of* trunk *is in* $(1,2)$. *Moreover, if* $I = (s_0, s_1]$ *is an interval such that a.s. the Hausdorff dimension of any scaling limit of LERW is in* I, *then the Hausdorff dimension of* $\mathsf{trunk}(\mathfrak{T})$ *is in* I.

Proof: The second statement follows immediately from Theorem 10.2. The first is now a consequence of the result of [ABNW], showing that there are $s_0, s_1 \in (1,2)$ such that a.s. the Hausdorff dimension of LERW scaling limit is in $[s_0, s_1]$.[5] ∎

Remark 10.5: The above-mentioned lower bound in [ABNW] is based on the ideas of [BJPP97]. Kenyon [Ken] can prove that we may take $s_1 = 5/4$. In earlier work [Ken98b] he showed that $n^{5/4}$ times the expected number of edges in a LERW from $(0,0)$ to the boundary of the square $[-n,n]^2$ tends to a finite positive constant as $n \to \infty$. This supports the conjecture that the Hausdorff dimension of the scaling limit of LERW is a.s. $5/4$, and the same would apply to $\mathsf{trunk}(\mathfrak{T})$.

The **degree** of a point p in a topological tree T is the number of connected components of $T - \{p\}$. The following corollary is a strong form of the statement that the maximum degree of points in $\mathsf{trunk}_0(1/n)$ is 3. From this and the fact that trunk is a tree (which we prove further below) it immediately follows that the

5 From Remark 3.2 follows the weaker result that the area measure of any subsequential scaling limit of LERW is zero, hence that the area of trunk is zero. It is likely that with a bit more effort the proof of Remark 3.2 is sufficient for the stronger claim that the Hausdorff dimension is smaller than 2.

maximum degree in trunk is 3, because every finite subset of trunk is contained in some $\mathsf{trunk}_0(1/n)$.

Given a point $p \in \mathbb{S}^2$ and two numbers $0 < r_1 < r_2 < 1$, let $A_{\mathrm{sp}}(p, r_1, r_2)$ denote the annulus with center p, inner radius r_1, and outer radius r_2, in the spherical metric.

COROLLARY 10.6: *Given every $\epsilon \in (0,1)$, there is an $r \in (0,\epsilon)$ with the following property. For every sufficiently small $\delta > 0$, the probability that there is a point $p \in \mathbb{S}^2$ such that there are 4 disjoint crossings in $\mathsf{trunk}_\delta(\epsilon)$ of the annulus $A_{\mathrm{sp}}(p, r, \epsilon)$ is at most ϵ.*

By having 4 disjoint crossings in $\mathsf{trunk}_\delta(\epsilon)$ of an annulus A, we mean that there are 4 disjoint connected subsets of $\mathsf{trunk}_\delta(\epsilon) \cap A$ that intersect both boundary components of A. Below, Corollary 10.11 gives a strengthening of Corollary 10.16.

Proof: By Theorem 10.2, it is enough to prove the statement with $Q_\delta(W)$ replacing $\mathsf{trunk}_\delta(\epsilon)$, where $W \subset \mathbb{R}^2$ is a set of bounded size, provided that the value of r does not depend on δ. Again, induction on $|W|$ can be used together with Wilson's algorithm. One needs note the following easy facts. The tree T_{k-1} spanned by $k-1$ points of W is unlikely to pass close to the other points of W, and when adding a further point, it is unlikely that the attachment point of the new branch on T_{k-1} will be close to another branch point. Also, once a random walk from the new point gets close to T_{k-1} it will hit T_{k-1} close by, with high likelihood. The easy details are left to the reader. ∎

We now turn to the central issue in the proof of Theorem 1.6, which is,

THEOREM 10.7: *In any subsequential scaling limit of UST in $\mathbb{S}^2$, a.s. the trunk and dual trunk do not intersect.*

LEMMA 10.8: *Given $\delta, \epsilon > 0$, let $T^3_\delta(\epsilon)$ be the set of points of degree 3 in $\mathsf{trunk}_\delta(\epsilon)$. Let $\mathsf{trunk}^\dagger_\delta(\epsilon)$ be the ϵ-trunk of the dual tree $\widehat{\mathsf{T}}^\dagger_\delta$. Let D be the spherical distance from $T^3_\delta(\epsilon)$ to $\mathsf{trunk}^\dagger_\delta(\epsilon)$, that is, the least spherical distance between a point in $T^3_\delta(\epsilon)$ to a point in $\mathsf{trunk}^\dagger_\delta(\epsilon)$. Then $\lim_{t\to 0} \mathbf{P}[D < t] \to 0$ uniformly in δ.*

The following simple observation is used in the proof. Suppose that we condition on a set of edges S to appear in the UST tree in a planar graph. For the dual tree, this is the same as deleting the edges dual to the edges in S. Consequently, one can perform a variation on Wilson's algorithm for a planar graph, where one switches back and forth from building the tree by adding LERW branches

and building the dual tree. When building the tree, the LERW acts with the constructed tree as a wired absorbing boundary and the constructed dual tree as a free boundary, and conversely when building the dual tree.

Proof: We first choose a large but finite collection of points Q in $\delta\mathbb{Z}^2$ so that with high probability the subtree T of $\widehat{\mathsf{T}}_\delta$ spanned by Q contains $\mathsf{trunk}_\delta(\epsilon)$ (and $|Q|$ does not depend on δ). This can be done, by Theorem 10.2. Let $Q^\dagger$ be a set of vertices of the dual graph $(\delta\mathbb{Z}^2)^\dagger$, such that with high probability the subtree $T^\dagger$ spanned by $Q^\dagger$ in the dual graph contains $\mathsf{trunk}^\dagger_\delta(\epsilon)$. Let $a_1, a_2, a_3 \in Q$ and $a_1^\dagger, a_2^\dagger \in Q^\dagger$ be distinct points. It suffices to show that the probability that the arc β joining $a_1^\dagger$ and $a_2^\dagger$ in $T^\dagger$ comes within distance t of the meeting point m of a_1, a_2 and a_3 in T goes to zero as $t \to 0$, uniformly in δ. This is easy. We condition on the subtree T_0 of T spanned by a_1, a_2, a_3. Let $z^\dagger$ be a dual vertex close to $a_1^\dagger$. Then with high probability the dual tree path β from $z^\dagger$ to $a_1^\dagger$ has diameter not much larger than the distance from $z^\dagger$ to $a_1^\dagger$. In particular, it does not go close to T_0. By the next lemma, conditioned on β and T_0, the probability that a simple random walk starting at $a_2^\dagger$, with T_0 acting as a reflecting boundary, will get to within distance t of m before hitting β tends to zero with t. Consequently, the same is true for the loop-erasure of this walk, which can be taken as the path joining β and $a_2^\dagger$ in the dual tree. ∎

A simple random walk on $\delta\mathbb{Z}^2$ is likely to hit a connected subgraph of fixed diameter> 0 before visiting a specified vertex at fixed positive distance away. The next lemma gives a uniform version of this statement, in a slightly more general setting, where there is also a reflecting boundary.

LEMMA 10.9: *Let D be a domain in $\mathbb{S}^2$ such that $\mathbb{S}^2 - D$ has two connected components, B_1, B_2, and assume that neither is a single point. Consider a sequence δ_j, $j \in \mathbb{N}$, of positive numbers tending to zero. Suppose that to each $j \in \mathbb{N}$ there are two connected subgraphs B_1^j, B_2^j of the grid $\delta_j\mathbb{Z}^2$, and that $B_1^j \to B_1$ and $B_2^j \to B_2$ in the Hausdorff metric on compact subsets of $\mathbb{S}^2$. Let $m \in B_2$ be some point, and for each $t > 0$ and $z \in \delta_j\mathbb{Z}^2 - B_2^j$, let $h_j(z,t)$ be the probability that simple random walk on $\delta_j\mathbb{Z}^2 - B_2^j$ starting at z (with reflecting boundary conditions on B_2^j) will get to within distance t of m before hitting B_1^j. Let $K \subset \mathbb{S}^2 - B_2$ be compact. Then*

$$\lim_{\substack{t\to 0\\ j\to\infty}} \sup\{h_j(z,t)\colon z \in K \cap \delta_j\mathbb{Z}^2\} = 0.$$

Proof: Set $G_j := \delta_j\mathbb{Z}^2 - B_2^j$, let $S(t)$ be the vertices of G_j that are within distance t from m, and let $V_j(t)$ be the set of vertices of $G_j - B_1^j - S(t)$. Then

$h_j(z,t)$ is discrete-harmonic in $V_j(t)$. Recall that the Dirichlet energy of $h_j(z,t)$ is $\sum\big(h_j(z,t) - h_j(z',t)\big)^2$, with the sum extending over all edges $[z,z']$ in G_j. Let $\lambda = \lambda_j$ be the minimum of $h_j(z,t)$ on K, and let z be where the minimum is achieved. Then there is a path β from z to $S(t)$ such that $h_j(z,t) \geqslant \lambda$ on β, by the maximum principle for discrete harmonic functions. Note that one can find a collection of at least $1/(c\delta_j)$ disjoint paths in G_j which join B_1^j and β and each path in the collection has combinatorial length bounded by c/δ_j, where $c < \infty$ depends only on K and D. The Dirichlet energy of $h_j(z,t)$ restricted to each such path is at least λ/δ_j times a positive constant, and therefore the Dirichlet energy of $h_j(z,t)$ is at least $C\lambda$, where $C > 0$ is a constant depending only on K and D.

Let d_j be the distance from m to B_1^j. Since $h_j(z,t)$ is harmonic in $V_j(t)$, it minimizes the Dirichlet energy among functions on G_j that are 1 on $S(t)$ and 0 on B_1^j. Therefore, the Dirichlet energy of $h_j(z,t)$ is at most the Dirichlet energy of the function $f\colon G_j \to \mathbb{R}$, which is 1 on $S(t)$, 0 outside of $S(d_j)$, and equal to $\log\big(d_j/|z-m|\big)/\log(d_j/t)$ elsewhere, which is $O(1)/\log(d_j/t)$, as $j \to \infty$. This gives, $\lambda_j = O(1)/\log(d_j/t)$, and the lemma follows. ∎

Proof of Theorem 10.7: Before we go into the actual details, the overall plan of the proof will be given (in a somewhat imprecise manner). Let $t_0 > 0$. It is not hard to reduce the theorem to the claim that with probability close to 1 the path $\gamma \subset \widehat{\mathsf{T}}_\delta$ which joins two fixed points a_1, a_2 does not have points p close to it such that the path $\alpha_p \subset \widehat{\mathsf{T}}_\delta$ joining p to γ is not contained in a small neighborhood of γ. Let Z be the set of points p such that α_p does not stay close to γ. When we condition on γ, the probability that $p \in Z$ goes to zero as p tends to a point in γ, by a simple harmonic measure estimate. However, this is not enough, since there are many different p's close to γ. We fix some collection L_1 of points close to γ, and take a thick collection of points L_2 which are much closer to γ. What we show is that conditioned on γ and on $Z \cap L_2 \neq \emptyset$, the expectation of $N := |Z \cap L_1|$ is much larger than $\mathbf{E}[N \mid \gamma]$. This is established by observing that when $p \in L_1 \cap Z$ is appropriately chosen, the expected number of points $p' \in L_2$ such that $\alpha_{p'}$ is contained in α_p, except for a small initial segment of $\alpha_{p'}$, is quite large. It follows that

$$\mathbf{P}[Z \cap L_2 \neq \emptyset \mid \gamma] \leqslant \frac{\mathbf{E}[N \mid \gamma]}{\mathbf{E}[N \mid \gamma,\ Z \cap L_1 \neq \emptyset]}$$

is small, which suffices to prove the theorem.

We now give the details. Fix four distinct points $a_1, a_2, b_1, b_2 \in \mathbb{R}^2$. Given $\delta > 0$, let a_1' and a_2' be points of $\delta\mathbb{Z}^2$ that are closest to a_1 and a_2, respectively,

and let b_1' and b_2' be vertices of the grid dual to $\delta\mathbb{Z}^2$ that are closest to b_1 and b_2, respectively. Let γ be the path in $\widehat{\mathsf{T}}_\delta$ that joins a_1' and a_2', and given any $p \in \delta\mathbb{Z}^2$, let α_p denote the path in $\widehat{\mathsf{T}}_\delta$ from p to γ. Let β be the path of $\widehat{\mathsf{T}}_\delta^\dagger$ that joins b_1' and b_2'.

Since $\mathsf{trunk} = \bigcup_n \mathsf{trunk}_0(1/n)$, and $\mathsf{trunk}^\dagger = \bigcup_n \mathsf{trunk}_0^\dagger(1/n)$, Theorem 10.2 shows that it suffices to prove that the probability that the distance between γ and β is less than t goes to zero, as t goes down to zero, uniformly in δ. We know that with probability close to one, β does not come close to $\{a_1, a_2\}$, and γ does not come close to $\{b_1, b_2\}$ (Remark 3.2). Therefore, we need only consider the situation where there is a point q on γ, which is close to β, but not close to $\{a_1, a_2, b_1, b_2\}$. Since β and γ cannot cross, and since β locally separates the sphere near every point of $\beta - \{b_1', b_2'\}$, such a situation implies that there is a point q' in $\delta\mathbb{Z}^2$, which is near q, but in order to get to γ from q' one must either cross β, or go "around" it. Consequently, $\operatorname{diam}(\alpha_{q'})$ must be bounded away from zero, as $\alpha_{q'}$ cannot cross β. It therefore suffices to rule out the existence of a point $q' \in \delta\mathbb{Z}^2$ close to γ but with $\operatorname{diam}(\alpha_{q'})$ bounded away from zero. More precisely, let K be a compact set disjoint from $\{a_1, a_2\}$, and let $\epsilon_1 \in (0,1)$. Let $\tilde{h}$ be the least distance from γ to some point $q' \in K \cap \delta\mathbb{Z}^2$ such that $\operatorname{diam}(\alpha_{q'}) \geqslant \epsilon_1$. It suffices to show that

$$\inf_{h_0>0} \limsup_{\delta\to 0} \mathbf{P}[\tilde{h} < h_0] = 0. \tag{10.2}$$

Given any $p \in \delta\mathbb{Z}^2$, let $h(p)$ be the distance from p to γ and let $k(p)$ be the maximal distance from a point on α_p to γ. By Corollary 10.6, the probability that there is an arc α in $\widehat{\mathsf{T}}_\delta$, which is disjoint from γ, satisfies $\operatorname{diam}(\alpha) \geqslant \epsilon_1$, and every point of α is within distance t of γ, goes to zero as $t \to 0$, uniformly in δ. Hence, to prove (10.2), it suffices to establish that

$$\forall t > 0 \quad \inf_{h_0>0} \limsup_{\delta\to 0} \mathbf{P}[\exists p \in K \cap \delta\mathbb{Z}^2 \ h(p) < h_0,\ k(p) \geqslant t] = 0. \tag{10.3}$$

Since the proof is somewhat involved, we consider first the simpler situation in which

$$\gamma = [0,1] \times \{0\}, \text{ and } K = [1/3, 2/3] \times [0,1] \tag{10.4}$$

(notwithstanding that this is an unrealistic situation, of extremely low probability). Obviously, it suffices to prove (10.3) for small $t > 0$.

In the following arguments, several small positive quantities appear. Their dependence differs from the natural flow of the proof. In order to make it clear

that the proof is logically sound, we state now that the dependence order is as follows:

$$\epsilon_0, t_0, r_0, t_1, h_1, r_1, h_2, \delta\,;$$

that is, each of these quantities may depend only on those appearing before it in the list, and should be thought of as much smaller than its predecessors.

Set $h_1' := \max\{k\delta\colon k \in \mathbb{Z},\ k\delta \leqslant h_1\}$, and let

$$L_1 := \bigl\{(k\delta, h_1')\colon k \in \mathbb{Z},\ 1/6 < k\delta < 5/6\bigr\}.$$

Given γ and $p \in \delta\mathbb{Z}^2$, we may choose α_p by loop-erasing a simple random walk from p to γ. Consequently, an easy harmonic measure estimate shows that $\mathbf{P}[k(p) > t] = O(h_1/t)$ for all $p \in L_1$.

Set $h_2' := \max\{k\delta\colon k \in \mathbb{Z},\ k\delta \leqslant h_2\}$, and

$$L_2 := \bigl\{(k\delta, h_2')\colon k \in \mathbb{Z},\ 1/4 < k\delta < 3/4\bigr\}.$$

Again, for all $p \in L_2$, $\mathbf{P}[k(p) \geqslant t_0] = O(h_2/t_0)$, so we may assume that the event $\mathcal{Q}$ that the leftmost point in L_2 satisfies $k(p) < t_0$ has probability at least $1 - \epsilon_0$. Let $\mathcal{K}$ be the event that there is some $p \in L_2$ with $k(p) \geqslant t_0$, and let $\mathcal{K}' := \mathcal{K} \cap \mathcal{Q}$. For proving (10.3) in the simpler situation (10.4), it suffices to show that $\mathbf{P}[\mathcal{K}'] = O(\epsilon_0)$ for all sufficiently small $\delta > 0$.

Consider the following procedure for generating $\widehat{\mathsf{T}}_\delta$ given γ. Perform Wilson's algorithm starting with the vertices in L_2, in left-to-right order. If we encounter in this procedure some vertex $p \in L_2$ such that $k(p) \geqslant t_0$, we stop, and let p_0 denote that vertex. Let T_0 be the tree constructed up to that point (including α_{p_0}). On the event $\mathcal{K}'$, let p_1 be the first point on α_{p_0} whose distance to γ is at least t_0, and let α be the arc of α_{p_0} from p_0 to p_1. Let $\mathcal{A}_1$ be the event that α is not contained in the rectangle $[1/5, 4/5] \times [0, t_0]$. Note that $\mathcal{A}_1$ implies that there is an arc in $\alpha \cap ([1/5, 4/5] \times [0, t_0])$ with diameter at least $1/15$. By considering this arc and γ, Corollary 10.6 shows that $\mathbf{P}[\mathcal{A}_1 \cap \mathcal{K}'] < \epsilon_0$, assuming that t_0 is sufficiently small.

On the event $\mathcal{K}' - \mathcal{A}_1$, let

$$x_1 := \max\bigl\{x \in \mathbb{R}\colon (x, h_1/2) \in \alpha\bigr\},$$

let U be the component of $\mathbb{R}^2 - \Bigl(\alpha \cup \bigl([x_1, \infty) \times \{h_1/2\}\bigr) \cup \bigl(\mathbb{R} \times \{t_1\}\bigr)\Bigr)$ that contains $(x_1, \infty) \times \{h_1/2\}$, and let U' be the set of points in U that are within distance t_1 from α. See Figure 10.1. Let $\mathcal{A}_2$ be the event that $\mathcal{K}' - \mathcal{A}_1$ occurs and U' intersects T_0.

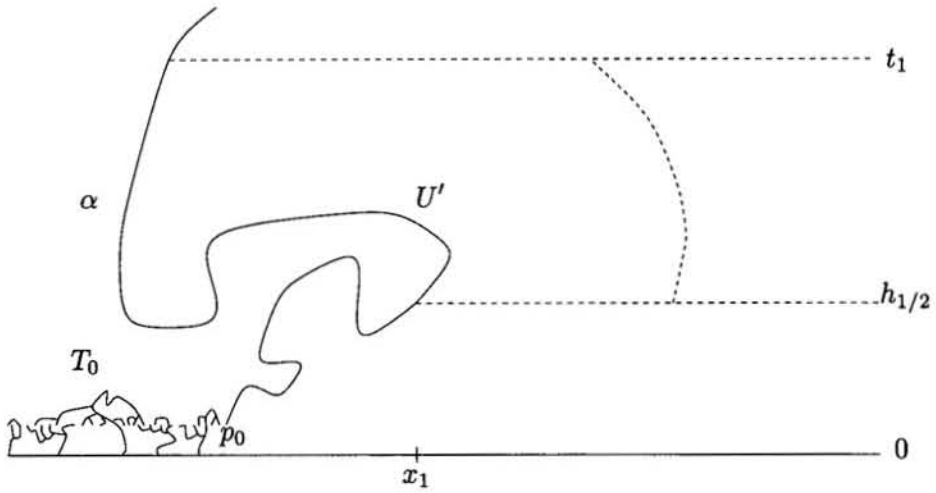


Figure 10.1.

We now prove that $\mathbf{P}[\mathcal{K} - \mathcal{A}_1 - \mathcal{A}_2] = O(\epsilon_0)$. Let N be the number of points $p \in L_1$ such that $k(p) \geqslant t_0$. For a given $p \in L_1$, the probability of $k(p) \geqslant t_0$ (given (10.4), but otherwise unconditioned) is $O(h_1/t_0)$. Therefore,

$$\mathbf{E}[N] \leqslant O(1)h_1/(\delta t_0). \tag{10.5}$$

On the other hand, condition on the event $\mathcal{K}' - \mathcal{A}_1 - \mathcal{A}_2$ and on T_0. Let L_1' be the set of $p \in L_1 \cap U$ such that the distance from p to α is at most $t_1/2$. Note that conditioned on $p \in L_1'$, the probability that α_p joins with α_{p_0} within distance $2t_1$ from p is at least

$$O(1)^{-1}\frac{h_1}{\operatorname{dist}(p,\alpha) + h_1}, \tag{10.6}$$

since after generating T_0, we may continue by running Wilson's algorithm starting at p, and the probability that the random walk starting at p will hit α before the ray $(x_1, \infty) \times \{h_1/2\}$ is at least (10.6). It therefore follows that conditioned on $\mathcal{K} - \mathcal{A}_1 - \mathcal{A}_2$, we have

$$\begin{aligned}\mathbf{E}[N \mid \mathcal{K} - \mathcal{A}_1 - \mathcal{A}_2] &\geqslant O(1)^{-1}\sum\{h_1/(k\delta)\colon k \in \mathbb{Z},\ h_1 \leqslant k\delta \leqslant t_1/2\} \\ &\geqslant O(1)^{-1}\delta^{-1}h_1\log\bigl(t_1/(2h_1)\bigr).\end{aligned}$$

Combining this with (10.5) gives

$$\mathbf{P}[\mathcal{K} - \mathcal{A}_1 - \mathcal{A}_2] \leqslant \frac{\mathbf{E}[N]}{\mathbf{E}[N \mid \mathcal{K} - \mathcal{A}_1 - \mathcal{A}_2]} \leqslant O(1)\Bigl(t_0\log\bigl(t_1/(2h_1)\bigr)\Bigr)^{-1} \leqslant \epsilon_0,$$

provided that h_1 is sufficiently small.

It remains to establish that $\mathbf{P}[\mathcal{A}_2] \leqslant O(\epsilon_0)$. First consider the case that there is some $p \in L_2$, to the left of p_0, such that $\alpha_p \cap U \neq \emptyset$. Then this must be the case for p being the left neighbor of p_0, namely $p = p_0' := p_0 - (\delta, 0)$, because when $p \in L_2$ is to the left of p_0', the path α_p cannot cross $\alpha_{p_0} \cup [p_0, p_0'] \cup \alpha_{p_0'}$, and α_p does not get to $\mathbb{R} \times \{t_0\}$. If $\alpha_{p_0'}$ intersects U, then $\alpha_{p_0'}$ must first get to some point z in $[x_1, \infty) \times \{h_1/2\}$. Near z there must be two points $z_1^\dagger, z_2^\dagger$, which are vertices of the dual grid $(\delta\mathbb{Z}^2)^\dagger$, and are locally separated from each other by α_{p_0}. The path in the dual tree that joins $z_1^\dagger$ and $z_2^\dagger$ has to contain the edge dual to the edge $[p_0', p_0]$. Now consider another dual vertex $z_3^\dagger$ just left of α near the point $(x_1, h_1/2)$. Let $m^\dagger$ be the meeting point of $z_1^\dagger, z_2^\dagger$ and $z_3^\dagger$ in the dual tree. If $m^\dagger$ is not within distance r_1 of p_0, we get in the $r_1/10$ trunk of the dual tree at least four disjoint crossings of the annulus $A(p_0, 2h_2, r_1/2)$. We may assume that this has probability $\leqslant \epsilon_0$, by Corollary 10.6. Similarly, Lemma 10.8, with the role of the tree and dual tree reversed, shows that we may take the event that $m^\dagger$ is within distance r_1 of p_0 to have probability $\leqslant \epsilon_0$, provided that r_1 is sufficiently small when compared with h_1.

To establish that $\mathbf{P}[\mathcal{A}_2] \leqslant O(\epsilon_0)$, it now suffices to prove that on the event $\mathcal{K}' - \mathcal{A}_1$, the probability that α_{p_0} intersects U' is $O(\epsilon_0)$. The argument is similar here, but occurs on a larger scale. Suppose that $w \in \alpha_{p_0} \cap U'$. Then w must be on the segment of α_{p_0} from p_1 to the point p_2 in $\alpha_{p_0} \cap \gamma$. Because $w \in \alpha_{p_0} - \alpha$ and w is within distance t_1 to α, there is a point near w that is in the $(t_0/2)$-trunk of the dual tree, namely, some point on the path connecting dual two vertices on opposite sides of α_{p_0} near p_1. Consequently, by Lemma 10.8, we may rule out the possibility that p_2 is within distance r_0 of w as having small probability, since p_2 is a point of degree 3 of the $(t_0/2)$-trunk. But if the distance between p_2 and w is more than r_0, then there are in the $(t_0/2)$-trunk at least four disjoint crossings of the annulus $A(w, 2t_1, r_0)$: two on α_{p_0} and two on γ. An appeal to Corollary 10.6 now establishes $\mathbf{P}[\mathcal{A}_2] \leqslant O(\epsilon_0)$. This completes the proof in the situation (10.4).

We now explain how to modify the above proof to deal with the general case. First note that the restriction on K is entirely inconsequential; we could in the same way deal with any compact set disjoint from the endpoints of γ. More significant is the special selection of γ. Observe that under the assumption of conformal invariance of the LERW scaling limit, the general case can be reduced to the case where $\gamma = [0, 1] \times \{0\}$, because after γ is generated, the rest of the UST is just unconditioned UST on the complement of γ with wired boundary

conditions. We may then transform γ by a conformal homeomorphism to $[0,1] \times \{0\}$, and refer to the above result.

Although we do not assume conformal invariance of the scaling limit, it turns out that the proof above is itself conformally invariant. With some care, one can apply the conformal map to the proof, in a manner of speaking. This is actually not very surprising, because the proof is ultimately based on a simple (discrete) harmonic measure estimate, which is conformally invariant.

Let us turn to the details. We may couple the UST for a subsequence of δ tending to zero so that γ tends to some path γ_0 as $\delta \to 0$ along that subsequence (see the discussion of the Prohorov metric in Section 2). Let

$$f_\delta \colon \mathbb{S}^2 - ([0,1] \times \{0\}) \to \mathbb{S}^2 - \gamma$$

be the conformal map normalized to take the endpoints of $[0,1] \times \{0\}$ to the endpoints of γ and so that $f_\delta(\infty)$ is on the line which is the set of points at equal distance from both endpoints of γ, say. (The latter normalization is necessary to make f_δ unique, but otherwise, it is quite arbitrary.) It follows that f_δ tends to the similarly normalized conformal map $f \colon \mathbb{S}^2 - ([0,1] \times \{0\}) \to \mathbb{S}^2 - \gamma_0$. We may assume that δ is so small that f and f_δ are very close on compact subsets disjoint from $[0,1] \times \{0\}$.

For each $p \in L_1$, where L_1 is as before, we let $\widehat{p}$ denote a point in $\delta\mathbb{Z}^2$ that is closest to $f(p)$. For the general case, we consider $\widehat{N}$, the number of $p \in L_1$ such that $k(\widehat{p})$ is not small, in place of N. Let $\widehat{L}_2$ denote the set of points in $\delta\mathbb{Z}^2$ that are within distance 3δ of $f\big([1/4,3/4] \times \{h_2\}\big)$. The proof for the general case uses $\widehat{L}_2$ in place of L_2. For traversing $\widehat{L}_1$, there is no clear notion of the left-right order. But any ordering that starts near $f\big((1/4,h_2)\big)$, and later does not visit any vertex before visiting an immediate neighbor, will do. Instead of the left neighbor p_0' of a vertex $p_0 \in L_1$, we use for $\widehat{p}_0 \in \widehat{L}_1$ that neighbor of $\widehat{p}_0$ in $\widehat{L}_1$ that is "most counterclockwise", in the appropriate sense. The rest of the proof proceeds with essentially no modifications, except that the coordinate system used is transformed by f. ∎

Remark 10.10: In [BLPS98] it has been asked whether the free USF on every planar proper bounded degree graph is a tree. The proof of Theorem 10.7 seems to be relevant. It is plausible that with a similar argument one can prove that for proper planar graphs with bounded degree and a bounded number of sides per face, the free USF is a tree.

We may now strengthen Corollary 10.6, as follows

COROLLARY 10.11: *Given every $\epsilon \in (0,1)$, there is an $r \in (0,\epsilon)$ with the following property. For every sufficiently small $\delta > 0$, the probability that there is a point $p \in \mathbb{S}^2$ such that there are 4 disjoint crossings in $\widehat{\mathsf{T}}_\delta$ of the annulus $A_{\mathrm{sp}}(p,r,\epsilon)$ is at most ϵ.*

Proof: Consider an annulus $A = A_{\mathrm{sp}}(p,r,\epsilon)$, and suppose that there are four disjoint paths $\alpha_0, \alpha_1, \alpha_2, \alpha_3$ in $\widehat{\mathsf{T}}_\delta$ that cross it. Let B_1 be the component of $\mathbb{S}^2 - A$ inside the inner boundary component of A, and let B_2 be the outside component. Without loss of generality, we suppose that $\alpha_0 \cup \alpha_2$ separate α_1 from α_3 inside A, that is, the circular order of these paths around A agrees with the order of the indices.

Assume first that there are no paths in $\widehat{\mathsf{T}}_\delta \cap A$ that join two of the paths α_j, $j = 0,1,2,3$. Then there must be paths in the dual tree $\beta_0, \beta_1, \beta_2, \beta_3$, such that β_j is between α_{j-1} and α_j (indices mod 4), for each $j = 0,1,2,3$. If $\alpha_0, \alpha_1, \alpha_2$ and α_3 can all be connected to each other by paths in $A \cup B_1$, it follows that there are four crossings of the annulus $A_{\mathrm{sp}}(p,r,\epsilon/3)$ in the $\epsilon/3$ trunk of $\widehat{\mathsf{T}}_\delta$, and we know that has small probability to happen anywhere, if r is small, by Corollary 10.6. If neither of the paths α_j connects to another in $A \cup B_1$, the same argument applies to the dual tree, because the paths β_j must all connect inside $A \cup B_1$. However, if two of the paths α_j connect in $A \cup B_1$, and one of the others does not connect to them, then also two of the paths β_j connect. This implies that the $\epsilon/3$ trunk gets within distance of r from the dual trunk, and again this can be discarded as having small likelihood.

We are left to deal with the situation where there is a simple path γ in $A \cap \widehat{\mathsf{T}}_\delta$ that connects two of the paths α_j. Note that for each pair of paths α_j there can be at most one such γ connecting them. Also note that any path connecting α_j and α_{j+2} (indices mod 4) must cross either α_{j+1} or α_{j+3}. Consequently, if we consider any four concentric annuli $A_j := A_{\mathrm{sp}}(p, r_j, r_{j+1})$, $j = 0,1,2,3,4$, with $r_j < r_{j+1}$ for each $j = 0,1,2,3,4$, $r_0 = r$, and $r_5 = \epsilon$, at least one of them will have the property that inside it there is no path joining any two paths among the α_j's. This allows a reduction to the previous case, and completes the proof. ∎

THEOREM 10.12: *A.s., every simple path ϕ: $[0,1) \to$ trunk has a limit $\lim_{t\to 1} \phi(t)$ in $\mathbb{S}^2$, and for every point $z \in \mathbb{S}^2$ there is a simple path ϕ: $[0,1) \to$ trunk such that $\lim_{t\to 1} \phi(t) = z$.*

Proof: Suppose that there are two distinct accumulation points, x and y, of $\phi(t)$ as $t \to 1$, and let $m \in \mathbb{N}_+$ satisfy $d_{sp}(x,y) > 9/m$. Then for each $t \in (0,1)$

the (spherical) diameter of $\phi([t,1))$ is greater than $9/m$. Let $a \in (0,1)$ be such that the diameter of $\phi([0,a])$ is at least $5/m$. It easily follows that $\phi([a,1)) \subset \mathsf{trunk}_0(1/m)$. But since $\mathsf{trunk}_0(1/m)$ is a compact finite tree (Corollary 10.3), and the restriction of ϕ to $[a,1)$ is in $\mathsf{trunk}_0(1/m)$, it follows that $\lim_{t\to 1}\phi(t)$ exists. Contradiction.

Let $z \in \mathbb{S}^2$ and $z' \in \mathsf{trunk}_0(1)$, $z' \neq z$. We want to produce a simple path $\gamma \subset \mathsf{trunk}$ starting at z' and tending to z. If $z \in \mathsf{trunk}$, then $z \in \mathsf{trunk}_0(1/n)$ for some $n \in \mathbb{N}_+$, and the existence of γ is clear. So suppose that $z \notin \mathsf{trunk}$. For each $n \in \mathbb{N}_+$, there is a point $z_n \in \mathsf{trunk}_0(1/n)$ which is within distance $2/n$ from z. Let β_n be the arc from z' to z_n in $\mathsf{trunk}_0(1/n)$. For each n and m, the intersection $\gamma_n \cap \mathsf{trunk}_0(1/m)$ is a simple path. Since $\mathsf{trunk}_0(1/m)$ is a compact finite topological tree, there is a subsequence γ_{n_j} such that for each m the Hausdorff limit $\lim_j(\gamma_{n_j} \cap \mathsf{trunk}_0(1/m))$ exists, and is a simple path. Because $\gamma_n - B(z,3/m) \subset \mathsf{trunk}_0(1/m)$ when $n \geqslant m$, it now follows that $\gamma := \lim_j \gamma_{n_j}$ is a simple path. Moreover, it is clear that $z \in \gamma$ and $\gamma - \{z\} \subset \mathsf{trunk}$. ∎

Proof of Theorem 1.6: We first prove that trunk is a topological tree. Clearly, the trunk is arcwise connected, since $\mathsf{trunk} := \bigcup_n \mathsf{trunk}_0(1/n)$, and each $\mathsf{trunk}_0(1/n)$ is arcwise connected. It is also clear that the trunk is dense in $\mathbb{S}^2$. Let $x, y \in \mathsf{trunk}$. Then there is some $n \in \mathbb{N}$ such that $x, y \in \mathsf{trunk}_0(1/n)$, and there is a unique arc γ_1 joining x and y in $\mathsf{trunk}_0(1/n)$. Let γ_2 be an arc joining x and y in trunk. Since the dual trunk is dense, it must intersect all connected components of $\mathbb{S}^2 - (\gamma_1 \cup \gamma_2)$. Since the dual trunk is disjoint from the trunk, it does not intersect $\gamma_1 \cup \gamma_2$. Because the dual trunk is connected, it now follows that $\mathbb{S}^2 - (\gamma_1 \cup \gamma_2)$ is connected. Consequently, $\gamma_1 = \gamma_2$, and the trunk is uniquely arcwise connected.

Let $n \in \mathbb{N}$ and let t be the spherical distance between $\mathsf{trunk}_0(1/n)$ and $\mathsf{trunk}_0^\dagger(1/n)$. Since these are compact and disjoint, $t > 0$. If $x \in \mathsf{trunk}_0(1/n)$ and there is a $y \in \mathsf{trunk}$ such that there is no path in $\mathsf{trunk} \cap B(x,3/n)$ joining x and y, then there must be a path in $\mathsf{trunk}_0^\dagger(1/n)$ separating x and y in $B(x,1/n)$. (Indeed, if γ is the path joining x and y and $p \in \gamma - B(x,3/n)$, then the path $\beta \subset \mathsf{trunk}^\dagger$ connecting two points p' and p'' that are near p and are separated from each other by γ near p, will have a subarc in $\mathsf{trunk}_0^\dagger(1/n)$ separating x and y in $B(x,1/n)$.) Consequently, for every point $x \in \mathsf{trunk}_0(1/n)$ and every $y \in \mathsf{trunk} \cap B(x,t) \cap B(x,1/n)$ there is a path in $\mathsf{trunk} \cap B(x,3/n)$ joining x and y. Hence, the union of all arcs that contain x and are contained in $\mathsf{trunk} \cap B(x,3/n)$ is an arcwise connected subset of $\mathsf{trunk} \cap B(x,3/n)$ which contains $\mathsf{trunk} \cap B(x,t) \cap B(x,1/n)$. This implies that trunk is locally arcwise connected, and so it is a topological tree. It is obviously dense in $\mathbb{S}^2$, and the

proof of part (iii) is complete.

It is clear that for every $a, b \in \mathbb{S}^2$ there is some ω such that $(a, b, \omega) \in \mathfrak{T}$.

Let $\mathfrak{T}(\epsilon)$ be a subsequential scaling limit of $\mathfrak{T}_\delta(\epsilon)$. We prove that a.s. every ω such that $(a, b, \omega) \in \mathfrak{T}(\epsilon)$ for some $a, b \in \mathbb{S}^2$ is a simple path. Let $\epsilon_1 > 0$. It suffices to prove that the above statement holds with probability at least $1 - \epsilon_1$. Let $V_1 \subset \mathbb{S}^2$ be a finite set of points, and for each $\delta > 0$ let V_1^δ be a collection of vertices of $\delta\mathbb{Z}^2$, each close to one point of V_1, and with $|V_1| = |V_1^\delta|$. By Theorem 10.2, V_1 may be chosen so that with probability at least $1 - \epsilon_1$ the subtree of $\widehat{\mathsf{T}}_\delta$ spanned by V_1^δ contains $\mathsf{trunk}_\delta(\epsilon)$, for all sufficiently small δ. This implies that each $\omega_{a,b}(\epsilon)$ is a subarc of $\omega_{v,u}$ for some $v, u \in V_1^\delta$. Because for every pair of points $v, u \in V_1$ the scaling limit of the LERW from v to u is a simple path, it follows that with probability at least $1 - \epsilon_1$ for each $(a, b, \omega) \in \mathfrak{T}(\epsilon)$, ω is a simple path.

We may now conclude that a.s. for every $(a, b, \omega) \in \mathfrak{T}$, the set

$$\omega - \left(\overline{B}(a, \epsilon) \cup \overline{B}(b, \epsilon)\right)$$

is a 1-manifold, that is, a disjoint union of simple paths. Therefore, $\omega' := \omega - \{a, b\}$ is a 1-manifold. This means that each component of ω' is an arc with endpoints in $\{a, b\}$. It is clear that ω may be oriented as a path from a to b. Suppose that ω visits a more than once. If $a \neq b$, it then follows that $a \in \mathsf{trunk}$, and there is a simple closed path in trunk containing a. This is impossible, since trunk is a topological tree. Hence a, and similarly b, are each visited only once in ω, which implies that ω is a simple path if $a \neq b$. If $a = b$, the only possibility is that $\omega = \{a\}$ or that ω is a simple closed path. This proves the first and second statements in (ii).

Observe that if there is simple curve $\alpha \subset \mathsf{trunk}$ such that $\overline{\alpha} = \alpha \cup \{a\}$, then a must be in the dual trunk, for the dual trunk is connected, intersects both components of $\mathbb{S}^2 - \overline{\alpha}$, and is disjoint from trunk. This is a rare event, by Remark 3.2 (or Corollary 10.4). This proves (ii).

Corollary 10.11 proves (iv).

The first claim in (i) is obvious. Suppose that $a \neq b$ are such that there are two sets ω and ω' with $(a, b, \omega), (a, b, \omega') \in \mathfrak{T}$. We know that ω and ω' are simple paths. If $\omega \neq \omega'$, then there is a simple closed path, say γ, contained in $\overline{\omega} \cup \overline{\omega}'$. But as above, γ must intersect the dual trunk, since the dual trunk is connected and dense. This implies that a or b are in the dual trunk. This completes the proof of (i), and of the theorem. ∎

Remark 10.13: Uniqueness of paths. We have seen in the above proof that the path in trunk from a to b is unique when $\{a,b\} \cap \mathsf{trunk}^\dagger = \emptyset$. The converse is also easily established.

Remark 10.14 (Reconstructing $\mathsf{trunk}^\dagger$): It can be shown that the scaling limit dual trunk can be reconstructed from the trunk. This can be seen from Remark 10.13. Another description of the dual trunk from the trunk is as follows. Given distinct $x, y \in \mathbb{S}^2 - \mathsf{trunk}$, let $\gamma(x,y)$ be the (unique) arc in $\mathbb{S}^2 - \mathsf{trunk}$ with endpoints x, y. Then

$$\mathsf{trunk}^\dagger = \bigcup \Big\{ \gamma(x,y) - \{x,y\} \colon x \neq y, x, y \in \mathbb{S}^2 - \mathsf{trunk} \Big\}.$$

To prove this, it suffices to establish that $\gamma(x,y)$ is unique, which follows from the fact that trunk is connected and dense.

Remark 10.15: Consider the metric d^* on trunk, where $d^*(x,y)$ is the spherical diameter of the unique (possibly degenerate) arc joining x and y in trunk. Since trunk is locally arcwise connected, this new metric on trunk is compatible with the topology of trunk as a subset of $\mathbb{S}^2$. Let trunk_* denote the completion of this metric. Then trunk_* is a compact topological tree, and is naturally homeomorphic with the ends compactification of trunk. Since d^* majorizes the spherical metric, there is a natural projection $\pi \colon \mathsf{trunk}_* \to \mathbb{S}^2$, whose restriction to trunk is the identity. It is easy to see that every point $p \in \mathbb{S}^2 - \mathsf{trunk}^\dagger$ has a unique preimage under π, and for points $p \in \mathsf{trunk}^\dagger$, the degree of p in $\mathsf{trunk}^\dagger$ is equal to $|\pi^{-1}(p)|$.

Consider some $o \in \mathbb{S}^2$, and let $\mathfrak{T}^o$ be the appropriate "slice" of $\mathfrak{T}$, that is, $\mathfrak{T}^o := \{(b,\omega) \colon (o,b,\omega) \in \mathfrak{T}\}$. One can show that if $o \notin \mathsf{trunk}^\dagger$, then $\mathfrak{T}^o$ is homeomorphic with trunk_*, and $\pi \colon \mathfrak{T}^o \to \mathbb{S}^2$ is the projection onto the first coordinate, when $\mathfrak{T}^o$ is identified with trunk_* through this homeomorphism.

11. Free and wired trunks and conformal invariance

We now want to give a precise formulation to a conformal invariance conjecture for the UST scaling limit, and prove that it follows from the conjectured conformal invariance of the LERW scaling limit. (Such conformal invariance conjectures seem to be floating in the air these days, with roots in the physics community.) The conformal automorphisms of $\mathbb{S}^2$ are Möbius transformations. We conjecture that the different notions of scaling limits of UST in $\mathbb{S}^2$, which were introduced in the previous section, exist (without a need to pass to a subsequence) and are invariant under Möbius transformations. Moreover, the scaling limits in subdomains $D \subset \mathbb{S}^2$ should be invariant under conformal homeomorphisms $f \colon D \to D' \subset \mathbb{S}^2$. This is a significantly stronger statement, since the

Möbius transformations of $\mathbb{S}^2$ form a 6-dimensional group, while the space of conformal homeomorphisms from the unit disk onto subdomains of $\mathbb{S}^2$ is infinite dimensional.

To formulate more precisely the invariance under conformal homeomorphisms of subdomains $f\colon D \to D'$, we need first to discuss UST scaling limits in subdomains of $\mathbb{S}^2$. This will be now explained.

For simplicity, we restrict our attention to simply connected domains $D \subsetneq \mathbb{R}^2$ whose boundary ∂D is a simple closed path. Let $p_0 \in D$ be some basepoint. Let FT_δ^D be the uniform spanning tree of $G(D,\delta)$, with free boundary conditions, and let WT_δ^D be the uniform spanning tree of $G(D,\delta)$ with wired boundary conditions. Let $\mathbb{S}_D^2$ be the metric space obtained from $\mathbb{S}^2$ by contracting $\mathbb{S}^2 - D$ to a single point. Then we may think of WT_δ^D as a random point in $\mathcal{H}(\mathbb{S}_D^2)$, which is a.s. a tree. The tree FT_δ^D may be thought of as a random point in $\mathcal{H}(\overline{D})$, which is a.s. a tree.

Let $\mathfrak{WT}_\delta^D := \mathfrak{T}(\mathsf{WT}_\delta^D)$ and $\mathfrak{FT}_\delta^D := \mathfrak{T}(\mathsf{FT}_\delta^D)$, that is, the **wired paths ensemble** $\mathfrak{WT}_\delta$ is defined from the wired tree WT_δ in exactly the same way that the ordinary paths ensemble $\mathfrak{T}_\delta$ was defined from $\widehat{\mathsf{T}}_\delta$, and similarly for $\mathfrak{FT}_\delta$. Note that $\mathfrak{WT}_\delta \in \mathcal{H}\big(\mathbb{S}_D^2 \times \mathbb{S}_D^2 \times \mathcal{H}(\mathbb{S}_D^2)\big)$ and $\mathfrak{FT}_\delta \in \mathcal{H}\big(\overline{D} \times \overline{D} \times \mathcal{H}(\overline{D})\big)$. Also the definitions of the scaling limits and the trunk are the same as in the previous section.

THEOREM 11.1: *Let $D \subset \mathbb{R}^2$ be a domain whose boundary is a C^1-smooth simple closed curve.*

(i) *Theorem 10.2, with $\mathbb{S}^2$ replaced by D, holds for the free and wired spanning trees in D.*

(ii) *The free scaling limit trunk in D is disjoint from ∂D, in every (subsequential) scaling limit.*

(iii) *The free scaling limit trunk in D is disjoint from the scaling limit trunk of the dual tree (which is wired), in every (subsequential) scaling limit.*

There are simply connected domains where ∂D is not a simple closed curve and (ii) fails: the domain $(0,2) \times (0,2) - \bigcup_{n=1}^{\infty} (0,1] \times \{1/n\}$ is an example.

LEMMA 11.2: *There is an absolute constant $C > 0$ such that the following holds true. Let D be as in Theorem 11.1. Then there is a $\delta_0 = \delta_0(D) > 0$ with the following property. Suppose that δ and δ_1 are numbers satisfying $0 < \delta \leqslant \delta_1 \leqslant \delta_0$ and A is a connected subgraph of $G(D,\delta)$ with diameter at least δ_1. Further suppose that $p \in \mathsf{V}\big(G(D,\delta)\big)$ has distance δ_1 to A. Then the probability that a random walk on $G(D,\delta)$ starting at p will get to distance $C\delta_1$ from p before hitting A is less than $1/2$.*

Proof: The proof is similar to the proof of Lemma 10.9. Let $Z = Z(t)$ be the set of vertices with distance at least t from p, where we take $t > 4\delta_1$. Let A' be a component of $A \cap B(p, 3\delta_1)$ containing some point at distance δ_1 to p and having diameter at least δ_1. For $v \in \mathsf{V}\big(G(D,\delta)\big)$, let $h(v)$ be the probability that a random walk starting from v will reach Z before hitting A'. Then h is discrete-harmonic, and minimizes Dirichlet energy among functions that are 1 on Z and 0 on A'. As in the proof of Lemma 10.9, it follows that the Dirichlet energy of h is at most $O(1)/\log(t/\delta_1)$. Let B be the set of vertices v at distance at most $2\delta_1$ from p such that $h(v) \geqslant h(p)$, and let B' be the component of B containing p. Note that the diameter of B' is at least δ_1, as B' must neighbor with some vertex with distance $2\delta_1$ from p, by the maximum principle for h. As in the proof of Lemma 10.9, it can be shown that when δ_1 is sufficiently small (how small depends on the scale in which ∂D appears smooth), one can find $O(1)/\delta$ disjoint paths in $\delta\mathbb{Z}^2 \cap D$ connecting A' to B', each of combinatorial length $O(1)/\delta$. Because h is zero on A' and at least $h(p)$ on B', it follows that the Dirichlet energy of h is at least $O(1)h(p)$. We conclude that $h(p) \leqslant O(1)/\log(t/\delta_1)$, which proves the lemma. ∎

Proof of Theorem 11.1: The proof for (i) in the wired case is the same as the proof of Theorem 10.2 (and we don't need to assume anything about D). The free case is the same, except that one needs to appeal to Lemma 11.2. Assuming (ii), the proof of (iii) is identical to the proof of Theorem 10.7. The proof of (ii) is also the same as the proof of 10.7, except that one needs to find the appropriate substitutes for Lemma 10.8 and Corollary 10.6; namely, for every $\epsilon > 0$ there is an $\epsilon_0 > 0$ such that for all $\delta \in (0, \epsilon_0)$ with probability at least $1 - \epsilon$ all points of degree three in the ϵ-trunk of FT_δ^D have distance at least ϵ_0 from ∂D, and the probability that there is a point $p \in \partial D$ such that there are two disjoint crossings of the annulus $A(p, \epsilon_0, \epsilon)$ in the ϵ-trunk of WT_δ^D is at most ϵ. The latter statement follows from the proof of Corollary 10.6.

It remains to prove the appropriate substitute for 10.8. Consider three distinct points in D, p_1, p_2, p_3, and for $\delta > 0$ let p_1', p_2', p_3' be a triple of points in $\delta\mathbb{Z}^2$ which is close to p_1, p_2, p_3, respectively. Let m be the meeting point of p_1', p_2', p_3' in FT_δ^D, and let B be any disk whose center is in ∂D and which does not intersect $\{p_1', p_2', p_3'\}$. Let B' and B'' be disks concentric with B of $1/2$ and $1/4$ of its size, respectively. By part (i), it suffices to prove that with probability going to 1 as $\epsilon_0 \to 0$ and $\delta \to 0$, m is not within distance ϵ_0 from $\partial D \cap B''$.

Suppose that $m \in B$. Let $\gamma_1, \gamma_2, \gamma_3$ be the arcs of FT_δ^D that join m to p_1', p_2', p_3', respectively, and let γ_j' be the largest initial segment of γ_j that is contained in B,

$j = 1, 2, 3$. There is a unique $k \in \{1, 2, 3\}$ such that $\bigcup_{j \neq k} \gamma'_j$ separates γ'_k from ∂D in B. By symmetry, it suffices to estimate the probability that m is close to $B \cap \partial D$ and $k = 3$. We generate m in the following way. Let γ be a LERW from p'_1 to p'_2 in $\delta\mathbb{Z}^2 \cap D$. Let X be a random walk on $\delta\mathbb{Z}^2 \cap D$ starting at p'_3, let τ_γ be the first time t where $X(t) \in \gamma$, and let τ_1 be the first time t when $X(t)$ is incident with an edge intersecting $\partial D \cap B'$. Then we may take $m = X(\tau_\gamma)$. Consider the event $\mathcal{A}$ where $m \in B$, $k = 3$ and $\tau_1 \geqslant \tau_\gamma$. With high probability, the points $X(t)$, $t \leqslant \tau_1$, which are close to $B' \cap \partial D$, are also close to $X(\tau_1)$. This is just a property of simple random walk absorbed at $B' \cap \partial D$. Consequently, on $\mathcal{A}$, with high probability, if m is close to $B \cap \partial D$, then γ passes close to $X(\tau_1)$. Since the probability that γ passes near any point, which is not too close to p'_1 and p'_2, is small (Remark 3.2 applies here), and $X(\tau_1)$ is independent of γ, we see that $\mathcal{A}$ has arbitrarily small probability.

We now need to consider the case $k = 3$ and $\tau_1 < \tau_\gamma$. Let τ'_1 be the first $t \geqslant \tau_1$ such that $X(t) \notin B$, and let τ_2 be the first $t \geqslant \tau'_1$ such that $X(t)$ is incident with an edge intersecting $\partial D \cap B'$. Inductively, let τ'_n be the first $t \geqslant \tau_n$ such that $X(t) \notin B$ and let τ_{n+1} be the first $t \geqslant \tau'_n$ such that $X(t)$ is incident with an edge intersecting $\partial D \cap B'$. Note that if $k = 3$ and $\tau_\gamma > \tau_1$, then $\tau_\gamma > \tau'_1$, for the random walk X must go around $\gamma'_1 \cup \gamma'_2$ before hitting γ. Similarly, if $k = 3$ and $\tau_\gamma > \tau_1$, then $\tau_\gamma \in (\tau'_n, \tau_{n+1})$ for some $n \in \mathbb{N}$. If we fix a finite $k \in \mathbb{N}$, then the same argument as above shows that with high probability, γ does not pass close to the set $\{X(\tau_n)\colon n = 1, 2, \ldots, k\}$. Because $\mathbf{P}[\tau'_k < \gamma] \to 0$ as $k \to \infty$, uniformly in δ, the required result follows. ∎

Suppose that D and D' are two domains in $\mathbb{R}^2$ such that the boundaries $\partial D, \partial D'$ are simple closed paths in $\mathbb{R}^2$. Then there is a conformal homeomorphism $f\colon D \to D'$. Moreover, f extends continuously to a homeomorphism of $\overline{D}$ onto $\overline{D}'$, which we will also denote by f. It follows that f induces maps

$$f_W\colon \mathcal{H}\big(\mathbb{S}^2_D \times \mathbb{S}^2_D \times \mathcal{H}(\mathbb{S}^2_D)\big) \to \mathcal{H}\big(\mathbb{S}^2_{D'} \times \mathbb{S}^2_{D'} \times \mathcal{H}(\mathbb{S}^2_{D'})\big),$$
$$f_F\colon \mathcal{H}\big(\overline{D} \times \overline{D} \times \mathcal{H}(\overline{D})\big) \to \mathcal{H}\big(\overline{D} \times \overline{D} \times \mathcal{H}(\overline{D})\big).$$

THEOREM 11.3: *Let $D \subset \mathbb{R}^2$ be a domain whose boundary is a C^1-smooth simple closed path. Assuming Conjecture 1.2, the following is true.*

(i) *The free and the wired UST scaling limits, $\mathfrak{FT}, \mathfrak{WT}$, in D exist. (That is, they do not depend on the sequence of δ tending to 0.)*

(ii) *If $f\colon D \to D'$ is a conformal homeomorphism between such domains, then f_W is measure preserving from the law of $\mathfrak{WT}$ in D to the law of $\mathfrak{WT}$ in D', and similarly for free boundary conditions.*

Proof: The proof for wired boundary conditions follows from Wilson's algorithm and Theorem 11.1. The easy details are left to the reader.

For the free boundary conditions, observe that WT_δ^D is dual to FT_δ^D (on the dual grid). Remark 10.14 is also valid in the present setting, and shows that the free scaling limit trunk can be reconstructed from the wired scaling limit trunk. It is easy to see that the free scaling limit $\mathfrak{FT}$ can be reconstructed from the trunk. Hence, conformal invariance of the wired UST implies conformal invariance of the free. ∎

12. Speculations about the Peano curve scaling limit

This section will discuss the Peano curve winding between the UST and its dual. From here on, the discussion will be somewhat speculative, and we omit proofs, not because the proofs are particularly hard, but because the paper is long enough as it is, and it is not clear when another paper on this subject will be produced.

This Peano curve was briefly mentioned in [BLPS98]. Consider the set of points $\theta_\delta \subset \mathbb{R}^2$ which have the same Euclidean distance from $\widehat{\mathsf{T}}_\delta$ as from its dual $\widehat{\mathsf{T}}_\delta^\dagger$. It is easy to verify that θ_δ is a simple path in a square grid $G_P(\delta)$, of mesh $\delta/2$, which visits all the vertices in that grid. Set $\widehat{\theta}_\delta := \theta_\delta \cup \{\infty\}$. Then $\widehat{\theta}_\delta$ is a.s. a simple closed path in $\mathbb{S}^2$ passing through ∞.

To consider the scaling limit of $\widehat{\theta}_\delta$, it is no use to think of it as a set of points in $\mathbb{S}^2$, because then the scaling limit will be all of $\mathbb{S}^2$. Rather, one needs to parameterize $\widehat{\theta}_\delta$ in some way. One natural parameterization would be by the area of its $\delta/2$-neighborhood, but there are several other plausible parameterizations. Another, more sophisticated approach, would be to think of $\widehat{\theta}_\delta$ as defining a circular order on the set $\widehat{\theta}_\delta$. The circular order R_δ is a closed subset of $(\mathbb{S}^2)^4$, and $(a,b,c,d) \in R_\delta$ iff $a,b,c,d \in \widehat{\theta}_\delta$ and $\{a,c\}$ separates b from d on $\widehat{\theta}_\delta$. Then the (subsequential) scaling limit of $\widehat{\theta}_\delta$ may be taken as the weak limit of the law of R_δ in $\mathcal{H}\Big((\mathbb{S}^2)^4\Big)$.

Let θ denote the Peano curve scaling limit, defined as a path, or as a circular order, or some other reasonable definition. Here is what we believe to be a description of θ, in terms of the scaling limit of the UST. Recall that in Remark 10.15, we have introduced a completion trunk_* of the trunk, in the metric d^*, where the distance between any two points of trunk is the diameter of the arc connecting them, and that π: $\mathsf{trunk}_* \to \mathbb{S}^2$ is the natural projection. Consider the joint distribution of trunk_* and the dual $\mathsf{trunk}_*^\dagger$, and let $\pi^\dagger$ denote the projection $\pi^\dagger$: $\mathsf{trunk}_*^\dagger \to \mathbb{S}^2$. Let $\tilde{\theta}$ be the set of points $(p,q) \in \mathsf{trunk}_* \times \mathsf{trunk}_*^\dagger$ such that $\pi p = \pi^\dagger q$. Then $\tilde{\theta}$ is a simple closed path, and the map $\tilde{\theta} \mapsto \mathbb{S}^2$ defined by

$(p,q) \mapsto \pi p$ gives the scaling limit θ.

Fix some $a, b \in \mathbb{R}^2$, $a \neq b$. Let $\omega_{a,b}$ be the path such that $(a, b, \omega_{a,b}) \in \mathfrak{T}$ (this $\omega_{a,b}$ is a.s. unique), and let $\omega^\dagger_{a,b}$ be such that $(a, b, \omega^\dagger_{a,b}) \in \mathfrak{T}^\dagger$. Then $\omega_{a,b}$ and $\omega^\dagger_{a,b}$ are a.s. simple paths. Let D_1 and D_2 be the two components of $\mathbb{S}^2 - (\omega_{a,b} \cup \omega^\dagger_{a,b})$. A.s. $\infty \in D_1 \cup D_2$, and without loss of generality take $\infty \in D_1$. It is then clear that the part of θ which is between a, b and does not contain ∞ is $\overline{D}_2$, and that the part which does contain ∞ is $\overline{D}_1$. Suppose that we **condition** on $\omega_{a,b}$ and on $\omega^\dagger_{a,b}$, and look at some point $c \in D_2$. We'd like to know the distribution of the part of θ between c and a which does not include b, say. Recall that on a finite planar graph, we may generate the UST and the dual UST by a modification of Wilson's algorithm, where at each step in which we start from a vertex in the graph, the dual tree built up to that point acts as a free boundary component, and the tree built up to that point acts as an absorbing wired boundary component, and at steps in which we start from a dual vertex, the tree built up to that point acts as a free boundary component, and the dual tree built up to that point acts as an absorbing wired boundary component. Consequently, we let α be the scaling limit of LERW on $\delta\mathbb{Z}^2 - \omega^\dagger_{a,b}$ starting at c that stops when it hits $\omega_{a,b}$. Then we let $\alpha^\dagger$ be the scaling limit of LERW on $\delta\mathbb{Z}^2 - (\alpha \cup \omega_{a,b})$ starting at c that stops when it hits $\omega^\dagger a, b$. (Since $c \in \alpha$, to define this requires taking a limit as the starting point tends to c.) Then $\alpha \cup \alpha^\dagger$ separates D_2 into two regions, say D_2' and D_2'', and if $b \notin \overline{D}_2'$, then D_2' is the part of θ "separated" from b by $\{a, c\}$.

In the above construction, the domain D_2 was considered with mixed boundary conditions. One arc of $\partial D_2 - \{a, b\}$ was taken as wired, while the other was free. The resulting Peano path scaling limit θ is a path joining a and b in $\overline{D}_2$. From Conjecture 1.2 should follow a conformal invariance result for UST in such domains with mixed boundary conditions. Therefore, having an understanding of the law of the Peano curve for one triplet (D, a, b), where a and b are distinct points in ∂D, which is a simple closed curve, suffices for any other such triplet. This suggests that we should take the simplest possible such configuration, that is, $D = \mathbb{H}$, the upper half plane, $a = 0, b = \infty$. Suppose that we then take a point $c \in \mathbb{H}$ and condition on the part of θ between 0 and c and separated from ∞. The effect of that on θ in the remaining subdomain D_c of $\mathbb{H}$ is all in the boundary ∂D_c. This is a kind of Markovian property for the Peano curve, similar to the property given by Lemma 4.3. By taking the conformal map from D_c to $\mathbb{H}$, which fixes ∞, takes c to 0, and is appropriately normalized at ∞, we may return to base one. This suggests that, as we have claimed in the introduction, a representation of the Peano curve scaling limit is similar to the SLE representation of the LERW

from 0 to ∂U which we have introduced. The analogue of the Löwner differential equation for this situation is (1.6). Due to the Markovian nature of the Peano curve, the corresponding parameter ζ in (1.6) should have the form $\zeta = \mathsf{B}(\kappa t)$, where B is Brownian motion on $\mathbb{R}$ starting at 0, and κ is some constant. One can, in fact, show that $\kappa = 8$, by deriving an appropriate analogue of Cardy's [Car92] conjectured formula, using the representation (1.6) and the techniques of Section 9. The details will appear elsewhere.

References

[Aiz] M. Aizenman, *Continuum limits for critical percolation and other stochastic geometric models*, Preprint. http://xxx.lanl.gov/abs/math-ph/9806004.

[ABNW] M. Aizenman, A. Burchard, C. M. Newman and D. B. Wilson, *Scaling limits for minimal and random spanning trees in two dimensions*, Preprint. http://xxx.lanl.gov/abs/math/9809145.

[ADA] M. Aizenman, B. Duplantier and A. Aharony, *Path crossing exponents and the external perimeter in 2D percolation*, Preprint. http://xxx.lanl.gov/abs/cond-mat/9901018.

[Ald90] D. J. Aldous, *The random walk construction of uniform spanning trees and uniform labelled trees*, SIAM Journal on Discrete Mathematics **3** (1990), 450–465.

[Ben] I. Benjamini, *Large scale degrees and the number of spanning clusters for the uniform spanning tree*, in *Perplexing Probability Problems: Papers in Honor of Harry Kesten* (M. Bramson and R. Durrett, eds.), Boston, Birkhäuser, to appear.

[BLPS98] I. Benjamini, R. Lyons, Y. Peres and O. Schramm, *Uniform spanning forests*, Preprint. http://www.wisdom.weizmann.ac.il/~schramm/papers/usf/.

[BJPP97] C. J. Bishop, P. W. Jones, R. Pemantle and Y. Peres, *The dimension of the Brownian frontier is greater than* 1, Journal of Functional Analysis **143** (1997), 309–336.

[Bow] B. H. Bowditch, *Treelike structures arising from continua and convergence groups*, Memoirs of the American Mathematical Society, to appear.

[Bro89] A. Broder, *Generating random spanning trees*, in *30th Annual Symposium on Foundations of Computer Science*, IEEE, Research Triangle Park, NC, 1989, pp. 442–447.

[BP93] R. Burton and R. Pemantle, *Local characteristics, entropy and limit theorems for spanning trees and domino tilings via transfer-impedances*, The Annals of Probability **21** (1993), 1329–1371.

[Car92] J. L. Cardy, *Critical percolation in finite geometries*, Journal of Physics A **25** (1992), L201–L206.

[DD88] B. Duplantier and F. David, *Exact partition functions and correlation functions of multiple Hamiltonian walks on the Manhattan lattice*, Journal of Statistical Physics **51** (1988), 327–434.

[Dur83] P. L. Duren, *Univalent Functions*, Springer-Verlag, New York, 1983.

[Dur84] R. Durrett, *Brownian Motion and Martingales in Analysis*, Wadsworth International Group, Belmont, California, 1984.

[Dur91] R. Durrett, *Probability*, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1991.

[EK86] S. N. Ethier and T. G. Kurtz, *Markov Processes*, Wiley, New York, 1986.

[Gri89] G. Grimmett, *Percolation*, Springer-Verlag, New York, 1989.

[Häg95] O. Häggström, *Random-cluster measures and uniform spanning trees*, Stochastic Processes and their Applications **59** (1995), 267–275.

[Itô61] K. Itô, *Lectures on Stochastic Processes*, Notes by K. M. Rao, Tata Institute of Fundamental Research, Bombay, 1961.

[Jan12] Janiszewski, Journal de l'Ecole Polytechnique **16** (1912), 76–170.

[Ken98a] R. Kenyon, *Conformal invariance of domino tiling*, Preprint. http://topo.math.u-psud.fr/~kenyon/confinv.ps.Z.

[Ken98b] R. Kenyon, *The asymptotic determinant of the discrete laplacian*, Preprint. http://topo.math.u-psud.fr/~kenyon/asymp.ps.Z.

[Ken99] R. Kenyon, *Long-range properties of spanning trees*, Preprint.

[Ken] R. Kenyon, in preparation.

[Kes87] H. Kesten, *Hitting probabilities of random walks on* $\mathbb{Z}^d$, Stochastic Processes and their Applications **25** (1987), 165–184.

[Kuf47] P. P. Kufarev, *A remark on integrals of Löwner's equation*, Doklady Akademii Nauk SSSR (N.S.) **57** (1947), 655–656.

[LPSA94] R. Langlands, P. Pouliot and Y. Saint-Aubin, *Conformal invariance in two-dimensional percolation*, Bulletin of the American Mathematical Society (N.S.) **30** (1994), 1–61.

[Law93] G. F. Lawler, *A discrete analogue of a theorem of Makarov*, Combinatorics, Probability and Computing **2** (1993), 181–199.

[Law] G. F. Lawler, *Loop-erased random walk*, in *Perplexing Probability Problems: Papers in Honor of Harry Kesten* (M. Bramson and R. Durrett, eds.), Boston, Birkhäuser, to appear.

[Löw23] K. Löwner, *Untersuchungen über schlichte konforme abbildungen des einheitskreises, I*, Mathematische Annalen **89** (1923), 103–121.

[Lyo98] R. Lyons, *A bird's-eye view of uniform spanning trees and forests*, in *Microsurveys in Discrete Probability (Princeton, NJ, 1997)*, American Mathematical Society, Providence, RI, 1998, pp. 135–162.

[MR] D. E. Marshall and S. Rohde, in preparation.

[MMOT92] J. C. Mayer, L. K. Mohler, L. G. Oversteegen and E. D. Tymchatyn, *Characterization of separable metric $\mathbb{R}$-trees*, Proceedings of the American Mathematical Society **115** (1992), 257–264.

[MO90] J. C. Mayer and L. G. Oversteegen, *A topological characterization of $\mathbb{R}$-trees*, Transactions of the American Mathematical Society **320** (1990), 395–415.

[New92] M. H. A. Newman, *Elements of the Topology of Plane Sets of Points*, second edition, Dover, New York, 1992.

[Pem91] R. Pemantle, *Choosing a spanning tree for the integer lattice uniformly*, The Annals of Probability **19** (1991), 1559–1574.

[Pom66] C. Pommerenke, *On the Loewner differential equation*, The Michigan Mathematical Journal **13** (1966), 435–443.

[Rus78] L. Russo, *A note on percolation*, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **43** (1978), 39–48.

[SD87] H. Saleur and B. Duplantier, *Exact determination of the percolation hull exponent in two dimensions*, Physical Review Letters **58** (1987), 2325–2328.

[Sch] O. Schramm, in preparation.

[Sla94] G. Slade, *Self-avoiding walks*, The Mathematical Intelligencer **16** (1994), 29–35.

[SW78] P. D. Seymour and D. J. A. Welsh, *Percolation probabilities on the square lattice*, in *Advances in Graph Theory (Cambridge Combinatorial Conference, Trinity College, Cambridge, 1977)*, Annals of Discrete Mathematics **3** (1978), 227–245.

[TW98] B. Tóth and W. Werner, *The true self-repelling motion*, Probability Theory and Related Fields **111** (1998), 375–452.

[Wil96] D. B. Wilson, *Generating random spanning trees more quickly than the cover time*, in *Proceedings of the Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996)*, ACM, New York, 1996, pp. 296–303.

Acta Math., 187 (2001), 237–273
© 2001 by Institut Mittag-Leffler. All rights reserved

Values of Brownian intersection exponents, I: Half-plane exponents

by

GREGORY F. LAWLER
Duke University
Durham, NC, U.S.A.

ODED SCHRAMM
Microsoft Research
Redmond, WA, U.S.A.

and

The Weizmann Institute of Science
Rehovot, Israel

WENDELIN WERNER
Université Paris-Sud
Orsay, France

1. Introduction

Theoretical physics predicts that conformal invariance plays a crucial role in the macroscopic behavior of a wide class of two-dimensional models in statistical physics (see, e.g., [4], [6]). For instance, by making the assumption that critical planar percolation behaves in a conformally invariant way in the scaling limit, and using ideas involving conformal field theory, Cardy [7] produced an exact formula for the limit, as $N\to\infty$, of the probability that, in two-dimensional critical percolation, there exists a cluster crossing the rectangle $[0,aN]\times[0,bN]$. Also, Duplantier and Saleur [13] predicted the "fractal dimension" of the hull of a very large percolation cluster. These are just two examples among many such predictions.

In 1988, Duplantier and Kwon [12] suggested that the ideas of conformal field theory can also be applied to predict the intersection exponents between random walks in $\mathbf{Z}^2$ (and Brownian motions in $\mathbf{R}^2$). They predicted, for instance, that if B and B' are independent planar Brownian motions (or simple random walks in $\mathbf{Z}^2$) started from distinct points in the upper half-plane $\mathbf{H}=\{(x,y):y>0\}=\{z\in\mathbf{C}:\mathrm{Im}(z)>0\}$, then when $n\to\infty$,

$$\mathbf{P}[B[0,n]\cap B'[0,n]=\varnothing]=n^{-\zeta+o(1)} \tag{1.1}$$

and

$$\mathbf{P}[B[0,n]\cap B'[0,n]=\varnothing \text{ and } B[0,n]\cup B'[0,n]\subset\mathbf{H}]=n^{-\tilde{\zeta}+o(1)}, \tag{1.2}$$

where

$$\zeta=\tfrac{5}{8},\quad \tilde{\zeta}=\tfrac{5}{3}.$$

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_28, © Springer Science+Business Media, LLC 2011

Very recently, Duplantier [11] gave another physical derivation of these conjectures based on "quantum gravity".

In 1982, Mandelbrot [35] suggested that the Hausdorff dimension of the Brownian frontier (i.e., the boundary of a connected component of the complement of the path) is $\frac{4}{3}$, based on simulations and the analogy with the conjectured value for the fractal dimension of self-avoiding walks predicted by Nienhuis (also $\frac{4}{3}$; see, e.g., [33]).

To date, none of the physicists' arguments have been made rigorous, and it seems very difficult to use their methods to produce proofs. Very recently, Kenyon [16], [17], [18] managed to derive the exact values of critical exponents for "loop-erased random walk" that theoretical physicists had predicted (Majumdar [34], Duplantier [10]). Kenyon's methods involve the relation of the loop-erased walk to the uniform spanning tree and to domino tilings. Kenyon shows that the equations relating probabilities of some domino tiling events are discrete analogues of the Cauchy–Riemann equations, and therefore the probabilities can be approximated by analytic functions with prescribed boundary behavior. These methods do not seem applicable for the goals of the present paper.

For planar Brownian motions, it is easy to show, using subadditivity arguments and the scaling property, that there exist positive finite numbers ζ and $\tilde{\zeta}$ such that (1.1) and (1.2) are true. Up to the present paper, there was not even a mathematical heuristic arguing that the values of ζ and $\tilde{\zeta}$ are $\frac{5}{8}$ and $\frac{5}{3}$. Burdzy–Lawler [5] (see also [9], [24]) showed that the intersection exponents were indeed the same for simple random walks as for Brownian motions; Lawler [21] proved that the Hausdorff dimension of the set of cut points of a Brownian path is $2-2\zeta$. He also showed (see [22], [23]) that the dimension of the Brownian frontier (and more generally the whole multifractal spectrum of the Brownian frontier) can be expressed in terms of exponents defined analogously to ζ. As part of that work, he showed that the right-hand side of (1.1) can be replaced with $n^{-\zeta}g(n)$ where g is bounded away from 0 and infinity; we expect that the argument can be adapted to show that the same is true for (1.2).

Recently, Lawler and Werner [28] extended the definition of intersection exponents in a natural way to "non-integer packets of Brownian motions", and derived certain functional relations between these exponents. These relations indicate that Mandelbrot's conjecture that the dimension of the Brownian frontier is $\frac{4}{3}$ is indeed compatible with the predictions of Duplantier–Kwon. It turned out that intersection exponents in the half-plane play an important role in understanding exponents in the whole plane. Conformal invariance of planar Brownian motion is a crucial tool in the derivation of these relations. In particular, there is a measure on Brownian excursions in domains that has some strong conformal invariance properties, including a "restriction" (or "locality") property.

In another paper, Lawler and Werner [29] showed that intersection exponents as-

sociated to any conformally invariant measure on sets with this restriction property are very closely related to the Brownian exponents. This provides a rigorous justification to the link between the conjectures regarding intersection exponents for planar Brownian motions and conjectures for intersection exponents of critical percolation clusters (see [13], [8], [3]), because percolation clusters are conjectured to be conformally invariant in the scaling limit—see, e.g., [19], [2]—and they should also have a restriction property (because of the independence properties of percolation). The question of how to compute these exponents remained open.

Independently, Schramm [42] defined a new class of conformally invariant stochastic processes indexed by a real parameter $\kappa \geqslant 0$, called SLE_κ (for stochastic Löwner evolution process with parameter κ). The definition of these processes is based on Löwner's ordinary differential equation that encodes in a conformally invariant way a continuous family of shrinking domains (see, e.g., [32], [37]). More precisely, [42] defines a family of conformal maps g_t from subsets D_t of $\mathbf{H}$ onto $\mathbf{H}$ by the equation

$$\partial_t g_t(z) = \frac{-2}{\beta_{\kappa t} - g_t(z)}, \tag{1.3}$$

where β is a standard Brownian motion on the real line. (Actually, in [42], instead of (1.3), the corresponding equation for the inverse maps g_t^{-1} is considered.) The domain D_t can be defined as the set of $z_0 \in \mathbf{H}$ such that a solution $g_s(z_0)$ of this equation exists for $s \in [0,t]$. When t increases, the set $K_t = \mathbf{H} \setminus D_t$ increases: Loosely speaking, $(K_t, t \geqslant 0)$ can be viewed as a growing "hull" that is penetrating the half-plane. By applying a conformal homeomorphism $f: \mathbf{H} \to D$, SLE_κ can similarly be defined in any simply-connected domain $D \subsetneq \mathbf{C}$.

In [42], the main focus is on the case $\kappa = 2$, which is conjectured there to correspond to the scaling limit of loop-erased random walks, but the conjecture that SLE_6 corresponds to the scaling limit of critical percolation cluster boundaries is also mentioned. In particular, it is possible to compute explicitly the probability that an SLE_6 crosses a rectangle of size $a \times b$. It turns out that this result is exactly Cardy's formula. This gives a mathematical proof for Cardy's formula, assuming the still open conjecture that SLE_6 is indeed the scaling limit of percolation cluster boundaries.

The main goal of the present paper is to prove some of the conjectured values of intersection exponents of Brownian motion in a half-plane.

THEOREM 1.1. *Let $B^1, \ldots, B^p$ denote p independent planar Brownian motions $(p \geqslant 2)$ started from distinct points in the upper half-plane $\mathbf{H}$. Then, when $t \to \infty$,*

$$\mathbf{P}[\forall i \neq j \in \{1, \ldots, p\},\ B^i[0,t] \cap B^j[0,t] = \varnothing \text{ and } B^i[0,t] \subset \mathbf{H}] = t^{-\tilde{\zeta}_p + o(1)},$$

where

$$\tilde{\zeta}_p = \tfrac{1}{6} p(2p+1).$$

These values have been predicted by Duplantier and Kwon [12]. In particular, $\tilde{\zeta}=\tilde{\zeta}_2=\frac{5}{3}$.

We also establish the exact value (and confirm some of the conjectures stated in [28], [11]) of more general intersection exponents between packets of Brownian motions in the half-plane; see Theorem 4.1.

The proof of Theorem 1.1 uses a combination of ideas from the papers [28], [29], [42]. However, to make the paper more accessible and self-contained, we attempt to review and explain all the necessary background. The reader who wishes to see complete proofs for all stated theorems has to be familiar with the basics of stochastic calculus and conformal mapping theory, and read about the excursion measure and the cascade relations from [28].

Although, at present, a proof of the conjecture that SLE_6 is the scaling limit of critical percolation cluster boundaries seems out of reach, this conjecture does lead one to believe that SLE_6 must satisfy a "locality" property; namely, it is not affected by the boundary of a domain when it is in the interior. This locality property for SLE_6 is stated more precisely and proved in §2. It is worthwhile to note that the locality property does not hold for the SLE_κ-processes when $\kappa\neq 6$.

In §3, we prove that SLE_6 satisfies a generalization of Cardy's formula for percolation crossings probabilities. From this, exponents associated with the SLE_6-process are computed.

In §4, universality ideas from [29] are used to compute the half-plane Brownian exponents from the SLE_6-exponents, which completes the proof of Theorem 1.1.

In a final short §5, the conjectured relationship between SLE_6 and critical percolation is discussed. It is demonstrated that this conjecture implies a formula from the physics literature [13], [8], [3] for the exponents corresponding to the event that there are k disjoint percolation crossings of a long rectangle.

In the subsequent papers [25], [26], [27], we determine the exponents in the full plane and the remaining half-plane exponents. In particular, we prove that $\zeta=\frac{5}{8}$, and also establish Mandelbrot's conjecture that the Hausdorff dimension of the frontier of planar Brownian motion is $\frac{4}{3}$.

It might be worthwhile to explain why the Brownian intersection exponents are accessible through SLE_6, but are difficult to compute directly. In a way, the SLE_6-process is simpler, since K_t continuously grows from its outer boundary. This means that when studying its evolution, one can essentially forget its interior, and only keep track of the exterior of K_t. By conformal invariance, this reduces problems to finitely many dimensions. The situation with planar Brownian motion is completely different, since it may enter holes it has surrounded and emerge to the exterior someplace else.

Many computations with SLE_κ are readily convertible to PDE problems, and in the presence of enough symmetry, some variables can often be eliminated, converting the PDE to an ODE.

2. SLE_6 and its locality property

2.1. The definition of chordal SLE_κ and some basic properties

Let $(\beta_t, t\geqslant 0)$ be a standard real-valued Brownian motion starting at $\beta_0=0$, let $\kappa>0$, and let $W_t^\kappa=\beta_{\kappa t}$. Consider the ordinary differential equation

$$\partial_t g_t(z)=\frac{-2}{W_t^\kappa-g_t(z)} \tag{2.1}$$

with $g_0(z)=z$. For every $z_0\in\mathbf{H}$ and every $T>0$, either there is a solution of (2.1) for $t\in[0,T]$ and for all z in a neighborhood of z_0, or there is some $t_0\in(0,T]$ such that the solution exists for $t\in[0,t_0)$ and $\lim_{t\nearrow t_0} g_t(z)=W_{t_0}^\kappa$. Let D_T be the (open) set of $z\in\mathbf{H}$ such that the former is true, and let K_T be the set of $z\in\mathbf{H}$ such that the latter holds. By considering the inverse flow $\partial_t G_t(z)=2(W_{T-t}^\kappa-G_t(z))^{-1}$, it is easy to see that $g_t(D_t)=\mathbf{H}$, and that $g_t: D_t\to\mathbf{H}$ is conformal. The process g_t, $t\geqslant 0$, will be called the *chordal stochastic Löwner evolution process* with parameter κ, or just SLE_κ; see [42]. In [42], a variation of this process, which we now call *radial* SLE_κ was also studied. In the current paper, we will not use radial SLE_κ, and therefore the word "chordal" will usually be omitted. (However, radial SLE_κ plays a major role in a subsequent paper [25].) The set $K_t=\mathbf{H}\setminus D_t$ will be called the *hull* of the SLE. The process W_t^κ will be called the *driving process* of the SLE.

It is easy to verify that each of the maps g_t satisfies the *hydrodynamic normalization* at infinity:

$$\lim_{z\to\infty} g(z)-z=0. \tag{2.2}$$

Remarks. It will be shown [40] that for all $\kappa\neq 8$ the hull K_t of SLE_κ is a.s. generated by a path. More precisely, a.s. the map $\gamma(t):=g_t^{-1}(W_t^\kappa)$ is a well-defined continuous path in $\overline{\mathbf{H}}$, and for every $t\geqslant 0$ the domain D_t is the unbounded connected component of $\mathbf{H}\setminus\gamma([0,t])$. There, it will also be shown that when $\kappa\leqslant 4$ a.s. K_t is a simple path for all $t>0$. This is not the case when $\kappa>4$ [42]. However, these results will not be needed for the current paper or for [25], [26], [27].

Löwner [32] considered the equation

$$\partial_t g_t(z)=g_t(z)\frac{\zeta(t)+g_t(z)}{\zeta(t)-g_t(z)},$$

with $g_0(z)=z$, where z is in the unit disk, and $\zeta(t)$ is a parameter. He used this equation in the study of extremal problems for classes of normalized conformal mappings. In Löwner's differential equation, the maps g_t satisfies $g_t(0)=0$. The equation (2.1) is an analogue of Löwner's equation in the half-plane, where the boundary point ∞ is fixed instead of 0, and $\zeta(t)$ is chosen to be scaled Brownian motion.

Marshall and Rohde [36] study conditions on $\zeta(t)$ which imply that K_t is a simple path.

We now note some basic properties of SLE_κ.

PROPOSITION 2.1. (i) [*Scaling*] SLE_κ *is scale-invariant in the following sense. Let* K_t *be the hull of* SLE_κ, *and let* $\alpha>0$. *Then the process* $t\mapsto\alpha^{-1/2}K_{\alpha t}$ *has the same law as* $t\mapsto K_t$.

(ii) [*Stationarity*] *Let* g_t *be an* SLE_κ*-process in* $\mathbf{H}$, *driven by* W_t^κ, *and let* τ *be any stopping time. Set* $\tilde{g}_t(z)=g_{\tau+t}\circ g_\tau^{-1}(z+W_\tau^\kappa)-W_\tau^\kappa$. *Then* $\tilde{g}_t$ *is an* SLE_κ*-process in* $\mathbf{H}$ *starting at* 0, *which is independent from* $\{g_t : t\in[0,\tau]\}$.

Proof. (i) If K_t is driven by W_t^κ, then $\alpha^{-1/2}K_{\alpha t}$ is driven by $\alpha^{-1/2}W_{\alpha t}^\kappa$, which has the same law as W_t^κ.

(ii) The process $\tilde{g}_t$ is driven by $W_{t+\tau}^\kappa-W_\tau^\kappa$. □

We now consider the definition of SLE_κ in domains other than $\mathbf{H}$.

Let $f: D\to\mathbf{H}$ be a conformal homeomorphism from some simply-connected domain D. Let f_t be the solution of (2.1) with $f_0(z)=f(z)$. Then $(f_t, t\geqslant 0)$ will be called the SLE_κ in D starting at f. If g_t is the solution of (2.1) with $g_0(z)=z$, then we have $f_t=g_t\circ f$. If K_t is the hull associated to g_t, then the hull associated with f_t is just $f^{-1}(K_t)$.

Suppose that ∂D is a Jordan curve in $\mathbf{C}$, and let $a,b\in\partial D$ be distinct. Then we may find such an $f: D\to\mathbf{H}$ with $f(a)=0$ and $f(b)=\infty$. Let K_t^f be the SLE_κ-hull associated with the SLE_κ-process starting at f. If $f^*: D\to\mathbf{H}$ is another such map with $f^*(a)=0$ and $f^*(b)=\infty$, then $f^*(z)=\alpha f(z)$ for some $\alpha>0$. By Proposition 2.1, the corresponding SLE_κ-hull $K_t^{f^*}$ has the same law as a linear time change of K_t^f. This makes it natural to consider K_t^f as a process from a to b in D, and to ignore the role of f. However, when D is not a Jordan curve, some care may be needed since the conformal map f does not necessarily extend continuously to the boundary. Partly for that reason, we have chosen to stress the importance of the conformal parameterization f.

2.2. The locality property

The main result of this section can be loosely described as follows: an SLE_6 process does not feel where the boundary of the domain lies as long as it does not hit it. This is consistent with the conjecture [42] that the SLE_6-process is the scaling limit of percolation cluster boundaries, which is explained in §5. This restriction property can therefore be viewed as additional evidence in favor of this conjecture. This feature is special to SLE_6; it is not shared by SLE_κ when $\kappa \neq 6$.

Such properties were studied in [29] and called "complete conformal invariance" (when combined with a conformal invariance property). As pointed out there, all processes with complete conformal invariance have closely related intersection exponents.

Let us first state a general local version of this result. We will say that the path γ is nice if it is a continuous simple path $\gamma:[0,1]\to\overline{\mathbf{H}}$ such that $\gamma(0),\gamma(1)\in\mathbf{R}\setminus\{0\}$ and $\gamma(0,1)\subset\mathbf{H}$. We then call the connected component $N=N(\gamma)$ of $\mathbf{H}\setminus\gamma[0,1]$ such that $0\in\partial N$ a nice neighborhood of 0 in $\mathbf{H}$. Note that N can be bounded or unbounded, depending on the sign of $\gamma(0)\gamma(1)$. When N is a nice neighborhood of 0, one can find a conformal homeomorphism $\psi=\psi_N$ from N onto $\mathbf{H}$ such that $\psi(0)=0$, $\psi'(0)=1$ and $\psi^{-1}(\infty)$ is equal to ∞ if N is unbounded and to $\gamma(1)$ if N is bounded.

THEOREM 2.2 (locality). *Let $f:D\to\mathbf{H}$ be a conformal homeomorphism from a domain $D\subset\mathbf{C}$ onto $\mathbf{H}$. Suppose that N is a nice neighborhood of 0 in $\mathbf{H}$. Define $D^*=f^{-1}(N)$ and let f^* be the conformal homeomorphism $\psi_N\circ f$ from D^* onto $\mathbf{H}$. Let $K_t\subset D$ be the hull of SLE_6 starting at f, and let $\tau:=\sup\{t:\overline{K}_t\cap\partial D^*\cap D=\varnothing\}$. Let K_t^* denote SLE_6 in D^* started at f^*, and let $\tau^*:=\sup\{t:\overline{K}_t^*\cap\partial D^*\cap D=\varnothing\}$.*

Then the law of $(K_t, t<\tau)$ is that of a time change of $(K_t^*, t<\tau^*)$.

Note that in this theorem, we have not made any regularity assumption on the boundary of the domain D.

A consequence of this result is that, modulo time change, one can define the hull of SLE_6 in a non-simply-connected domain with finitely many boundary components, since such a domain looks locally like a simply-connected domain.

This property implies the following "global" restriction properties. For convenience only, we will state them under some assumptions on the boundaries of the domains.

COROLLARY 2.3 (splitting property). *Let D denote a simply-connected domain such that ∂D is a Jordan curve. Let a, b and b' denote three distinct points on ∂D, and let I denote the connected component of $\partial D\setminus\{b,b'\}$ that does not contain a. Let $(K_t, t\geqslant 0)$ (resp. K_t') denote an SLE_6 in D from a to b (resp. from a to b'). Let T (resp. T') denote the first time at which $\overline{K}_t$ (resp. $\overline{K}_t'$) intersects I. Then $(K_t, t<T)$ and $(K_t', t<T')$ have the same law up to time change.*

COROLLARY 2.4 (restriction property). *Let $D^* \subset D$ denote two simply-connected domains, and assume that ∂D is a Jordan curve. Suppose that $I := \partial D^* \setminus \partial D$ is connected. Take two distinct points a and b in $\partial D \cap \partial D^* \setminus \bar{I}$.*

Let $(K_t, t \geqslant 0)$ denote SLE_6 *from a to b in D, and $T := \sup\{t : \bar{K}_t \cap I = \varnothing\}$. Similarly, let $(K_t^*, t \geqslant 0)$ be* SLE_6 *from a to b in D^*, and $T^* := \sup\{t : \bar{K}_t^* \cap I = \varnothing\}$. Then, $(K_t, t < T)$ and $(K_t^*, t < T^*)$ have the same law up to time change.*

In the present paper, we will use these results when D is a rectangle.

Proof of Corollary 2.3 (assuming Theorem 2.2). This is just a consequence of the fact that in Theorem 2.2 with $D = \mathbf{H}$ and bounded N, one can replace γ by $\beta(s) := \gamma(1-s)$. Then we get that the law of SLE_6 in N from 0 to $\gamma(0)$ is that of a time change of SLE_6 in N from 0 to $\gamma(1)$ up to their hitting times of γ. The result in a general domain follows by mapping it conformally onto a nice neighborhood N with a mapped to 0 and $\{b, b'\}$ to $\{\gamma(0), \gamma(1)\}$. □

Proof of Corollary 2.4 (assuming Theorem 2.2). By approximation, it suffices to consider the case where I is a simple path. Let f denote a conformal map from D onto $\mathbf{H}$, with $f(a) = 0$ and $f(b) = \infty$. Define γ in such a way that $\gamma[0, 1] = f(I)$; note that $D^* = f^{-1}(N(\gamma))$. As $b \in \partial D^* \setminus I$, $N(\gamma)$ is unbounded. Hence, by Theorem 2.2, the law of SLE_6 in D from a to b stopped when it hits I is (up to time change) the same as that of SLE_6 in D^* from a to b stopped when its closure hits I. □

In order to prove Theorem 2.2, we will establish

LEMMA 2.5. *Under the assumptions of Theorem* 2.2, *define for any fixed $s < 1$, $L_s = \gamma(0, s]$ and*

$$T = \sup\{t \geqslant 0 : \bar{K}_t \cap \bar{L}_s = \varnothing\}.$$

For all $t \leqslant T$, let $g_{s,t}$ denote the conformal homeomorphism taking $\mathbf{H} \setminus (K_t \cup L_s)$ onto $\mathbf{H}$ with the hydrodynamic normalization. Then, the process $(g_{s,t}, t < T)$ has the same law as a time change of SLE_6 *in $\mathbf{H} \setminus L_s$ starting at $g_{s,0} - g_{s,0}(0)$, up to the time when the closure of its hull intersects $\bar{L}_s$.*

Proof of Theorem 2.2 (assuming Lemma 2.5). In the setting of the lemma, let

$$h_{s,t}(z) = \frac{g_{s,t}(z) - g_{s,0}(0)}{g'_{s,0}(0)}.$$

By Lemma 2.5 and Proposition 2.1, $t \mapsto h_{s,t}$ has the same law as a time change of SLE_6 starting at $h_{s,0}$. Note that $h_{s,0}(0) = 0$, $h'_{s,0}(0) = 1$. Hence, it follows easily that for all $z \in N(\gamma)$,

$$\lim_{s \to 1} h_{s,0}(z) = \psi_N(z).$$

By continuity, if we let $h_{1,t}=\lim_{s\to 1} h_{s,t}$, then $t\mapsto h_{1,t}$ has the same law as a time-changed SLE_6 (in N) started from ψ_N. The proof is completed by noting that the hull of $h_{1,t}$ is K_t. □

The idea in the proof of Lemma 2.5 is to study how the process $g_{s,t}$ changes as s increases. For this, we will need to use some of the properties of solutions to (2.1) where W^κ is replaced by other continuous functions, and to study how (deterministic) families of conformal maps can be represented in this way with some driving function.

2.3. Deterministic expanding hulls

2.3.1. *Definition and first properties.* If $(U_t, t\in[0,a])$ is a continuous real-valued function, then the process defined by

$$\partial_t g_t(z) = \frac{-2}{U_t - g_t(z)} \tag{2.3}$$

and $g_0(z)=z$ will be called the *Löwner evolution with driving function* U_t. Note that g_t satisfies the hydrodynamic normalization (2.2). Moreover,

$$g_t(z) = z + 2tz^{-1} + a_2(t)z^{-2} + \dots, \quad z\to\infty, \tag{2.4}$$

for some functions $a_j(t)$, $j=2,3,\dots$. As above, we let $D_t\subset\mathbf{H}$ denote the domain of g_t, and let $K_t:=\mathbf{H}\setminus D_t$. K_t will be called the *expanding hull* of the process g_t.

We now address the question of which processes K_t can appear as the expanding hull driven by a continuous function U_t. We say that a bounded set $K\subset\mathbf{H}$ is a *hull* if $\mathbf{H}\setminus K$ is open and simply-connected. The Riemann mapping theorem tells us that for each hull K, there is a unique conformal homeomorphism $g_K:\mathbf{H}\setminus K\to\mathbf{H}$ which satisfies the hydrodynamic normalization (2.2). Let

$$A(K) = A(g_K) := \tfrac{1}{2}\lim_{z\to\infty} z(g_K(z)-z);$$

that is, $g(z)=z+2A(g)z^{-1}+\dots$ near ∞. Observe that $A(K)$ is real, because $g_K(x)$ is real when $x\in\mathbf{R}$ and $|x|$ is sufficiently large. Moreover, $A(K)\geqslant 0$, because $\mathrm{Im}(z-g_K(z))$ is a harmonic function which vanishes at infinity and has non-negative boundary values. Note that

$$A(g\circ h) = A(g)+A(h)$$

if g and h satisfy the hydrodynamic normalization. It follows that $A(K)\leqslant A(L)$ when $K\subset L$, since $g_L=g_{g_K(L\setminus K)}\circ g_K$.

The quantity $A(g)$ is similar to capacity, and plays an analogous role for the equation (2.1) as capacity plays for Löwner's equation.

THEOREM 2.6. *Let $(K_t, t\in[0,a])$ be an increasing family of hulls. Then the following are equivalent:*

(1) *For all $t\in[0,a]$, $A(K_t)=t$, and for each $\varepsilon>0$ there is a $\delta>0$ such that for each $t\in[0,a-\delta]$ there is a bounded connected set $S\subset\mathbf{H}\setminus K_t$ with $\operatorname{diam}(S)<\varepsilon$ and such that S disconnects $K_{t+\delta}\setminus K_t$ from infinity in $\mathbf{H}\setminus K_t$.*

(2) *There is some continuous $U:[0,a]\to\mathbf{R}$ such that K_t is driven by U_t.*

In [37] a similar theorem is proved for Löwner's differential equation in the disk.

Note that K_t may change discontinuously, in the Hausdorff metric, as t increases. For example, consider $\widehat{K}_t:=\{\exp(is):0<s\leqslant t\}$ when $t<\pi$, and $\widehat{K}_\pi:=\{z\in\mathbf{H}:|z|\leqslant 1\}$ and $\widehat{K}_{t+\pi}:=K_\pi\cup(-1,-1+it]$, $t>0$, say, and let $K_t:=\widehat{K}_{\phi(t)}$ where ϕ is chosen to satisfy $A(K_{\phi(t)})=t$.

LEMMA 2.7. *Let $r>0$ and $x_0\in\mathbf{R}$, and suppose that K is a hull contained in the disk $\{z:|z-x_0|<r\}$. Then*

$$\left|g_K^{-1}(z)-z+\frac{2A(K)}{z-x_0}\right|\leqslant\frac{CrA(K)}{|z-x_0|^2}$$

for all $z\in\mathbf{H}$ with $|z-x_0|>Cr$, where $C>0$ is an absolute constant.

Proof of Lemma 2.7. For notational simplicity, we assume that $x_0=0$. Clearly, this does not entail any loss of generality. By approximation, we may assume that K has smooth boundary. Let $I\subset\mathbf{R}$ be the smallest interval in $\mathbf{R}$ containing $\{g_K(x):x\in\partial K\cap\mathbf{H}\}$, and let $f:=g_K^{-1}$. Let f_I be the restriction of f to I. Let f^* denote the extension of f to $\mathbf{C}\setminus I$, by Schwarz reflection. The Cauchy formula gives

$$2\pi i f^*(w)=\int_{|z|=R}\frac{f^*(z)}{z-w}\,dz+\int_I\frac{f_I(x)-\overline{f_I(x)}}{x-w}\,dx,$$

provided that $R>|w|$, $R>\max\{|x|:x\in I\}$ and $w\in\mathbf{C}\setminus I$. Since $f^*(z)=z-2A(K)z^{-1}+\ldots$ near ∞,

$$\lim_{R\to\infty}\int_{|z|=R}\frac{f^*(z)}{z-w}\,dz=\lim_{R\to\infty}\int_{|z|=R}\frac{z}{z-w}\,dz=2\pi i w.$$

Consequently, we have

$$f^*(w)-w=\frac{1}{\pi}\int_I\frac{\operatorname{Im}(f_I(x))}{x-w}\,dx.$$

Multiplying by w and taking $w\to\infty$ gives

$$A(g_K)=-A(f^*)=\frac{1}{2\pi}\int_I\operatorname{Im}(f_I(x))\,dx. \tag{2.5}$$

Moreover,

$$f^*(w)-w-2A(f^*)w^{-1}=\frac{1}{\pi}\int_I \mathrm{Im}(f_I(x))\left(\frac{1}{x-w}+\frac{1}{w}\right)dx,$$

and therefore

$$|f^*(w)-w-2A(f^*)w^{-1}|\leqslant\frac{1}{\pi}\int_I \mathrm{Im}(f_I(x))\sup\{|(x-w)^{-1}+w^{-1}|:x\in I\}\,dx$$

$$=-2A(f^*)\sup\left\{\left|\frac{x}{(x-w)w}\right|:x\in I\right\}.$$

Hence, the proof will be complete once we demonstrate that there is some constant c_0 such that $I\subset[-c_0r,c_0r]$. This is easily done, as follows. Define $G(z):=g_K(rz)/r$ for $|z|>1$, and write $G(z)=z+a_1z^{-1}+a_2z^{-2}+\dots$. The Area Theorem (see, e.g., [41]) gives $1\geqslant\sum_{j=1}^{\infty}j|a_j|^2$. In particular, $|a_j|\leqslant1$ for $j\geqslant1$. Consequently, we have $|G(z)-z|\leqslant1$ for $|z|\geqslant2$. By Rouché's theorem (e.g. [41]), it follows that $G(\{|z|\geqslant2\})\supset\{|z|>3\}$. Consequently, $g_K(\mathbf{H}\setminus K)\supset\{|z|>3r\}$, which gives $I\subset[-3r,3r]$. □

For convenience, we adopt the notation

$$K_{t,u}:=g_{K_t}(K_{t+u}\setminus K_t).$$

Proof of Theorem 2.6. We start with (1) implies (2). Let $R:=\sup\{|z|:z\in K_a\}$ and $Q:=\{z\in\mathbf{H}:|z|>R+2\}$. Let t, δ, ε and S be as in the statement of the theorem, and let $s\in\partial S$. Suppose that $\varepsilon<1$ and $r\in[\varepsilon,\sqrt{\varepsilon}]$. Then there is an arc β_r of the circle of radius r about s such that $\beta_r\subset\mathbf{H}\setminus K_t$ and $K_t\cup\mathbf{R}\cup\beta_r$ separates $K_{t+\delta}\setminus K_t$ from Q. It therefore follows that the extremal length of the set of arcs in $\mathbf{H}\setminus K_t$ which separate $K_{t+\delta}\setminus K_t$ from Q in $\mathbf{H}\setminus K_t$ is at most $\mathrm{const}/\log(1/\varepsilon)$. (For the definition and basic properties of extremal length, see [1], [31]. The terms extremal length and extremal distance have the same meaning.) Since extremal length is invariant under conformal maps, it follows that the extremal length of the set of paths in $\mathbf{H}$ that separate $K_{t,\delta}$ from $g_{K_t}(Q)$ is at most $\mathrm{const}/\log(1/\varepsilon)$. Because the diameter of $g_{K_t}(\mathbf{H}\setminus Q)$ is bounded by some function of R (this follows since g_{K_t} has the hydrodynamic normalization), we conclude that at least one of these arcs has length less than $\mathrm{const}/\log(1/\varepsilon)$. Consequently, this is a bound on the diameter of $K_{t,\delta}$. Observe that this bound is uniform for all $t\in[0,a-\delta]$. For each $t<a$, we then define U_t to be the point in the intersection $\bigcap_{u>0}\overline{K}_{t,u}$. We have an upper bound on $\mathrm{diam}(K_{t,\delta})$ which tends to zero uniformly as $\delta\to0$, and therefore $\lim_{\delta\to0}g_{K_{t,\delta}}(z)-z=0$ uniformly for $z\in\mathbf{H}\setminus K_{t,\delta}$ and $t\leqslant a-\delta$. This implies that U_t is uniformly continuous on $[0,a)$ and can be extended continuously to $[0,a]$.

Now let $z_0\in\mathbf{H}\setminus K_a$. Then there is some $c>0$ such that $\mathrm{Im}(g_{K_t}(z_0))>c$ for all $t\in[0,a]$. Lemma 2.7 applied with $K=K_{t,u}$, $z=g_{K_{t+u}}(z_0)$ and $x_0=U_{t+u}$ gives

$$\frac{g_{K_t}(z_0)-g_{K_{t+u}}(z_0)}{u}+\frac{2}{g_{K_{t+u}}(z_0)-U_{t+u}}\to0$$

as $\delta \to 0$. As $g_{K_t}(z_0)$ and U_t are continuous in t, we may therefore conclude that

$$\partial_t g_{K_t}(z_0) = \frac{2}{g_{K_t}(z_0) - U_t},$$

which gives (2).

The proof that (2) implies (1) is easy. Let $\varepsilon > 0$. Given $0 \leqslant t \leqslant t+u < a$, let $\varrho(t,u) := u + \max\{U(t') - U(t'') : t', t'' \in [t, t+u]\}$. Observe that $\operatorname{diam}(K_{t,u}) \to 0$ if $\varrho(t,u) \to 0$, and $\varrho(t,u) \to 0$ if $u \to 0$. Consequently, the extremal length of the set of paths separating $K_{t,u}$ from $\{z \in \mathbf{H} : |z| > 1\}$ in $\mathbf{H}$ goes to zero as $\varrho(t,u) \to 0$. This implies that there is a path β in this set such that $\operatorname{diam}(g_{K_t}^{-1}(\beta)) < \varepsilon$, provided that u is small. We then just take $S = g_{K_t}^{-1}(\beta)$. □

2.3.2. *Time-modified expanding hulls and restriction.* Let $(K_t, t \in [0,a])$ denote a family of hulls, and suppose that there is a monotone increasing homeomorphism ϕ: $[0,a] \to [0,\hat{a}]$ such that $(K_{\phi(t)}, t \in [0,\hat{a}])$ is an expanding hull driven by some function $t \mapsto \hat{U}_t$. If additionally ϕ is continuously differentiable in $[0,a]$ and $\phi'(t) > 0$ for each $t \in [0,a]$, then we call $(K_t, t \in [0,a])$ a *time-modified* expanding hull, with driving function $U_t := \hat{U}_{\phi^{-1}(t)}$. Note that, in this case, $\phi^{-1}(t) = A(K_t)$, and that

$$\partial_t g_{K_t}(z) = \frac{2\partial_t A(K_t)}{g_{K_t}(z) - U_t}. \tag{2.6}$$

Note that in our terminology, an expanding hull is always a time-modified expanding hull.

LEMMA 2.8. *Let $(K_t, t \in [0,a])$ be a time-modified expanding hull, with driving function $(U_t, t \in [0,a])$. Let D be a relatively open subset of $\overline{\mathbf{H}}$ which contains $\overline{K}_a$, and set $D_{\mathbf{R}} := D \cap \mathbf{R}$. Let $G : D \to \overline{\mathbf{H}}$ be conformal in $D \setminus D_{\mathbf{R}}$ and continuous in D, and suppose that $G(D_{\mathbf{R}}) \subset \mathbf{R}$. Then $(G(K_t), t \in [0,a])$ is a time-modified expanding hull. Moreover,*

$$\partial_t A(G(K_t)) = G'(U_0)^2 \partial_t A(K_t) \quad \text{at } t = 0. \tag{2.7}$$

Proof. We first prove (2.7). The proof will be based on (2.5). Note first that if $K' = aK$ then $A(K') = a^2 A(K)$. Therefore, we may assume that $G'(U_0) = 1$. Similarly, with no loss of generality, we assume that $U_0 = G(U_0) = 0$. By the reflection principle, G is analytic in D.

Set $\hat{K}_t := G(K_t)$. Let $I_t \subset \mathbf{R}$ be the interval corresponding to $\overline{\partial K_t \cap \mathbf{H}}$ under g_{K_t}, and let $\hat{I}_t$ be the interval corresponding to $\overline{\partial \hat{K}_t \cap \mathbf{H}}$ under $g_{\hat{K}_t}$. Let $\varepsilon > 0$, and let $D_\varepsilon := \{z \in D : |1 - G'(z)| < \varepsilon\}$. Let β be some arc in $D_\varepsilon \setminus \{0\}$ that separates 0 from ∞ in $\overline{\mathbf{H}}$. Consider the map

$$h_t = g_{\hat{K}_t} \circ G \circ g_{K_t}^{-1}.$$

It is well defined in a neighborhood of I_t provided that $K_t \cap \beta = \varnothing$ (for instance), and this holds when t is small. This map may be continued analytically by reflecting in the real axis, and therefore the maximum principle implies that

$$\sup\{h_t'(x) : x \in I_t\} \leqslant \sup\{|h_t'(z)| : z \in g_{K_t}(\beta)\}$$

when $K_t \cap \beta = \varnothing$. Note that $g_{K_t}(z) - z \to 0$ and $g_{\widehat{K}_t}(z) - z \to 0$ as $t \searrow 0$, and therefore $g_{K_t}'(z) \to 1$ and $g_{\widehat{K}_t}'(z) \to 1$ on β. Consequently, for small t we have

$$\sup\{h_t'(x) : x \in I_t\} < 1 + 2\varepsilon. \tag{2.8}$$

Note that for z close to $U_0 = 0$ we have $\operatorname{Im}(G(z)) \leqslant (1+\varepsilon)\operatorname{Im}(z)$. Using (2.5), this inequality and (2.8), we get

$$\begin{aligned} A(G(K_t)) &= \frac{1}{2\pi} \int_{\hat{I}_t} \operatorname{Im}(g_{\widehat{K}_t}^{-1}(x))\, dx \\ &= \frac{1}{2\pi} \int_{I_t} \operatorname{Im}(G \circ g_{K_t}^{-1}(x)) h_t'(x)\, dx \\ &\leqslant \frac{1}{2\pi} \int_{I_t} (1+\varepsilon) \operatorname{Im}(g_{K_t}^{-1}(x))(1+2\varepsilon)\, dx \\ &= (1+\varepsilon)(1+2\varepsilon) A(K_t) \end{aligned}$$

for small $t > 0$. (Note that $g_{K_t}^{-1}(x)$ is not defined for every $x \in I_t$, but it is defined for almost every $x \in I_t$.) By symmetry, we also have a similar inequality in the other direction. This proves (2.7).

By Theorem 2.6, to show that $G(K_t)$ is a time-modified expanding hull, it suffices to show that $A(G(K_t))$ is continuously differentiable in t, with derivative bounded away from 0. Let $G_t := g_{\widehat{K}_t} \circ G \circ g_{K_t}^{-1}$. Then G_t is analytic in $g_{K_t}(D \setminus K_t)$ and depends continuously on t. Hence $G_t'(U_t)$ is continuous in t. Since $A(G(K_{t+u})) = A(G(K_t)) + A(G_t(K_{t,u}))$, it follows that $\partial_t A(G(K_t)) = G_t'(U_t)^2 \partial_t A(K_t)$, which completes the proof. □

For future reference, we note that when $g_t = g_{K_t}$ satisfies the differential equation (2.6), we have the formula

$$\partial_t \log g_t'(z) = -\frac{2\partial_t A(g_t)}{(g_t(z) - U_t)^2}, \tag{2.9}$$

which is obtained by differentiating (2.6) with respect to z.

2.3.3. *Pairs of time-modified expanding hulls.* We now discuss the situation where there are two disjoint expanding hulls.

Let $(L_s, s\in[0,s_0])$ and $(K_t, t\in[0,t_0])$ be a pair of time-modified expanding hulls such that $\bar{L}_{s_0}\cap\bar{K}_{t_0}=\varnothing$. Let $g_{s,t}:=g_{L_s\cup K_t}$, $g_t:=g_{K_t}$, $\hat{g}_s:=g_{L_s}$ and $a(s,t):=A(g_{s,t})$. Then for each $s\in[0,s_0]$ and $t\in[0,t_0]$ we have

$$g_{s,t}=g_{g_t(L_s)}\circ g_t.$$

Therefore,

$$\partial_s g_{s,t}(z)=\frac{2\partial_s a(s,t)}{g_{s,t}(z)-U^1(s,t)},$$

where $s\mapsto U^1(s,t)$ is the driving function for the time-modified expanding hulls $s\mapsto g_t(L_s)$. Similarly,

$$\partial_t g_{s,t}(z)=\frac{2\partial_t a(s,t)}{g_{s,t}(z)-U^2(s,t)},$$

where $t\mapsto U^2(s,t)$ is the driving function for the time-modified expanding hulls $t\mapsto\hat{g}_s(K_t)$. Although we do not know that $g_t^{-1}(U^2(0,t))$ is well defined, $g_{s,t}\circ g_t^{-1}$ is analytic in a neighborhood of $U^2(0,t)$, by the reflection principle. Hence, it is clear that $U^2(s,t)=g_{s,t}\circ g_t^{-1}(U^2(0,t))$ (see, for example, the construction of U_t in the proof of Theorem 2.6), and therefore

$$\partial_s U^2(s,t)=\frac{2\partial_s a(s,t)}{U^2(s,t)-U^1(s,t)}. \tag{2.10}$$

We will now prove the formula

$$\partial_s\partial_t a(s,t)=\frac{-4\partial_s a(s,t)\partial_t a(s,t)}{(U^2(s,t)-U^1(s,t))^2}. \tag{2.11}$$

From (2.7) we have

$$\partial_t a(0,s)=\hat{g}_s'(U^2(0,0))^2\partial_t a(0,0)=\hat{g}_s'(U^2(0,0))^2,$$

and using (2.9), we obtain

$$\partial_s\log\partial_t a(0,s)=\frac{-4\partial_s A(\hat{g}_s)}{(\hat{g}_s(U^2(0,0))-U^1(s,0))^2}=\frac{-4\partial_s a(0,s)}{(U^2(s,0)-U^1(s,0))^2}.$$

This verifies (2.11) for the case $t=0$. The general case is similarly obtained.

2.4. Proof of Lemma 2.5

We will now prove Lemma 2.5; this is the core of the proof of the locality property. We assume that $\gamma:[0,s_1]\to\overline{\mathbf{H}}$ is a continuous simple path with $\gamma(0)\in\mathbf{R}\setminus\{0\}$ and $L_s:=\gamma(0,s_1]\subset\mathbf{H}$. With no loss of generality, assume that γ is parameterized so that $A(L_s)=s$. By Theorem 2.6, $(L_s, s\in[0,s_1])$ is a time-modified expanding hull, and by Lemma 2.8, for each s, $t\mapsto g_{L_s}(K_t)$ is a time-modified expanding hull, and for each t, $s\mapsto g_{K_t}(L_s)$ is a time-modified expanding hull. Let $t\mapsto W(s,t)$ be the process driving $t\mapsto g_{L_s}(K_t)$, let $U(s,t)$ be the process driving $s\mapsto g_{K_t}(L_s)$, and let $Y_t=W(0,t)$ be the process driving K_t. As above, let $a(s,t)=A(g_{s,t})$. For simplicity, $W(s,t)$ will be abbreviated to W, $U(s,t)$ to U, $a(s,t)$ to a, etc.

Our aim is to show that $(W(s_1,t), t\geqslant 0)$ is a continuous martingale (up to the stopping time T), and that its quadratic variation (for background on stochastic calculus, see, e.g., [15], [39]) is

$$\langle W(s_1,\cdot)\rangle_t = 6(a(s_1,t)-a(s_1,0)).$$

Indeed, if this is true, let $\phi(t)$ be the inverse of the map $t\mapsto a(s_1,t)-a(s_1,0)$, and define $\widetilde{W}(t)=W(s_1,\phi(t))$. Then $\widetilde{W}(\frac{1}{6}t)$ is a Brownian motion, so that $t\mapsto g_{s_1,0}(K_{\phi(t)})-g_{s_1,0}(0)$ is an SLE_6-process, as required. Note that this will in fact give a precise expression for the time change in Lemma 2.5 and Theorem 2.2.

Before giving the mathematically rigorous proof, we first present a formal, non-rigorous derivation of the fact that $W(s,\cdot)$ is a martingale. In this derivation, κ will be kept as a variable, in order to stress where the assumption $\kappa=6$ plays a role (it will not be so apparent in our proof).

Non-rigorous argument. The first goal is to show that the quadratic variation $\langle W\rangle_t$ of $t\mapsto W(s,t)$ satisfies

$$\partial_t\langle W\rangle_t = \kappa\partial_t a, \tag{2.12}$$

for each s,t. It is clear that this holds when $s=0$, since K_t is SLE_6. We have

$$\begin{aligned}
\partial_s\partial_t\langle W\rangle_t &= \partial_t\partial_s\langle W\rangle_t \\
&= 2\partial_t\langle\partial_s W, W\rangle_t \\
&= 2\partial_t\langle 2(\partial_s a)(W-U)^{-1}, W\rangle_t \quad \text{(by (2.10))} \\
&= -4(\partial_s a)(W-U)^{-2}\partial_t\langle W\rangle_t \quad \text{(by Itô's formula)} \\
&= (\partial_t a)^{-1}(\partial_s\partial_t a)\partial_t\langle W\rangle_t \quad \text{(by (2.11))}.
\end{aligned}$$

Consequently,

$$(\partial_t a)^2\partial_s(\partial_t\langle W\rangle_t/\partial_t a) = \partial_t a\,\partial_s\partial_t\langle W\rangle_t - \partial_t\langle W\rangle_t\partial_s\partial_t a = 0,$$

which means that $\partial_t\langle W\rangle_t/\partial_t a$ does not depend on s. Since (2.12) holds when $s=0$, this proves (2.12).

We now show that $t\mapsto W(s,t)$ is a martingale. The dt-term in Itô's formula for the ∂_t-derivative of

$$\partial_s W(s,t)=\frac{2\partial_s a}{W-U}$$

is

$$\frac{2\partial_t\partial_s a}{W-U}+2\frac{\partial_s a}{(W-U)^3}\partial_t\langle W\rangle_t-4\frac{\partial_s a}{(W-U)^3}\partial_t a,$$

where the first summand comes from differentiating $\partial_s a$, the second summand is the diffusion term in Itô's formula, and the last summand comes from differentiating with respect to U and using (2.10) for $\partial_t U$. Using (2.11) and (2.12), this becomes

$$\left(3-\tfrac{1}{2}\kappa\right)\frac{\partial_t\partial_s a}{W-U},$$

which vanishes when $\kappa=6$. Hence $t\mapsto\partial_s W(s,t)$ is a martingale. As

$$W(s,t)=Y_t+\int_0^s \partial_s W(s',t)\,ds',$$

it follows that $t\mapsto W(s,t)$ is a martingale. This completes the informal proof.

The problem with the above argument is that we do not know that $t\mapsto W(s,t)$ is a semi-martingale, and hence cannot apply stochastic calculus to it. Moreover, we need to check that there is sufficient regularity to justify the equality $\partial_s\partial_t\langle W\rangle_t=\partial_t\partial_s\langle W\rangle_t$.

To rectify the situation, set

$$V(s,t'):=W(s,0)+\int_0^{t'}\sqrt{\partial_t a(s,t)}\,dY_t.$$

Then $t\mapsto V(s,t)$ is clearly a martingale. The rest of this subsection will be devoted to the proof of the fact that $V=W$. Recall that

$$T=\sup\{t\geqslant 0:\overline{K}_t\cap\bar{L}_{s_1}=\varnothing\}.$$

We will need the following fact:

LEMMA 2.9. *There exists a continuous version of V on $[0,s_1]\times[0,T)$.*

Proof. In order to keep some quantities bounded, we have to stop the processes slightly before T. Let us fix $\varepsilon\in(0,1)$, and define

$$T_1^\varepsilon=\inf\left\{t>0:\inf_{s\leqslant s_1}|W(s,t)-U(s,t)|\leqslant\varepsilon\right\}.$$

Define for any $s \leqslant s_1$ and $t_0 \geqslant 0$

$$\widetilde{V}(s,t_0) = \widetilde{V}^{\varepsilon}(s,t_0) := \int_0^{\min(t_0, T_1^{\varepsilon})} \sqrt{\partial_t a}\, dY_t.$$

As

$$\sup_{n \geqslant 1} T_1^{1/n} = T,$$

it is sufficient to show existence of a continuous version (on $[0,s_1] \times \mathbf{R}_+$) of $\widetilde{V}$.

Let

$$\tilde{a} = \tilde{a}(s,t) := 1_{\{t \leqslant T_1^{\varepsilon}\}} \sqrt{\partial_t a}\,.$$

Note that $\partial_t a(0,t) = \partial_s a(s,0) = 1$. Hence, from (2.11) it follows that $\partial_s a \leqslant 1$ and $\partial_t a \leqslant 1$ for all $s \leqslant s_1$, $t < T$. Using (2.11) again, we get

$$|\partial_s \tilde{a}| = 1_{\{t \leqslant T_1^{\varepsilon}\}} \frac{2\sqrt{\partial_t a}\, \partial_s a}{(W-U)^2} \leqslant 2\varepsilon^{-2}.$$

Hence, for all $t \geqslant 0$, for all s, s' in $[0, s_1]$,

$$|\tilde{a}(s,t) - \tilde{a}(s',t)| \leqslant 2\varepsilon^{-2} |s - s'|.$$

But

$$\mathbf{E}[(\widetilde{V}(s,t_0) - \widetilde{V}(s',t_0'))^4] \leqslant 16\mathbf{E}\left[\left(\int_{t_0}^{t_0'} \tilde{a}(s',t)\, dY_t\right)^4\right] + 16\mathbf{E}\left[\left(\int_0^{t_0} (\tilde{a}(s,t) - \tilde{a}(s',t))\, dY_t\right)^4\right],$$

and using, for instance, the Burkholder–Davis–Gundy inequality for $p=4$ (see, e.g., [39, IV.4]), we see that there exists a constant $c_1 = c_1(\varepsilon)$ such that for all $t_0, t_0' \geqslant 0$ and $s, s' \in [0, s_1]$,

$$\begin{aligned}\mathbf{E}[(\widetilde{V}(s,t_0) - \widetilde{V}(s',t_0'))^4] &\leqslant c_1 \mathbf{E}[(t_0 - t_0')^2] + c_1 \mathbf{E}\left[\left(\int_0^{t_0} (s-s')^2\, dt\right)^2\right] \\ &\leqslant c_1 (t_0 - t_0')^2 + c_1 t_0^2 (s-s')^4,\end{aligned}$$

and the existence of a continuous version of $\widetilde{V}$ then easily follows from Kolmogorov's lemma (see, e.g., [39, I.(1.8)]). □

From now on, we will use a version of V that is continuous on $[0,s_1] \times [0,T)$. Define

$$\begin{aligned}T_2^{\varepsilon} &:= \inf\{t \geqslant 0 : \sup_{s \leqslant s_1} |V(s,t) - W(s,t)| \geqslant 1\}, \\ T_3^{\varepsilon} &:= \inf\{t \geqslant 0 : \inf_{s \leqslant s_1} |V(s,t) - U(s,t)| \leqslant \varepsilon\}, \\ T^{\varepsilon} &:= \min(T_1^{\varepsilon}, T_2^{\varepsilon}, T_3^{\varepsilon}).\end{aligned}$$

Note that for all $s \leqslant s_1$ and $t < T$,

$$\partial_s \sqrt{\partial_t a} = \frac{-2\sqrt{\partial_t a}\,\partial_s a}{(W-U)^2}.$$

The process $\partial_s \sqrt{\partial_t a}$ remains bounded before T^ϵ (uniformly in $s \leqslant s_1$), and it is a measurable function of (s,t). By Fubini's theorem for stochastic integrals ([15, Lemma III.4.1]), we have that for all $s_0 \leqslant s_1$, for all $t_0 \geqslant 0$, almost surely

$$\begin{aligned}\int_0^{s_0}\int_0^{t_0'} \frac{-2\sqrt{\partial_t a}\,\partial_s a}{(W-U)^2}\,dY_t\,ds &= \int_0^{t_0'}\left(\int_0^{s_0} \partial_s\sqrt{\partial_t a}\,ds\right) dY_t \\ &= \int_0^{t_0'} \left(\sqrt{\partial_t a(s_0,t)} - \sqrt{\partial_t a(0,t)}\,\right) dY_t \\ &= V(s_0,t_0') - V(0,t_0') - (W(s_0,0) - W(0,0)),\end{aligned} \tag{2.13}$$

where $t_0' := \min(t_0, T^\epsilon)$. On the other hand, using Itô's formula, we now compute

$$\begin{aligned}\frac{2\partial_s a(s,t_0')}{U(s,t_0')-V(s,t_0')} &= \frac{2}{U(s,0)-V(s,0)} + \int_0^{t_0'} \frac{2\partial_s a\sqrt{\partial_t a}}{(U-V)^2}\,dY_t \\ &\quad + 2\int_0^{t_0'}\left(\frac{\partial_t\partial_s a}{U-V} - \frac{\partial_s a\,\partial_t U}{(U-V)^2} + \frac{\partial_s a\,\partial_t\langle V\rangle_t}{(U-V)^3}\right) dt \\ &= -\partial_s W(s,0) + \int_0^{t_0'} \tilde{b}_1\,dY_t + \int_0^{t_0'} (V-W)b_2\,dt,\end{aligned} \tag{2.14}$$

where (using $\partial_t\langle V\rangle_t = \partial_t a\,\partial_t\langle Y\rangle_t = \kappa\partial_t a = 6\partial_t a$)

$$\tilde{b}_1(s,t) := \frac{2\partial_s a\sqrt{\partial_t a}}{(U-V)^2},$$

$$b_2(s,t) := \frac{4\partial_s a\,\partial_t a}{(U-W)^2(U-V)^2}\left(5 + 3\frac{W-V}{V-U}\right).$$

Note that for all $s \leqslant s_1$ and $t \leqslant T^\epsilon$,

$$|b_2(s,t)| \leqslant 16\epsilon^{-5}.$$

By integrating (2.14) with respect to s and subtracting (2.13) from it, we get

$$\begin{aligned}&V(s_0,t_0') - V(0,t_0') + \int_0^{s_0} \frac{2\partial_s a(s,t_0')}{U(s,t_0')-V(s,t_0')}\,ds \\ &\quad = \int_0^{s_0}\int_0^{t_0'} (V-W)b_2(s,t)\,dt\,ds \\ &\qquad + \int_0^{s_0}\left(\int_0^{t_0'} \frac{-2\sqrt{\partial_t a}\,\partial_s a}{(W-U)^2}\,dY_t + \int_0^{t_0} \tilde{b}_1(s,t)\,dY_t\right) ds \\ &\quad = \int_0^{s_0}\int_0^{t_0'} (V-W)b_2\,dt\,ds + \int_0^{s_0}\int_0^{t_0'} (V-W)b_1\,dY_t\,ds,\end{aligned} \tag{2.15}$$

where (after some simplifications)

$$b_1(s,t) := \frac{2\partial_s a\sqrt{\partial_t a}}{(V-U)^2(W-U)^2}((U-W)+(U-V)).$$

Note that for all $s\in[0,s_1]$ and $t\leqslant T^\varepsilon$,

$$|b_1(s,t)| \leqslant 4\varepsilon^{-3}.$$

But we know on the other hand that

$$W(s_0,t_0')-W(0,t_0') = \int_0^{s_0} \frac{2\partial_s a(s,t_0')}{W(s,t_0')-U(s,t_0')}\,ds. \tag{2.16}$$

Subtracting this equation from (2.15), one gets

$$V(s_0,t_0')-W(s_0,t_0') = \int_0^{s_0} b_3(s,t_0')(V(s,t_0')-W(s,t_0'))\,ds + \int_0^{s_0}\int_0^{t_0'}(V-W)b_2\,dt\,ds + \int_0^{s_0}\int_0^{t_0'}(V-W)b_1\,dY_t\,ds,$$

where

$$b_3(s,t) := \frac{-2\partial_s a}{(U-W)(U-V)}.$$

Again b_3 remains uniformly bounded before T^ε.

We now define

$$H(s,t) = V(s,t)-W(s,t).$$

Hence, for all $t_0\leqslant T$,

$$H(s_0,t_0) = \int_0^{s_0} b_3(s,t_0)H(s,t_0)\,ds + \int_0^{s_0}\int_0^{t_0} b_2 H\,ds\,dt + \int_0^{s_0}\int_0^{t_0} b_1 H\,ds\,dY_t$$

and $|b_1|, |b_2|, |b_3|$ are all bounded by some constant $c_2=c_2(\varepsilon)$ on $[0,s_1]\times[0,T^\varepsilon]$. This equation and an argument similar to Gronwall's lemma will show that $H=0$.

Let us fix $t_1>0$. For any $t\geqslant 0$, define

$$\tau(t) = \min(t,t_1,T^\varepsilon).$$

We will use the notation $\tau_0=\tau(t_0)$. It is easy to see that there exists a $c_3=c_3(\varepsilon,t_1,s_1)$ such that for all $t_0\geqslant 0$ and $s_0\in[0,s_1]$,

$$H(s_0,\tau_0)^2 \leqslant c_3\int_0^{s_0} H(s,\tau_0)^2\,ds + c_3\int_0^{\tau_0}\int_0^{s_0} H^2\,ds\,dt + c_3\left(\int_0^{\tau_0}\int_0^{s_0} b_1 H\,ds\,dY_t\right)^2.$$

The Burkholder–Davis–Gundy inequality for $p=2$ (see, e.g., [39, IV.4]) shows that there exist constants $c_4=c_4(\varepsilon,t_1,s_1)$ and $c_5=c_5(\varepsilon,t_1,s_1)$ such that for all $s\in[0,s_1]$ and $t_0\geqslant 0$,

$$\mathbf{E}\left[\sup_{u\leqslant t_0}\left(\int_0^{\tau(u)}\int_0^{s_0} b_1 H\,ds\,dY_t\right)^2\right]\leqslant c_4\mathbf{E}\left[\int_0^{\tau_0}\left(\int_0^{s_0} b_1 H\,ds\right)^2 dt\right]$$
$$\leqslant c_5\int_0^{t_0}\int_0^{s_0}\mathbf{E}[H(s,\tau(t))^2]\,ds\,dt.$$

Let us now define

$$h(s_0,t_0)=\mathbf{E}\big[\sup_{t\leqslant t_0} H(s_0,\tau(t))^2\big].$$

Then

$$h(s_0,t_0)\leqslant c_6\left(\int_0^{s_0} h(s,t_0)\,ds+\int_0^{t_0}\int_0^{s_0} h(s,t)\,ds\,dt\right).$$

We also know that $h(s,t)$ is bounded by 1 (because $|H|\leqslant 1$). Hence, it is straightforward to prove by induction that for all $s_0\in[0,s_1]$, $t_0\geqslant 0$ and $p=1,2,\ldots$,

$$h(s_0,t_0)\leqslant\frac{c_6^p s_0^p(1+t_0)^p}{p!},$$

so that $h(s_0,t_0)=0$. In particular (using the continuity of V and W), this shows that $W=V$ almost surely on all sets $[0,s_1]\times[0,\min(t_1,T^\varepsilon)]$. As this is true for all ε and t_1, we conclude that $V=W$ on $[0,s_1]\times[0,T)$. Lemma 2.5 follows, and thereby also Theorem 2.2. □

3. Exponents for SLE_6

3.1. Statement

In the present section, we are going to compute intersection exponents associated with SLE_6.

Suppose that $D\subset\mathbf{C}$ is a Jordan domain; that is, ∂D is a simple closed curve in $\mathbf{C}$. Let $a,b\in\partial D$ be two distinct points on the boundary of D. As explained in §2.1, the SLE_6 $(K_t, t\geqslant 0)$ from a to b in D is well defined, up to a linear time change.

Now suppose that $I\subset\partial D$ is an arc with $b\in I$ but $a\notin I$. Let

$$\tau_I:=\sup\{t\geqslant 0:\overline{K}_t\cap I=\varnothing\}.$$

By Corollary 2.3, up to a time change, the law of the process $(K_t, t<\tau_I)$ does not change if we replace b by another point $b'\in I$. Set

$$S=S(a,I,D):=\bigcup_{t<\tau_I} K_t,$$

and call this set the *hull from a to I in D*. It does not depend on b.

Suppose that $L>0$, and let $\mathcal{R}=\mathcal{R}(L)$ denote the rectangle with corners

$$A_1 := 0, \quad A_2 := L, \quad A_3 := L+i\pi, \quad A_4 := i\pi. \tag{3.1}$$

Let $\mathcal{S}$ denote the closure of the hull from A_4 to $[A_1, A_2]\cup[A_2, A_3]$ in $\mathcal{R}$.

In the following, we will use the terminology *π-extremal distance* instead of "π times the extremal distance". For instance, the π-extremal distance between the vertical sides of $\mathcal{R}$ in $\mathcal{R}$ is L.

When $\mathcal{S}\cap[A_1, A_2]=\varnothing$, let $\mathcal{L}$ be the π-extremal distance between $[A_1, A_4]$ and $[A_2, A_3]$ in $\mathcal{R}\setminus\mathcal{S}$. Otherwise, put $\mathcal{L}=\infty$.

In the sequel, we will use the function

$$u(\lambda) = \tfrac{1}{6}\bigl(6\lambda+1+\sqrt{24\lambda+1}\,\bigr). \tag{3.2}$$

The main goal of this section is to prove the following result.

THEOREM 3.1.

$$\mathbf{E}[1_{\{\mathcal{L}<\infty\}}\exp(-\lambda\mathcal{L})] = \exp(-u(\lambda)L+O(1)(\lambda+1)) \quad \textit{as } L\to\infty, \tag{3.3}$$

for any $\lambda\geqslant 0$ (where $O(1)$ denotes an arbitrary quantity whose absolute value is bounded by a constant which does not depend on L or λ).

In particular, when $\lambda=0$,

$$\mathbf{P}[\mathcal{S}\cap[A_1, A_2]=\varnothing] = \mathbf{P}[\mathcal{L}<\infty] = \exp\bigl(-\tfrac{1}{3}L+O(1)\bigr) \quad \text{as } L\to\infty. \tag{3.4}$$

3.2. Generalized Cardy's formula

By conformal invariance, we may work in the half-plane $\mathbf{H}$. Map the rectangle $\mathcal{R}$ conformally onto $\mathbf{H}$ so that A_1 is mapped to 1, A_2 is mapped to ∞, A_3 is mapped to 0, and then the image $x=x(L)\in(0,1)$ of A_4 is determined for us. Let K_t be the hull of an SLE_6-process $g_t=g_{K_t}$ in $\mathbf{H}$, with driving process $W(t)$, which is started at $W(0)=x$ (that is, K_t is a translation by x of the standard SLE_6 starting at 0). In order to emphasize the dependence on x, we will use the notation $\mathbf{P}_x$ and $\mathbf{E}_x$ for probability and expectation.

Set

$$\begin{aligned} T_0 &:= \sup\{t\geqslant 0 : \overline{K}_t\cap(-\infty, 0] = \varnothing\},\\ T_1 &:= \sup\{t\geqslant 0 : \overline{K}_t\cap[1, \infty) = \varnothing\},\\ T &:= \min\{T_0, T_1\}. \end{aligned}$$

As will be demonstrated, $T_0, T_1 < \infty$ a.s. Let

$$f_t(z) := \frac{g_t(z) - g_t(0)}{g_t(1) - g_t(0)},$$

for $t<T$, which is just g_t renormalized to fix 0, 1 and ∞. It turns out that $f_T := \lim_{t \nearrow T} f_t$ exists a.s. On the event $T_1 > T_0$, f_T uniformizes the quadrilateral

$$(\mathbf{H} \setminus K_T; 1, \infty, \min(\overline{K}_T \cap \mathbf{R}), \max(\overline{K}_T \cap \mathbf{R}))$$

to the form

$$(\mathbf{H}; 1, \infty, 0, f_T(\max(\overline{K}_T \cap \mathbf{R}))).$$

Therefore, we want to know the distribution of

$$1 - f_T(\max(\overline{K}_T \cap \mathbf{R})),$$

and how it depends on x (especially when $x \in (0,1)$ is close to 1). In this subsection, we will calculate something very closely related: the distribution of $f_T'(1)$ and how it depends on x.

Set

$$\Lambda(1-x, b) := \mathbf{E}_x[1_{\{T_0 < T_1\}} f_T'(1)^b]$$

for $b \geqslant 0$ and $x \in (0,1)$. Recall the definition of the hypergeometric function ${}_2F_1$ (see, e.g., [30]):

$${}_2F_1(a_0, a_1, a_2; x) = \sum_{n=0}^{\infty} \frac{(a_0)_n (a_1)_n}{(a_2)_n n!} x^n,$$

where $(a)_n = \prod_{j=1}^{n} (a+j-1)$ and $(a)_0 = 1$. Note that ${}_2F_1(a_0, a_1, a_2; 0) = 1$.

THEOREM 3.2. *For all $b \geqslant 0$, $x \in (0,1)$,*

$$\Lambda(x, b) = \frac{\sqrt{\pi}\, 2^{-2\hat{b}} \Gamma(\frac{5}{6} + \hat{b})}{\Gamma(\frac{1}{3}) \Gamma(1 + \hat{b})} x^{1/6 + \hat{b}} {}_2F_1(\tfrac{1}{6} + \hat{b}, \tfrac{1}{2} + \hat{b}, 1 + 2\hat{b}; x),$$

where

$$\hat{b} = \tfrac{1}{6}\sqrt{1 + 24b},$$

and ${}_2F_1$ is the hypergeometric function.

Setting $b=0$, we obtain Cardy's formula [7]. Thus, this result can be thought of as a generalization of Cardy's formula.

Note that Theorem 3.2 determines completely the law of $1_{\{T_0 < T_1\}} f_T'(1)$. In particular, the Laplace transform of the conditional law of $\log 1/f_T'(1)$ given $\{T_0 < T_1\}$ is $\Lambda(1-x, b)/\Lambda(1-x, 0)$.

Proof. We first observe that $T_1<\infty$ almost surely. Indeed, $g_t(1)-W(t)$ is a Bessel process with index $\frac{5}{3}$, with time linearly scaled, and hence hits 0 almost surely in finite time (e.g. [39]). Similarly, $g_t(0)-W(t)$ hits 0 almost surely. It is clear that T_1 is the time t when $g_t(1)-W(t)$ hits 0, and T_0 is the time when $g_t(0)-W(t)$ hits 0. It follows from Theorem 2.6 and $x\in(0,1)$ that almost surely $T_0\neq T_1$. We may also conclude that

$$\lim_{x\to 0}\mathbf{P}_x[T_0<T_1]=1, \quad \lim_{x\to 1}\mathbf{P}_x[T_0<T_1]=0. \tag{3.5}$$

The next goal is to prove that

$$\lim_{t\nearrow T} f_t'(1)>0 \quad \text{if and only if} \quad T_0<T_1. \tag{3.6}$$

If $T_0<T_1$, then $\overline{K}_T\cap[1,\infty]=\varnothing$. Therefore f_T is defined and conformal near 1, and $f_T'(1)>0$, by the reflection principle. On the other hand, if $T_1<T_0$, then $\overline{K}_T\cap[1,\infty)\neq\varnothing$. We claim that

$$1\notin\overline{K}_{T_1} \quad \text{almost surely;} \tag{3.7}$$

since $\overline{K}_{T_1}\cap[1,\infty)\neq\varnothing$, this means that K_{T_1} separates 1 from ∞ in $\mathbf{H}$. Indeed, let $\phi:\mathbf{H}\to\mathbf{H}$ be the anti-conformal automorphism that fixes x and exchanges 1 and ∞. $T_1<\infty$ a.s. and $\sup_{t<T_1}|W(t)|<\infty$ a.s. imply that K_{T_1} is bounded a.s., which is the same as saying that $\phi(K_t)$ stays bounded away from 1 as $t\nearrow T_1$. But Corollary 2.3 and invariance under reflection imply that up to time T_1 the law of K_t is the same as a time change of the law of $\phi(K_t)$. Hence, a.s. K_t stays bounded away from 1 as $t\nearrow T_1$, proving (3.7). It follows from (3.7) that $\lim_{t\nearrow T_1} f_t'(1)=0$ a.s. on the event $T_1<T_0$ (observe that, given (3.7), the extremal length from a neighborhood of 0 to a neighborhood of 1 in $\mathbf{H}\setminus K_t$ tends to ∞ as $t\nearrow T_1$), and (3.6) is established.

Define the renormalized version of $W(t)$,

$$Z(t):=\frac{W(t)-g_t(0)}{g_t(1)-g_t(0)},$$

and the new time parameter

$$s=s(t):=\int_0^t \frac{dt}{(g_t(1)-g_t(0))^2}, \quad t<T.$$

Set $s_0:=\lim_{t\nearrow T} s(t)$. Since $T_1\neq T_0$ a.s., $\inf\{g_t(1)-g_t(0):t<T\}>0$ a.s., and hence $s_0<\infty$ a.s. Let $t(s)$ denote the inverse to the map $t\mapsto s(t)$. A direct calculation gives

$$\partial_s(f_{t(s)}(z))=\frac{-2}{Z(t)-f_t(z)}+\frac{2(1-f_t(z))}{Z(t)}-\frac{2f_t(z)}{1-Z(t)}$$

and

$$dZ(t) = \frac{dW_t}{g_t(1)-g_t(0)} + \frac{2\,dt}{(g_t(1)-g_t(0))^2}\left(\frac{1}{Z(t)}+\frac{1}{Z(t)-1}\right).$$

We now use the notation

$$\widetilde{Z}(s) := Z(t(s)), \quad \tilde{f}_s(z) := f_{t(s)}(z).$$

Then,

$$d\widetilde{Z}_s = dX_s + \frac{2(1-2\widetilde{Z}(s))\,ds}{\widetilde{Z}(s)(1-\widetilde{Z}(s))} = dX_s + \left(\frac{2}{\widetilde{Z}(s)} - \frac{2}{1-\widetilde{Z}(s)}\right)ds, \tag{3.8}$$

where $(X_s, s\geqslant 0)$ has the same law as $(W(t), t\geqslant 0)$; i.e., it is a Brownian motion with time rescaled by a factor of 6. Also,

$$\partial_s(\tilde{f}_s(z)) = \frac{-2}{\widetilde{Z}(s)-\tilde{f}_s(z)} + \frac{2(1-\tilde{f}_s(z))}{\widetilde{Z}(s)} - \frac{2\tilde{f}_s(z)}{1-\widetilde{Z}(s)}. \tag{3.9}$$

These two equations describe the evolution of $\tilde{f}_s(z)$. Note that $s(T)=s_0$ is the first time at which $\widetilde{Z}(s)$ hits 0 or 1.

We now assume that $b>0$. Differentiating (3.9) with respect to z gives

$$\partial_s(\log \tilde{f}_s'(z)) = \frac{-2}{(\widetilde{Z}(s)-\tilde{f}_s(z))^2} - \frac{2}{\widetilde{Z}(s)} - \frac{2}{1-\widetilde{Z}(s)} \tag{3.10}$$

(the Cauchy integral formula, for example, shows that we may indeed differentiate, but this is also legitimate since ∂_s and ∂_z commute in this case). We are particularly interested in

$$\alpha(s) := \log \tilde{f}_s'(1) = \log f_{t(s)}'(1),$$

which satisfies

$$\partial_s\alpha(s) = \frac{-2}{(\widetilde{Z}(s)-1)^2} - \frac{2}{\widetilde{Z}(s)} - \frac{2}{1-\widetilde{Z}(s)}. \tag{3.11}$$

Note that the equations (3.8) and (3.11) describe the evolution of the Markov process $(\widetilde{Z}(s), \alpha(s))$. The process stops at s_0. Define

$$y(x,v) := \mathbf{E}[\exp(b\alpha(s_0)) \mid \widetilde{Z}(0)=x, \alpha(0)=v],$$

where the expectation corresponds to the Markov process started from $\widetilde{Z}(0)=x$ and $\alpha(0)=v$. From the definition of y it follows that

$$y(x,0) = \Lambda(1-x,b),$$

since $\lim_{t\nearrow T} f_t = f_T$ in a neighborhood of 1 on the event $T_0 < T_1$ and (3.6) holds. It is standard that such a function $y(x,v)$ is C^∞, and the strong Markov property ensures that the process

$$Y = y(\widetilde{Z}(s), \alpha(s))$$

is a local martingale. The drift term in Itô's formula for dY must vanish, which gives

$$0 = \frac{2(1-2x)}{x(1-x)}\partial_x y + 3\partial_{xx}^2 y + \left(\frac{-2}{(1-x)^2} - \frac{2}{x} - \frac{2}{1-x}\right)\partial_v y. \tag{3.12}$$

As

$$\alpha(s) = \alpha(0) + \int_0^u (\partial_s \alpha(s'))\, ds',$$

we get

$$y(x,v) = \exp(bv)\, y(x,0).$$

Set

$$h(x) := y(1-x,0) = \Lambda(x,b),$$

so that $y(x,v) = \exp(bv)h(1-x)$. Hence (3.12) becomes

$$-2bh(x) + 2x(1-2x)h'(x) + 3x^2(1-x)h''(x) = 0. \tag{3.13}$$

The second statement in (3.5) implies that

$$\lim_{x\searrow 0} h(x) = 0,$$

while

$$\lim_{x\nearrow 1} h(x) = 1$$

holds, since when x is close to 0, K_{T_0} is likely to be small, by scale invariance, for example. The differential equation (3.13) can be solved explicitly by looking for solutions of the type $h(x) = x^c z(x)$: two linearly independent solutions are $(i = 1, 2)$

$$h_i(x) = x^{1/6+b_i}\, {}_2F_1\left(\tfrac{1}{6}+b_i, \tfrac{1}{2}+b_i, 1+2b_i; x\right),$$

where

$$b_1 = -b_2 = \tfrac{1}{6}\sqrt{1+24b}\,.$$

Recall that ${}_2F_1(a_0, a_1, a_2; 0) = 1$. The function $h(x)$ must be a linear combination of h_1 and h_2. However, $\lim_{x\searrow 0} h(x) = 0 = \lim_{x\searrow 0} h_1(x)$, but $\lim_{x\searrow 0} h_2(x) = \infty$. Hence, $h(x) = ch_1(x)$ for some constant c. The equality $h(1) = 1$ and knowledge of the value at $x = 1$ of

hypergeometric functions (see, e.g., [30]) allows the determination of c, and establishes the theorem in the case $b>0$. The case $b=0$ follows by taking a limit as $b\searrow 0$. □

Remark. With the same proof, Theorem 3.2 generalizes to SLE_κ with $\kappa>4$, and gives

$$\Lambda_\kappa(x,b)=C(b,\kappa)x^{1/2-2/\kappa+\hat{b}_\kappa}{}_2F_1\left(\tfrac{1}{2}-2/\kappa+\hat{b}_\kappa, 6/\kappa-\tfrac{1}{2}+\hat{b}_\kappa, 1+2\hat{b}_\kappa; x\right),$$

where

$$\hat{b}_\kappa := \frac{\sqrt{(\kappa-4)^2+16\kappa b}}{2\kappa}, \quad C(b,\kappa) := \frac{\Gamma\left(\frac{3}{2}-6/\kappa+\hat{b}_\kappa\right)\Gamma\left(\frac{1}{2}+2/\kappa+\hat{b}_\kappa\right)}{\Gamma(1-4/\kappa)\Gamma(1+2\hat{b}_\kappa)}.$$

3.3. Determination of the SLE_6-exponents

For every $t\geqslant 0$, set

$$M_t := \max(\overline{K}_t\cap\mathbf{R}).$$

The following lemma shows that our understanding of the derivative $f_T'(1)$ gives information on $f_T(M_T)$ itself.

LEMMA 3.3. *In the above setting, let* $N_T := f_T(M_T)$. *For* $b\geqslant 0$, *set*

$$\Theta(x,b) := E_{1-x}[1_{\{T_0<T_1\}}(1-N_T)^b].$$

Then

$$\left(\tfrac{1}{2}x\right)^b\Lambda\left(\tfrac{1}{2}x,b\right)\leqslant\Theta(x,b)\leqslant x^b\Lambda(x,b). \tag{3.14}$$

Note that $\Theta(x,b)$ is close to the quantity we are after, since $-\log(1-N_T)$ is approximately the extremal length of the quadrilateral

$$(\mathbf{H}\setminus K_T; \min(\overline{K}_T\cap\mathbf{R}), \max(\overline{K}_T\cap\mathbf{R}), 1, \infty).$$

Proof. It follows easily from (3.10) that $f_t'(z)$ is non-decreasing in z (viewed as a real variable), as long as $z\geqslant M_t$. Therefore,

$$1-N_T=\int_{M_T}^1 f_T'(z)\,dz\leqslant(1-M_T)f_T'(1)\leqslant xf_T'(1).$$

This gives the right-hand inequality in (3.14).

To get the other inequality, consider some fixed $x^*>1$ (x^* should be thought of as close to 1; we will eventually take $x^*=1/(1-x)$). Let $\tilde{x}^*=f_T(x^*)$. Then

$$\tilde{x}^*-N_T\geqslant\tilde{x}^*-1=\int_1^{x^*}f_T'(z)\,dz\geqslant(x^*-1)f_T'(1).$$

This gives

$$E_{1-x}[1_{\{T_0<T_1\}}(\tilde{x}^*-N_T)^b] \geqslant (x^*-1)^b \Lambda(x,b). \tag{3.15}$$

A simple scaling argument will give an upper bound of the left-hand side of this inequality in terms of Θ. Let

$$T^* = \inf\{t \geqslant 0 : \overline{K}_t \cap \mathbf{R} \setminus (0,x^*) \neq \varnothing\}.$$

Note that $T \leqslant T^* < \infty$ a.s. and that $T=T^*$ if $T_0<T_1$. For each $t \leqslant T^*$, let f_t^* be the conformal map from the unbounded component of $\mathbf{H}\setminus K_t$ to $\mathbf{H}$, which fixes the points $\infty, 0, x^*$. For all $t \leqslant T$,

$$f_t^*(z) = \frac{x^*}{f_t(x^*)} f_t(z).$$

Note that

$$f_t(x^*) \leqslant x^* = f_t^*(x^*).$$

Then,

$$1_{\{T_0<T_1\}}(\tilde{x}^*-N_T) \leqslant 1_{\{T_0<T_1\}}(x^*-f_T^*(M_T)).$$

Hence,

$$1_{\{T_0<T_1\}}(\tilde{x}^*-N_T) \leqslant 1_{\{M_{T^*}<x^*\}}(x^*-f_{T^*}^*(M_{T^*})), \tag{3.16}$$

since on the event $T_0<T_1$, we have $T=T^*$ and $M_T=M_{T^*}$. However, by scale invariance, when $W(0)=1-x$ the random variable

$$1_{\{M_{T^*}<x^*\}}(x^*-f_{T^*}^*(M_{T^*}))$$

has the same law as the random variable

$$x^* 1_{\{M_T<1\}}(1-f_T(M_T))$$

does when $W(0)=(1-x)/x^*$. Thus, combining (3.16) and (3.15) gives

$$(x^*)^b \Theta(1-(1-x)/x^*, b) \geqslant (x^*-1)^b \Lambda(x,b).$$

We take $x^*=1/(1-x)$, say, and get

$$\Theta(2x,b) \geqslant \Theta(2x-x^2,b) \geqslant x^b \Lambda(x,b),$$

which gives the left-hand side of (3.14). □

Proof of Theorem 3.1. We are now ready to derive Theorem 3.1 by combining Theorem 3.2 and Lemma 3.3.

In the setting of the theorem, let $\phi: \mathcal{R}(L) \to \mathbf{H}$ be the conformal homeomorphism satisfying $\phi(A_1)=1$, $\phi(A_2)=\infty$ and $\phi(A_3)=0$. Set $x=x(L):=\phi(A_4)$. By conformal invariance, the law of $\mathcal{L}$ is the same as that of the π-extremal distance $\mathcal{L}^*$ from $(-\infty, 0]$ to $(N_T, 1)$ in $\mathbf{H}$ (with the notations of Lemma 3.3). Considering the map $z \mapsto \log(z-1)$ makes it clear that

$$L = -\log(1-x)+O(1), \tag{3.17}$$

for $L>1$, and similarly

$$\mathcal{L}^* = -\log(1-N_T)+O(1). \tag{3.18}$$

For $L>1$ (note also that $\mathcal{L} \geqslant L$),

$$\begin{aligned}
\mathbf{E}[1_{\{\mathcal{L}<\infty\}} \exp(-\lambda \mathcal{L})] &= \mathbf{E}[1_{\{\mathcal{L}^*<\infty\}} \exp(-\lambda \mathcal{L}^*)] \\
&= \exp(O(1))\, \mathbf{E}[1_{\{N_T<1\}} (1-N_T)^{\lambda}] \quad \text{(by (3.18))} \\
&= \exp(O(1))\, \Theta(1-x, \lambda) \\
&= \exp(O(\lambda+1))(1-x)^{u(\lambda)} \quad \text{(by Lemma 3.3 and Theorem 3.2)} \\
&= \exp(O(\lambda+1)) \exp(-u(\lambda) L) \quad \text{(by (3.17))},
\end{aligned}$$

which completes the proof of Theorem 3.1. □

4. The Brownian half-plane exponents

We are now ready to combine the results collected so far and a "universality" idea similar to that developed in [29] to compute the exact value of some Brownian intersection exponents in the half-plane.

4.1. Definitions and background

In this short subsection, we quickly review some results on intersection exponents between independent planar Brownian motions. For details and complete proofs of these results, see [28], [29] .

Suppose that $n+p$ independent planar Brownian motions $\beta^1, \ldots, \beta^n$ and $\gamma^1, \ldots, \gamma^p$ are started from points $\beta^1(0)=\ldots=\beta^n(0)=0$ and $\gamma^1(0)=\ldots=\gamma^p(0)=1$ in the complex plane, and consider the probability $f_{n,p}(t)$ that for all $j \leqslant n$ and $l \leqslant p$, the paths of β^j up to time t and of γ^l up to time t do not intersect; more precisely,

$$f_{n,p}(t) := \mathbf{P}\Big[\Big(\bigcup_{j=1}^{n} \beta^j[0,t]\Big) \cap \Big(\bigcup_{l=1}^{p} \gamma^l[0,t]\Big) = \varnothing\Big].$$

It is easy to see that as $t \to \infty$ this probability decays roughly like a power of t. The (n,p)-intersection exponent $\xi(n,p)$ is defined as twice this power, i.e.,

$$f_{n,p}(t) = \left(\sqrt{t}\right)^{-\xi(n,p)+o(1)}, \quad t \to \infty.$$

We call $\xi(n,p)$ the intersection exponent between one packet of n Brownian motions and one packet of p Brownian motions (for a list of references on Brownian intersection exponents, see [28]). Note that the exponent ζ described in the introduction is $\frac{1}{2}\xi(1,1)$. It turns out to be more convenient to use this definition as a power of $\sqrt{t}$, i.e., of the space parameter. A Brownian motion travels very roughly to distance $\sqrt{t}$ in time t: recall that if β is a planar Brownian motion started from 0, say, and T_R denotes its hitting time of the circle of radius R about 0, then for all $\delta > 0$, the probability that $T_R \notin (R^{2-\delta}, R^{2+\delta})$ decays as $R \to \infty$ faster than any negative power of R. This facilitates an easy conversion between the time-based definition of intersection exponents and a definition where the particles die when they exit a large ball.

Similarly, one can define corresponding probabilities for intersection exponents in a half-plane,

$$\tilde{f}_{n,p}(t) := \mathbf{P}[\forall j \leqslant n, \forall l \leqslant p, \beta^j[0,t] \cap \gamma^l[0,t] = \varnothing \text{ and } \beta^j[0,t] \cup \gamma^l[0,t] \subset \mathcal{H}],$$

where $\mathcal{H}$ is some half-plane containing the two starting points ($\tilde{f}_{n,p}(t)$ will depend on $\mathcal{H}$). In plain words, we are looking at the probability that all Brownian motions stay in the half-plane and that all β's avoid all γ's. It is also easy to see that there exists a $\tilde{\xi}(n,p)$ (which does not depend on $\mathcal{H}$) such that

$$\tilde{f}_{n,p}(t) = \left(\sqrt{t}\right)^{-\tilde{\xi}(n,p)+o(1)}, \quad t \to \infty.$$

Note that $\tilde{\zeta}$ described in the introduction is $\frac{1}{2}\tilde{\xi}(1,1)$.

One can also define intersection exponents $\xi(n_1, \ldots, n_p)$ and $\tilde{\xi}(n_1, \ldots, n_p)$ involving more packets of Brownian motions. (For a more detailed discussion of this, see [28]). For instance, if B^1, B^2, B^3, B^4 denote four Brownian motions started from different points, the exponent $\xi(2,1,1)$ is defined by

$$\mathbf{P}[\text{the three sets } B^1[0,t] \cup B^2[0,t],\ B^3[0,t],\ B^4[0,t] \text{ are disjoint}] = t^{-\xi(2,1,1)/2+o(1)}, \quad t \to \infty.$$

One of the results of [28] is that there is a natural and rigorous way to generalize the definition of intersection exponents between packets of Brownian motions to the case where each packet of Brownian motions is the union of a "non-integer number" of paths; for

the half-plane exponents, one can define the exponents $\tilde{\xi}(u_1, ..., u_p)$, where $u_1, ..., u_p \geqslant 0$. These generalized exponents satisfy the so-called cascade relations (see [28]): for any $1 \leqslant q \leqslant p-1$,

$$\tilde{\xi}(u_0, ..., u_p) = \tilde{\xi}(u_0, ..., u_{q-1}, \tilde{\xi}(u_q, ..., u_p)). \tag{4.1}$$

Moreover, $\tilde{\xi}$ is invariant under a permutation of its arguments.

There exists (see [28], [29]) a characterization of these exponents in terms of the so-called Brownian excursions that turns out to be useful. For any bounded simply-connected open domain D, there exists a Brownian excursion measure μ_D in D. This is an infinite measure on paths $(B(t), t \leqslant \tau)$ in D such that $B(0, \tau) \subset D$ and $B(0), B(\tau) \in \partial D$ (these can be viewed as prime ends if necessary). $x_s := B(0)$ and $x_e := B(\tau)$ are the starting point and terminal point of the excursion. One possible definition of μ_D is the following: Suppose first that D is the unit disc. For any $s > 0$ define the measure P^s on Brownian paths (modulo continuous increasing time change) started uniformly on the circle of radius $\exp(-s)$, and killed when they exit D. Note for any $s_0 > s$, the killed Brownian path defined under the probability measure P^s has a probability s/s_0 to intersect the circle of radius $\exp(-s_0)$. Then, define

$$\mu_D := \lim_{s \searrow 0} (2\pi/s) P^s.$$

One can then easily check that for any Möbius transformation ϕ from D onto D, $\phi(\mu_D) = \mu_D$. This makes it possible to extend the definition of μ_D to any simply-connected domain D, by conformal invariance. These Brownian excursions also have a "restriction" property [29], as the Brownian paths only feel the boundary of D when they hit it (and get killed).

Suppose for a moment that $\mathcal{R} = \mathcal{R}(L) \subset \mathbf{C}$ is the rectangle with corners given by (3.1), and that $\mathcal{B}$ is the trace of the Brownian excursion $(B(t), t \leqslant \tau)$ in $\mathcal{R}$. Define the event

$$E_1 = \{B(0) \in [A_1, A_4] \text{ and } B(\tau) \in [A_2, A_3]\},$$

i.e., B crosses the rectangle from the left to the right. (Although $\mu_{\mathcal{R}}$ is an infinite measure, $\mu_{\mathcal{R}}(E_1)$ is finite.) When E_1 holds, let $\mathcal{R}_B^+$ be the component of $\mathcal{R} \setminus \mathcal{B}$ above $\mathcal{B}$, and let $\mathcal{R}_B^-$ be the component of $\mathcal{R} \setminus \mathcal{B}$ below $\mathcal{B}$. Let $\mathcal{L}_B^-$ (resp. $\mathcal{L}_B^+$) denote the π-extremal distance between $[A_1, x_s]$ and $[A_2, x_e]$ in $\mathcal{R}_B^-$ (resp. $[x_s, A_4]$ and $[x_e, A_3]$) in $\mathcal{R}_B^+$.

Then, for any $\alpha \geqslant 0$ and $\alpha' \geqslant 0$, the exponent $\tilde{\xi}(\alpha, 1, \alpha') = \tilde{\xi}(1, \tilde{\xi}(\alpha, \alpha'))$ is characterized by

$$\mathbf{E}_{\mu_{\mathcal{R}}}[1_{E_1} \exp(-\alpha \mathcal{L}_B^+ - \alpha' \mathcal{L}_B^-)] = \exp(-\tilde{\xi}(\alpha', 1, \alpha) L + o(L)), \tag{4.2}$$

when $L \to \infty$, where $\mathbf{E}_{\mu_{\mathcal{R}}}$ denotes expectation (that is, integration) with respect to the measure $\mu_{\mathcal{R}}$. Similarly,

$$\mathbf{E}_{\mu_{\mathcal{R}}}[1_{E_1} \exp(-\alpha \mathcal{L}_B^+)] = \exp(-\tilde{\xi}(1, \alpha) L + o(L)), \quad L \to \infty. \tag{4.3}$$

See [29]. It will also be important later that $\tilde{\xi}$ is continuous in its arguments, and that $\lambda \mapsto \tilde{\xi}(1,\lambda)$ is strictly monotone.

4.2. Statement and proof

For any $p \geqslant 0$, we put

$$v_p = \tfrac{1}{6}p(p+1).$$

Let $\mathcal{V}$ denote the set of numbers $\{v_p : p \in \mathbf{N}\}$. Note that the smallest values in $\mathcal{V}$ are $0, \frac{1}{3}, 1, 2, \frac{10}{3}, 5, 7$.

We are now ready to prove the following result:

THEOREM 4.1. *For any $k \geqslant 2$, $\alpha_1, \ldots, \alpha_{k-1}$ in $\mathcal{V}$, and for all $\alpha_k \in \mathbf{R}_+$,*

$$\tilde{\xi}(\alpha_1, \ldots, \alpha_k) = \tfrac{1}{24}\left(\left(\sqrt{24\alpha_1+1} + \ldots + \sqrt{24\alpha_k+1} - (k-1)\right)^2 - 1\right). \tag{4.4}$$

It is immediate to verify that this theorem implies Theorem 1.1.

Remark. In [26], Theorem 4.1 is extended to all non-negative reals $\alpha_1, \ldots, \alpha_k$.

Theorem 4.1 is a consequence of the cascade relations and the following lemma, which is the special case of the theorem with $k=2$, $\alpha_1 = \frac{1}{3}$:

LEMMA 4.2. *For any $\lambda > 0$,*

$$\tilde{\xi}\left(\tfrac{1}{3}, \lambda\right) = u(\lambda),$$

where $u(\lambda)$ is given by (3.2).

Proof of Theorem 4.1 (assuming Lemma 4.2). Define $U(\lambda) = \sqrt{24\lambda+1} - 1$, for all $\lambda \geqslant 0$. Lemma 4.2 implies immediately that for all $\lambda \geqslant 0$,

$$U\left(\tilde{\xi}\left(\tfrac{1}{3}, \lambda\right)\right) = U(\lambda) + 2 = U(\lambda) + U\left(\tfrac{1}{3}\right)$$

and (for all integer p) $v_{p+1} = \tilde{\xi}\left(\frac{1}{3}, v_p\right)$. The cascade relations then imply that for all integers $p_1, \ldots, p_{k-1}$,

$$\tilde{\xi}(v_{p_1}, \ldots, v_{p_{k-1}}, \lambda) = U^{-1}(2(p_1 + \ldots + p_{k-1}) + U(\lambda)).$$

This is (4.4). □

Proof of Lemma 4.2. For convenience, we again work in a rectangle rather than in the upper half-plane. Let $\mathcal{R} = \mathcal{R}(L)$, and let $\mathcal{S}$ denote the closure of the hull of SLE_6 from A_4 to $[A_1, A_2] \cup [A_2, A_3]$ in $\mathcal{R}$, as in Theorem 3.1. Let $\mathcal{B}$ denote the trace of a Brownian

excursion in $\mathcal{R}$; we will call its starting point x_s and its terminal point x_e. Consider the events

$$E_1 = \{x_s \in [A_1, A_4] \text{ and } x_e \in [A_2, A_3]\},$$
$$E_2 = \{\mathcal{S} \cap [A_1, A_2] = \varnothing\},$$
$$E_3 = E_1 \cap E_2 \cap \{\mathcal{S} \cap \mathcal{B} = \varnothing\}.$$

When E_2 holds, let $\mathcal{L}_S$ denote the π-extremal distance between the vertical edges of $\mathcal{R}$ in $\mathcal{R}\backslash\mathcal{S}$ (that is, in the quadrilateral "below" $\mathcal{S}$). Otherwise, let $\mathcal{L}_S = \infty$.

When E_1 holds, let $\mathcal{R}_B^+$ be the component of $\mathcal{R}\backslash\mathcal{B}$ above $\mathcal{B}$, and let $\mathcal{R}_B^-$ be the component of $\mathcal{R}\backslash\mathcal{B}$ below $\mathcal{B}$. Let $\mathcal{L}_B^-$ (resp. $\mathcal{L}_B^+$) denote the π-extremal distance between the vertical edges of $\mathcal{R}$ in $\mathcal{R}_B^-$ (resp. in $\mathcal{R}_B^+$), as before. When E_3 holds, let $\mathcal{L}_{SB}$ denote the π-extremal distance between the vertical edges of $\mathcal{R}$ in $\mathcal{R}_B^+\backslash\mathcal{S}$ (that is, in the quadrilateral "below $\mathcal{S}$ and above $\mathcal{B}$").

Let $\lambda > 0$. We are interested in the asymptotic behavior of

$$f(L) = \mathbf{E}[1_{E_3} \exp(-\lambda \mathcal{L}_{SB})]$$

when $L \to \infty$. By first taking expectations with respect to B (with the measure $\mu_{\mathcal{R}}$), and using the restriction property (Corollary 2.4) for the domains $\mathcal{R}$ and $\mathcal{R}_B^+$, it follows that as $L \to \infty$,

$$\begin{aligned} f(L) &= \mathbf{E}_B[\mathbf{E}_S[\exp(-\lambda \mathcal{L}_{SB})]] \\ &= \mathbf{E}_B[\exp(-u(\lambda)\mathcal{L}_B^+ + O(1))] \quad \text{(by Theorem 3.1 and restriction to } \mathcal{R}_B^+) \\ &= \exp(-\tilde{\xi}(1, u(\lambda))L + o(L)) \quad \text{(by (4.3)).} \end{aligned}$$

On the other hand, we may first take expectation with respect to $\mathcal{S}$. Given $\mathcal{S}$, the law of $\mathcal{L}_{SB}$ is the same as that of $\mathcal{L}_B^-$, by complete conformal invariance of the excursion measure (which is the analogue of the restriction property to the excursion measure, see [28]). Hence, as $L \to \infty$,

$$\begin{aligned} f(L) &= \mathbf{E}_S[\mathbf{E}_B[1_{E_3} \exp(-\lambda \mathcal{L}_{SB})]] \\ &= \mathbf{E}_S[\mathbf{E}_B[1_{E_3} \exp(-\lambda \mathcal{L}_B^-)]] \\ &= \mathbf{E}_B[\mathbf{E}_S[1_{E_3} \exp(-\lambda \mathcal{L}_B^-)]] \\ &= \mathbf{E}_B[\mathbf{P}_S[E_3 | \mathcal{L}_B^+] \exp(-\lambda \mathcal{L}_B^-)] \\ &= \mathbf{E}_B\big[\exp\big(-\tfrac{1}{3}\mathcal{L}_B^+ + O(1)\big) \exp(-\lambda \mathcal{L}_B^-)\big] \quad \text{(by (3.4))} \\ &= \exp\big(-\tilde{\xi}\big(\tfrac{1}{3}, 1, \lambda\big)L + o(L)\big) \quad \text{(by (4.2))} \\ &= \exp\big(-\tilde{\xi}\big(1, \tilde{\xi}\big(\tfrac{1}{3}, \lambda\big)\big)L + o(L)\big), \end{aligned}$$

by the cascade relations (4.1). Comparing with (4.5) gives $\tilde{\xi}(1,\tilde{\xi}(\frac{1}{3},\lambda))=\tilde{\xi}(1,u(\lambda))$. Finally,

$$\tilde{\xi}(\tfrac{1}{3},\lambda)=u(\lambda)$$

follows, since $\lambda'\mapsto\tilde{\xi}(1,\lambda')$ is strictly increasing. □

5. Crossing exponents for critical percolation

It has been conjectured [42] that SLE_6 corresponds to the scaling limit of critical percolation clusters. As additional support for this conjecture, we now show that it implies the conjectured formula for the exponents corresponding to the probability that a long rectangle is crossed by p disjoint paths or clusters of critical percolation ([13], [8], [3]).

Let us first explain the conjectured relation between SLE_6 and critical percolation. Let $D\subset\mathbf{C}$ be a domain whose boundary $\partial D\subset\mathbf{C}$ is a simple closed curve. Let $a,b\in\partial D$ be distinct points. Let γ_1 be the counterclockwise arc on ∂D from a to b, and let γ_2 be the clockwise arc on ∂D from a to b. Let $\delta>0$, and consider a fine hexagonal grid H in the plane with mesh δ; that is, each face of the grid is a regular hexagon with edges of length δ, and each vertex has degree 3. For simplicity, assume that ∂D does not pass through a vertex of H, and that a and b do not lie on edges of H. Color each hexagon of H independently, black or white, with probability $\frac{1}{2}$. Then the union of the black hexagons forms one of the standard models for critical percolation (see Grimmett [14] for percolation background and references).

Given the random coloring, there is a unique path $\beta\subset\overline{D}$ that starts at a and ends at b, such that whenever β is not on γ_1 it has a black hexagon on its "right", and whenever β is not on γ_2 it has a white hexagon on its "left". This path is the boundary between the union of the white clusters in D touching γ_2 and the black clusters in D touching γ_1. Let $f\colon D\to\mathbf{H}$ be a conformal homeomorphism such that $f(a)=0$ and $f(b)=\infty$, and parameterize β in such a way that $A(f(\beta[0,t]))=t$. Let D_t be the component of $D\setminus\beta[0,t]$ that has b on its boundary, and let $K_t=D\setminus D_t$. The conjecture from [42] (stated a bit differently) is that as $\delta\to0$ the process $(K_t,\,t\geqslant0)$ converges to SLE_6 from a to b in D. In light of this conjecture, the Locality Theorem 2.2 and its corollaries are very natural.

Now consider an arc $I\subset\partial D$, which contains b but not a. Let b_1 and b_2 be the endpoints of I, labeled in such a way that the triplet a,b_1,b_2 is in counterclockwise order around D. Let $\gamma_1'\subset\gamma_1$ be the counterclockwise arc from a to b_1, and let $\gamma_2'\subset\gamma_2$ be the clockwise arc from a to b_2. Let T be the first time such that $\beta(t)\in I$, and set $S:=\bigcup_{t<T}K_t$. Then the component α_1 of $\partial S\cap D$ joining γ_1' to I is a crossing in $\overline{\mathcal{B}}$ from γ_1' to I which is "maximal", in the sense that any other crossing $\alpha\subset\overline{\mathcal{B}}$ from γ_1' to I is separated by α_1 from b_2 in D.

Let L be large, and recall the definition of the rectangle $\mathcal{R}=\mathcal{R}(L)$ with corners given by (3.1). Let $p\in\mathbf{N}_+$ and $\sigma=(\sigma_1,\dots,\sigma_p)\in\{\text{black},\text{white}\}^p$. Consider the event $C_\sigma(\mathcal{R})$ that there are paths $\alpha_1,\dots,\alpha_p$ from $[A_4,A_1]$ to $[A_2,A_3]$ in $\mathcal{R}$ such that each α_j is contained in the union of the hexagons of color σ_j, there is no hexagon which intersects more than one of these paths, and α_{j+1} separates α_j from $[A_1,A_2]$ in $\mathcal{R}$ when $j=1,2,\dots,p-1$.

Take α_1 to be the topmost crossing with color σ_1, if such exists, let α_2 be the topmost crossing with color σ_2 which is below all the hexagons meeting α_1, etc. Then $C_\sigma(\mathcal{R})$ holds if and only if these specific $\alpha_1,\dots,\alpha_p$ exist. Note that after we condition on α_1, the hexagons "below" it are still independent and are black or white with probability $\frac{1}{2}$. Hence the following formula holds:

$$\mathbf{P}[C_\sigma(\mathcal{R})]=\mathbf{P}[\alpha_1\text{ exists}]\,\mathbf{E}[\mathbf{P}[C_{\sigma_{+1}}(\mathcal{R}_{\alpha_1})\,|\,\alpha_1]\,|\,\alpha_1\text{ exists}],$$

where $\sigma_{+1}=(\sigma_2,\sigma_3,\dots,\sigma_p)$, $\mathcal{R}_{\alpha_1}$ is the union of the hexagons below α_1, and $C_{\sigma_{+1}}(\mathcal{R}_{\alpha_1})$ is the event that there are multiple crossings with colors specified by σ_{+1} from $[A_4,A_1]$ to $[A_2,A_3]$ in $\mathcal{R}_{\alpha_1}$.

It is clear that $\mathbf{P}[C_\sigma(\mathcal{R})]$ does not depend on the choice of the sequence σ, but only its length. Moreover, the conjectured conformal invariance (or the conjecture that SLE_6 is the scaling limit) implies that $\lim_{\delta\to0}\mathbf{P}[C_\sigma(\mathcal{D})]$ depends on the quadrilateral $\mathcal{D}$ only through its conformal modulus. Hence define

$$f_p(L):=\lim_{\delta\to0}\mathbf{P}[C_\sigma(\mathcal{R}(L))],\quad \sigma\in\{\text{black},\text{white}\}^p.$$

We also set $f_p(\infty):=0$ and $f_0(L):=1_{\{L<\infty\}}$.

Let S be the SLE_6-hull from A_4 to $I:=[A_1,A_2]\cup[A_2,A_3]$ in $\mathcal{R}=\mathcal{R}(L)$, as defined in §3.1. Let $\mathcal{R}_-$ be the component of $\mathcal{R}\setminus S$ which has A_1 on its boundary, and let $\mathcal{L}$ denote the π-extremal length from $[A_4,A_1]$ to $[A_2,A_3]$ in $\mathcal{R}_-$. Note that $\mathcal{L}=\infty$ if $\bar{S}\cap[A_1,A_2]\neq\varnothing$. Then we have

$$f_p(L)=\mathbf{E}[f_{p-1}(\mathcal{L})],\quad p=1,2,\dots\,.$$

To completely justify this step requires more work, which we omit, since this whole discussion depends on a conjecture anyway. The slight difficulty has to do with the fact that having a crossing of a closed rectangle is a closed condition, and the probability of a closed event can go up when taking a weak limit of measures. One simple way to deal with this is to note that when the continuous process has a crossing in the rectangle $\mathcal{R}(L+\varepsilon)$, every sufficiently close discrete approximation of it has a crossing of the rectangle $\mathcal{R}(L)$.

Consequently, induction and Theorem 3.1 give

$$f_p(L)=\exp(-(L+O(1))v_p),\quad L\to\infty, \tag{5.1}$$

where

$$v_p = u^{\circ p}(0) = \tfrac{1}{6}p(p+1), \tag{5.2}$$

as before. Here, the constant implicit in the $O(1)$-notation may depend on p.

Note also that if

$$\sigma = (\text{white, black, white, black, ..., white}) \in \{\text{black, white}\}^{2k-1},$$

then (in the discrete setting) the event $C_\sigma(\mathcal{R})$ is identical to the event that the rectangle $\mathcal{R}$ is crossed from left to right by k disjoint white clusters.

The exponents (5.2) are those predicted in [8], [3].

References

[1] AHLFORS, L. V., *Conformal Invariants: Topics in Geometric Function Theory.* McGraw-Hill, New York, 1973.

[2] AIZENMAN, M., The geometry of critical percolation and conformal invariance, in *STATPHYS* 19 (Xiamen, 1995), pp. 104–120. World Sci. Publishing, River Edge, NJ, 1996.

[3] AIZENMAN, M., DUPLANTIER, B. & AHARONY, A., Path crossing exponents and the external perimeter in 2D percolation. *Phys. Rev. Lett.*, 83 (1999), 1359–1362.

[4] BELAVIN, A. A., POLYAKOV, A. M. & ZAMOLODCHIKOV, A. B., Infinite conformal symmetry in two-dimensional quantum field theory. *Nuclear Phys. B*, 241 (1984), 333–380.

[5] BURDZY, K. & LAWLER, G. F., Non-intersection exponents for Brownian paths. Part I: Existence and an invariance principle. *Probab. Theory Related Fields*, 84 (1990), 393–410.

[6] CARDY, J. L., Conformal invariance and surface critical behavior. *Nuclear Phys. B*, 240 (1984), 514–532.

[7] — Critical percolation in finite geometries. *J. Phys. A*, 25 (1992), L201–L206.

[8] — The number of incipient spanning clusters in two-dimensional percolation. *J. Phys. A*, 31 (1998), L105.

[9] CRANSTON, M. & MOUNTFORD, T., An extension of a result of Burdzy and Lawler. *Probab. Theory Related Fields*, 89 (1991), 487–502.

[10] DUPLANTIER, B., Loop-erased self-avoiding walks in two dimensions: Exact critical exponents and winding numbers. *Phys. A*, 191 (1992), 516–522.

[11] — Random walks and quantum gravity in two dimensions. *Phys. Rev. Lett.*, 81 (1998), 5489–5492.

[12] DUPLANTIER, B. & KWON, K.-H., Conformal invariance and intersection of random walks. *Phys. Rev. Lett.*, 61 (1988), 2514–2517.

[13] DUPLANTIER, B. & SALEUR, H., Exact determination of the percolation hull exponent in two dimensions. *Phys. Rev. Lett.*, 58 (1987), 2325–2328.

[14] GRIMMETT, G., *Percolation.* Springer-Verlag, New York, 1989.

[15] IKEDA, N. & WATANABE, S., *Stochastic Differential Equations and Diffusion Processes*, 2nd edition. North-Holland Math. Library, 24. North-Holland, Amsterdam, 1989.

[16] KENYON, R., Conformal invariance of domino tiling. *Ann. Probab.*, 28 (2000), 759–795.

[17] — The asymptotic determinant of the discrete Laplacian. *Acta Math.*, 185 (2000), 239–286.

[18] — Long-range properties of spanning trees. Probabilistic techniques in equilibrium and nonequilibrium statistical physics. *J. Math. Phys.*, 41 (2000), 1338–1363.

[19] LANGLANDS, R., POULIOT, P. & SAINT-AUBIN, Y., Conformal invariance in two-dimensional percolation. *Bull. Amer. Math. Soc. (N.S.)*, 30 (1994), 1–61.
[20] LAWLER, G. F., *Intersections of Random Walks*. Birkhäuser Boston, Boston, MA, 1991.
[21] — Hausdorff dimension of cut points for Brownian motion. *Electron. J. Probab.*, 1:2 (1996), 1–20 (electronic).
[22] — The dimension of the frontier of planar Brownian motion. *Electron. Comm. Probab.*, 1:5 (1996), 29–47 (electronic).
[23] — The frontier of a Brownian path is multifractal. Preprint, 1997.
[24] LAWLER, G. F. & PUCKETTE, E. E., The intersection exponent for simple random walk. *Combin. Probab. Comput.*, 9 (2000), 441–464.
[25] LAWLER, G. F., SCHRAMM, O. & WERNER, W., Values of Brownian intersection exponents, II: Plane exponents. *Acta Math.*, 187 (2001), 275–308. `http://arxiv.org/abs/math.PR/0003156`.
[26] — Values of Brownian intersection exponents, III: Two-sided exponents. To appear in *Ann. Inst. H. Poincaré Probab. Statist.* `http://arxiv.org/abs/math.PR/0005294`.
[27] — Analyticity of intersection exponents for planar Brownian motion. To appear in *Acta Math.*, 188 (2002). `http://arxiv.org/abs/math.PR/0005295`.
[28] LAWLER, G. F. & WERNER, W., Intersection exponents for planar Brownian motion. *Ann. Probab.*, 27 (1999), 1601–1642.
[29] — Universality for conformally invariant intersection exponents. *J. Eur. Math. Soc. (JEMS)*, 2 (2000), 291–328.
[30] LEBEDEV, N. N., *Special Functions and Their Applications*. Dover, New York, 1972.
[31] LEHTO, O. & VIRTANEN, K. I., *Quasiconformal Mappings in the Plane*, 2nd edition. Grundlehren Math. Wiss., 126. Springer-Verlag, New York–Heidelberg, 1973.
[32] LÖWNER, K., Untersuchungen über schlichte konforme Abbildungen des Einheitskreises, I. *Math. Ann.*, 89 (1923), 103–121.
[33] MADRAS, N. & SLADE, G., *The Self-Avoiding Walk*. Birkhäuser Boston, Boston, MA, 1993.
[34] MAJUMDAR, S. N., Exact fractal dimension of the loop-erased random walk in two dimensions. *Phys. Rev. Lett.*, 68 (1992), 2329–2331.
[35] MANDELBROT, B. B., *The Fractal Geometry of Nature*. Freeman, San Francisco, CA, 1982.
[36] MARSHALL, D. E. & ROHDE, S., The Loewner differential equation and slit mappings. Preprint.
[37] POMMERENKE, CH., On the Loewner differential equation. *Michigan Math. J.*, 13 (1966), 435–443.
[38] — *Boundary Behaviour of Conformal Maps*. Grundlehren Math. Wiss., 299. Springer-Verlag, Berlin, 1992.
[39] REVUZ, D. & YOR, M., *Continuous Martingales and Brownian Motion*. Grundlehren Math. Wiss., 293. Springer-Verlag, Berlin, 1991.
[40] ROHDE, S. & SCHRAMM, O., Basic properties of SLE. Preprint.
[41] RUDIN, W., *Real and Complex Analysis*, 3rd edition. McGraw-Hill, New York, 1987.
[42] SCHRAMM, O., Scaling limits of loop-erased random walks and uniform spanning trees. *Israel J. Math.*, 118 (2000), 221–288.

GREGORY F. LAWLER
Department of Mathematics
Duke University
Box 90320
Durham, NC 27708-0320
U.S.A.
jose@math.duke.edu

ODED SCHRAMM
Microsoft Research
1, Microsoft Way
Redmond, WA 98052
U.S.A.
schramm@microsoft.com

WENDELIN WERNER
Département de Mathématiques
Université Paris-Sud
Bât. 425
FR-91405 Orsay Cedex
France
wendelin.werner@math.u-psud.fr

Received January 3, 2000

Acta Math., 187 (2001), 275–308
© 2001 by Institut Mittag-Leffler. All rights reserved

Values of Brownian intersection exponents, II: Plane exponents

by

GREGORY F. LAWLER
Duke University
Durham, NC, U.S.A.

ODED SCHRAMM
Microsoft Research
Redmond, WA, U.S.A.

and

WENDELIN WERNER
Université Paris-Sud
Orsay, France

and

The Weizmann Institute of Science
Rehovot, Israel

1. Introduction

This paper is the follow-up of the paper [27], in which we derived the exact value of intersection exponents between Brownian motions in a half-plane. In the present paper, we will derive the value of intersection exponents between planar Brownian motions (or simple random walks) in the whole plane.

This problem is very closely related to the more general question of the existence and value of critical exponents for a wide class of two-dimensional systems from statistical physics, including percolation, self-avoiding walks and other random processes. Theoretical physics predicts that these systems behave in a conformally invariant way in the scaling limit, and uses this fact to predict certain critical exponents associated to these systems. We refer to [27] for a more detailed account on this link and for more references on this subject.

Let us now briefly describe some of the results that we shall derive in the present paper. Suppose that $B^1, ..., B^n$ are $n \geqslant 2$ independent planar Brownian motions started from n different points in the plane. Then it is easy to see (using a subadditivity argument) that there exists a constant ζ_n such that

$$\mathbf{P}[\forall i \neq j \in \{1, ..., n\},\ B^i[0,t] \cap B^j[0,t] = \varnothing] = t^{-\zeta_n + o(1)} \tag{1.1}$$

when $t \to \infty$. We shall prove

The first author was supported by the National Science Foundation.

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_29, © Springer Science+Business Media, LLC 2011

THEOREM 1.1. *For all* $n \geqslant 2$,

$$\zeta_n = \tfrac{1}{24}(4n^2-1).$$

This result had been conjectured by Duplantier–Kwon [14] (see also, more recently, Duplantier [11]), using ideas from theoretical physics (conformal field theory, quantum gravity and analogies with some other models for which exponents had also been conjectured).

It was shown in [5] that the exponent ζ_2 equals the corresponding exponent for simple random walks (see also [10]). This result was sharpened in [21], [26], where estimates were derived up to multiplicative constants. It follows from these results and Theorem 1.1 that if S and S' denote two independent simple random walks started from neighboring vertices in $\mathbf{Z}^2$, then for some constant $c>0$,

$$c^{-1}k^{-5/8} \leqslant \mathbf{P}[S[0,k] \cap S'[0,k] = \varnothing] \leqslant ck^{-5/8}$$

for all $k \geqslant 1$. Similarly, [20] and Theorem 1.1 imply

$$c^{-1}t^{-5/8} \leqslant \mathbf{P}[B^1[0,t] \cap B^2[0,t] = \varnothing] \leqslant ct^{-5/8}$$

for all $t \geqslant 1$, assuming that the distance between $B^1(0)$ and $B^2(0)$ is 1, say.

One can define more general exponents, allowing intersection between some Brownian motions, but forbidding intersection between different packets of Brownian motions. For instance, there exists a constant $\zeta = \zeta(n,m)$ such that if $B^1, \ldots, B^n$ and $B'^1, \ldots, B'^m$ denote $n+m$ independent planar Brownian motions started from points such that $B^i(0) \neq B'^j(0)$ for all $i \leqslant n$ and $j \leqslant m$, then

$$\mathbf{P}[\forall i \leqslant n, \forall j \leqslant m, B^i[0,t] \cap B'^j[0,t] = \varnothing] = t^{-\zeta+o(1)} \tag{1.2}$$

when $t \to \infty$. Similarly, one can define exponents $\zeta(n_1, \ldots, n_k)$ corresponding to non-intersection between k packets of Brownian motions.

It is easy to see (e.g. [23]) that there is a natural extension of the definition of $\zeta(n,m)$ to pairs (n,λ), where n is a positive integer and λ is any positive real. In [31], it is shown that there is also a natural definition of $\zeta(\lambda_1, \ldots, \lambda_k)$ where the λ_j are positive reals with $\lambda_1, \lambda_2 \geqslant 1$. In the present paper, we shall derive the value of the exponents for a certain class of k-tuples $(\lambda_1, \ldots, \lambda_k)$ (see Theorem 5.3). In particular, we shall prove

THEOREM 1.2. *For all real* $\lambda \geqslant 2$,

$$\zeta(2,\lambda) = \tfrac{1}{96}\left(\left(5+\sqrt{24\lambda+1}\right)^2-4\right). \tag{1.3}$$

It has been shown by Lawler [20], [22], [23], [25] that some of these critical exponents are closely related to the Hausdorff dimension of exceptional subsets of a planar Brownian path. Recall that a *cut point* of a connected set K is the set of points $x \in K$ such that $K \setminus \{x\}$ is disconnected. The Hausdorff dimension of the set of cut points of the Brownian path $B[0,1]$ is $2-2\zeta_2$ almost surely [20]. Consequently, we get the following corollary from Theorem 1.1.

COROLLARY 1.3. *Let B be Brownian motion in the plane. Then the Hausdorff dimension of the set of cut points of $B[0,1]$ is $\frac{3}{4}$ almost surely.*

Recall that the *frontier* of a bounded set $K \subset \mathbf{R}^2$ is its outer boundary, i.e., the boundary of the unbounded connected component of $\mathbf{R}^2 \setminus K$. The exponents $\zeta(2,\lambda)$ are closely related to the multifractal spectrum of the Brownian frontier [23]. In particular [22], the Hausdorff dimension d_F of the frontier of $B[0,1]$ is almost surely $d_F = 2-\eta_2$, where $\eta_2 := \lim_{\lambda \searrow 0} 2\zeta(2,\lambda)$ is called the *disconnection exponent* for two Brownian paths. (This is not the definition of the disconnection exponent used in [22]; however, the two definitions are equivalent; see [24], [30].) It had been conjectured by Mandelbrot [35] (by analogy with the conjectures for planar self-avoiding walks) that $d_F = \frac{4}{3}$. Upper and lower bounds for η_2 from [6], [40], combined with the fact that $d_F = 2-\eta_2$ [22], showed that $1.015 < d_F < 1.5$. (See also [4] for another proof of $d_F > 1$.) In the present paper, we derive the values of $\zeta(2,\lambda)$ only for $\lambda \geqslant 2$, so that we cannot directly apply our results to show that $\eta_2 = \frac{2}{3}$. However, in the subsequent paper [29], we prove

THEOREM 1.4. *The function $\lambda \mapsto \zeta(2,\lambda)$ is real analytic on $(0,\infty)$.*

Combining this with Theorem 1.2 shows immediately that (1.3) holds for all $\lambda > 0$, and therefore $\eta_2 = \frac{2}{3}$. This completes the proof of Mandelbrot's conjecture:

COROLLARY 1.5. *The Hausdorff dimension of the Brownian frontier is almost surely $\frac{4}{3}$.*

The formula for the multifractal spectrum of the Brownian frontier also follows. This formula has been conjectured in [31] as a consequence of the conjectures of Duplantier–Kwon [14] and of the functional relations between generalized Brownian exponents derived in [31]; see also recent physics work on this subject by Duplantier [11], [12].

The *pioneer points* of B are the image under B of the set of times t such that $B(t)$ is in the frontier of $B[0,t]$. It has been shown [25] that the dimension of the set of pioneer points of B is $2-\eta_1$, where $\eta_1 := \lim_{\lambda \searrow 0} 2\zeta(1,\lambda)$. Below, we show that

$$\zeta(1,\lambda) = \tfrac{1}{96}\left(\left(3+\sqrt{24\lambda+1}\right)^2 - 4\right)$$

for all sufficiently large λ (see (5.10)). In [29] it will be proved that $\zeta(1,\lambda)$ is analytic for $\lambda>0$. Consequently, by analytic continuation of the above formula for $\zeta(1,\lambda)$, it follows that $\eta_1=\frac{1}{4}$. Hence, using the above-mentioned result of [25], we obtain

COROLLARY 1.6. *The Hausdorff dimension of the set of pioneer points of Brownian motion is* $\frac{7}{4}$ *almost surely.*

Let us briefly mention that there is a (non-rigorous) link between our results and the conjectures concerning two-dimensional self-avoiding walks. For instance, there is a heuristic argument (see [32]) which uses the Brownian intersection exponents and explains why the number of self-avoiding walks of length N on a planar lattice increases asymptotically like $N^{11/32}\mu^N$, for some (lattice-dependent) constant $\mu>1$, as conjectured by Nienhuis [36] (see also [34] for a mathematical account).

Just as in [27], a central role in the present paper will be played by SLE_6, the stochastic Löwner evolution process with parameter 6, which is conjectured [39] to correspond to the scaling limit of two-dimensional critical percolation cluster boundaries. In [39] the processes SLE_κ were introduced, and it was shown that SLE_2 is the scaling limit of loop-erased random walk, assuming the conjecture that the latter has a conformally invariant scaling limit.

Actually, there are two versions of SLE_κ. In the first version, which we now call *radial* SLE_κ, one has a set K_t growing from a boundary point of the unit disk to the interior point 0, while in *chordal* SLE_κ, the set K_t grows from a point in $\mathbf{R}$ to ∞ within the upper half-plane. (The precise definitions will be recalled in §2.2.) By applying conformal maps, these processes can be defined in any simply-connected proper subdomain of the plane.

After recalling the definition of SLE_κ, we study in §3 some of its properties. In particular, the SLE_κ-analogues of the exponents $\zeta(1,\lambda)$, $\lambda\geqslant 1$, are computed. From Cardy's formula for SLE_6 (that we proved in [27]) we then derive the asymptotic decay of the probability that SLE_6 crosses a long rectangle without touching the upper and lower boundaries of this rectangle, and show that chordal and radial SLE_6 are very closely related.

We then turn our attention to the Brownian intersection exponents. In §5, the definition and some properties of the exponents are recalled. In particular, it is explained how to formulate these exponents in terms of non-intersection between two-dimensional Brownian excursions. Then, these facts (properties of SLE_6, exponents for SLE_6, description of the Brownian exponents in terms of Brownian excursions, properties of these Brownian excursions) are combined to derive the value of the Brownian intersection exponents.

2. Preliminaries

2.1. Notation

Throughout this paper, $\mathbf{U}$ will denote the unit disk in the complex plane $\mathbf{C}$. $C=C_1=\partial\mathbf{U}$ will denote the unit circle. For any $r>0$, $C_r=rC_1$ will denote the circle of radius r centered at 0. $\mathbf{H}$ will denote the upper half-plane $\mathbf{H}=\{x+iy:y>0\}$. When $w\neq w'$ are two points on the unit circle C, then $a(w,w')$ will denote the counterclockwise arc of C from w to w'.

Just as in [31], [32], [27], it will be convenient to use π-extremal distances of quadrilaterals: this just means π times the usual extremal distance. (Extremal distance is also known as extremal length. See, e.g., [1] for the definition and basic properties of extremal length and conformal maps.) The π-extremal distance between sets A and B in a set O will be denoted by $\ell(A,B,O)$.

Let f and g be functions, and let $l\in\mathbf{R}$ or $l=\infty$. We say that $f(x)\sim g(x)$ when $x\to l$, if $f(x)/g(x)\to 1$. We write $f(x)\approx g(x)$, if $\log f(x)/\log g(x)\to 1$, and we write $f(x)\asymp g(x)$, if $f(x)/g(x)$ is bounded above and below by positive finite constants when x is sufficiently close to l.

2.2. Radial and chordal SLE

In [27], we studied chordal SLE_κ as a random increasing family $(K_t, t\geqslant 0)$ of bounded subsets of the upper half-plane $\mathbf{H}$ (or, more generally, their image under a conformal map). As t increases, the set K_t grows, and $\bigcup_t K_t$ is unbounded. One can say that K_t is growing towards ∞, which we think of as a boundary point of $\mathbf{H}$.

Similarly, when looking at the conformal image of $(K_t, t\geqslant 0)$ under the map $\varphi(z)=(z-i)/(z+i)$ that maps $\mathbf{H}$ onto the unit disk and ∞ to 1, we get an increasing family of subsets of the unit disk that is growing towards 1.

In the present paper, we will mainly use a variant of this process, called *radial* SLE_κ, where $(K_t, t\geqslant 0)$ is an increasing family of subsets of the unit disk that grows towards 0. The main distinction is that 0 is an interior point of $\mathbf{U}$, instead of a boundary point.

Suppose that $(\delta_t, t\geqslant 0)$ is a continuous function taking values on the unit circle $C_1=\partial\mathbf{U}$. Consider for each $z\in\mathbf{U}$ the solution $g_t=g_t(z)$ of the ordinary differential equation

$$\partial_t g_t = g_t \frac{\delta_t+g_t}{\delta_t-g_t}, \quad t\geqslant 0, \tag{2.1}$$

with $g_0=z$. This equation (and the corresponding equation for g_t^{-1}) was first considered by Löwner (see [33], also [37]) and is called Löwner's differential equation. For each $z\in\mathbf{U}$, it is well defined up to the time τ_z where $\lim_{t\nearrow\tau_z} g_t=\delta_{\tau_z}$, if there is such a τ_z,

and otherwise $\tau_z=\infty$. Let D_t be the set of $z\in\mathbf{U}$ such that $t<\tau_z$ (i.e., the set on which g_t is defined), and $K_t=\mathbf{U}\setminus D_t$. It is easy to check that g_t is the unique conformal map from D_t onto $\mathbf{U}$ such that $g_t(0)=0$ and $g_t'(0)$ is a positive real number. (The notation g' refers to differentiation with respect to z.) It is also easy to verify that $g_t'(0)=\exp(t)$ by differentiating both sides of (2.1) with respect to z at $z=0$ and noting that $g_t(0)=0$.

Let $(B_t, t\geqslant 0)$ be Brownian motion on the real line, starting at some point $B_0\in\mathbf{R}$, and let $\kappa\geqslant 0$. Set $\delta_t:=\exp(i\sqrt{\kappa}\,B_t)$ and consider the solution to Löwner's differential equation as defined above. (Note that δ_t is just Brownian motion on $\partial\mathbf{U}$ with time scaled by κ.) The resulting process was defined in [39] and called SLE_κ (this acronym stands for stochastic Löwner evolution with parameter κ). The set K_t is called the hull of SLE_κ, and $(\delta_t)_{t\geqslant 0}$ its driving process.

If $f\colon D\to\mathbf{U}$ is a conformal map, then radial SLE_κ in D starting from f can be defined as the composition $g_t\circ f$, where g_t is radial SLE_κ in $\mathbf{U}$. Its hull is $f^{-1}(K_t)$, where K_t is the hull of g_t. If ∂D is sufficiently tame, then f^{-1} extends continuously to $\partial\mathbf{U}$, and hence $f^{-1}(\delta_0)$ is well defined, where δ is the driving parameter of g_t. We may then refer to the resulting SLE_κ-process in D as SLE_κ from $f^{-1}(\delta_0)$ to $f^{-1}(0)$ in D.

In [39], another variant SLE was also defined (see also [27]), which we now call *chordal* SLE_κ. Let $\mathbf{H}$ be the upper half-plane, let $\delta_t=\sqrt{\kappa}\,B_t$, and consider for each $z\in\mathbf{H}$ the solution $\tilde{g}_t=\tilde{g}_t(z)$ of the differential equation

$$\partial_t\tilde{g}_t=\frac{-2}{\delta_t-\tilde{g}_t},\quad t\geqslant 0, \tag{2.2}$$

with $\tilde{g}_0(z)=z$. As before, let $\widetilde{D}_t$ be the set of points $z\in\mathbf{H}$ for which (2.2) has a solution in some interval $[0,t']$, $t'>t$, and let $\widetilde{K}_t:=\mathbf{H}\setminus\widetilde{D}_t$. Then the process $(\tilde{g}_t)_{t\geqslant 0}$ is chordal SLE_κ in $\mathbf{H}$, and $\widetilde{K}_t$ is its hull. Recall that $\widetilde{K}_t$ is bounded for each t, but $\bigcup_{t\geqslant 0}\widetilde{K}_t$ is unbounded. If $f\colon D\to\mathbf{U}$ is a conformal map, then $f\circ g_t$ is called chordal SLE_κ in D. Its hull is $f^{-1}(K_t)$, where K_t is the hull of the process g_t. If ∂D is sufficiently tame, $f^{-1}(\infty)$ and $f^{-1}(\delta_0)$ are well defined. In this case, we refer to the process $f\circ g_t$ as SLE_κ from $f^{-1}(\delta_0)$ to $f^{-1}(\infty)$ in D. This terminology is explained more fully in [27].

For the main results of this paper, only the case $\kappa=6$ will be used. As we shall see in §4.1, radial SLE_6 is essentially the same process as chordal SLE_6. For $\kappa\neq 6$, the radial and chordal SLE_κ-processes are closely related, but the equivalence is weaker.

3. Annulus-crossing exponents for SLE

3.1. Statement

Consider a radial SLE_κ-process (with $\kappa\geqslant 0$) in the unit disk $\mathbf{U}$, with driving element δ_t starting at $\delta_0=1$. As described before, let g_t be the conformal map $g_t\colon D_t\to\mathbf{U}$, with

$g_t(0)=0$ and $g_t'(0)=\exp(t)$. Let

$$A_t := \partial\mathbf{U}\setminus\overline{K}_t.$$

It is easy to see that A_t is either an arc on $\partial\mathbf{U}$, or $A_t=\varnothing$.

Let $b\geqslant 0$, and set

$$\nu=\nu(\kappa,b) := \tfrac{1}{16}\bigl(8b+\kappa-4+\sqrt{(\kappa-4)^2+16b\kappa}\,\bigr). \tag{3.1}$$

THEOREM 3.1. *Let $\kappa>0$ and $r\in(0,1)$, and let $T(r)$ be the least $t>0$ such that K_t intersects the circle C_r. Let $\mathfrak{L}(r)$ be the π-extremal distance from C_r to C_1 in $\mathbf{U}\setminus K_{T(r)}$ (we set $\mathfrak{L}(r):=\infty$ if $A_t=\varnothing$). Then for $b\geqslant 1$, as $r\to 0$,*

$$\mathbf{E}[\exp(-b\mathfrak{L}(r))] \asymp r^{\nu}.$$

This theorem gives an analogue of Theorem 3.1 in [27] for crossing an annulus. Its proof and usage will also be similar.

Remarks. The constants implicit in the $\asymp$-notation may depend on κ and b.

One can also show that the theorem holds when $b\in(0,1)$, but we only need the case $b\geqslant 1$ (and $\kappa=6$) in the present paper. Observe that the case $\kappa=0$ is also correct, and easy to verify, for then $K_{T(r)}=[r,1]$.

One should also note that the values of the exponents for $\kappa=2$ fit with the conjecture that SLE_2 is the scaling limit of two-dimensional loop-erased random walks [39] and the exponents computed by Kenyon [16], [17], [18] for loop-erased walks. For instance, the definition of loop-erased random walks suggests that the number of vertices in the loop erasure of an N^2-step walk is roughly $N^{2-\nu(2,1)}=N^{5/4}$ (loosely speaking, in order for one of the N^2 steps of the simple random walk to remain in the loop erasure, the future of the random walk beyond that step has to avoid the past loop-erased walk), in accordance with Kenyon's results.

Similarly, combining this result (with $\kappa=6$) with the non-intersection exponents between SLE_6 in a rectangle (see the discussion at the end of [27]) and the restriction property for SLE_6 (which was proved for chordal SLE_6 in [27] and will be proved for radial SLE_6 in §4) leads to the value of non-intersection exponents between independent SLE_6 in an annulus that agree with predictions for annulus crossings in a critical percolation cluster (see [15], [9], [2]).

The proof of Theorem 3.1 will consist of three steps. First, we will obtain an estimate on $\mathbf{E}[|g_t'(e^{ix})|^b]$ for large (deterministic) times. We deduce from it a result concerning the large-time behavior of the arc length of $g_t(A_t)$, and then show that this implies Theorem 3.1.

3.2. Derivative exponents

LEMMA 3.2. *Assume that $\kappa>0$ and $b>0$. Let $\mathcal{H}(x,t)$ denote the event $\{\exp(ix)\in A_t\}$, and set*

$$f(x,t) := \mathbf{E}[|g_t'(\exp(ix))|^b 1_{\mathcal{H}(x,t)}],$$

$$q = q(\kappa,b) := \frac{\kappa-4+\sqrt{(\kappa-4)^2+16b\kappa}}{2\kappa},$$

$$h^*(x,t) := \exp(-t\nu)\big(\sin(\tfrac{1}{2}x)\big)^q,$$

where ν is as in (3.1). *Then there is a constant $c>0$ such that*

$$\forall t\geqslant 1,\ \forall x\in(0,2\pi), \quad h^*(x,t)\leqslant f(x,t)\leqslant ch^*(x,t).$$

Proof. Let $\delta_t=\exp(i\sqrt{\kappa}\,B_t)$ be the driving process of the SLE_κ, with $B_0=0$. For all $x\in(0,2\pi)$, let Y_t^x be the continuous real-valued function of t which satisfies

$$g_t(e^{ix}) = \delta_t \exp(iY_t^x)$$

and $Y_0^x=x$. The function Y_t^x is defined on the set of pairs (x,t) such that $\mathcal{H}(x,t)$ holds. Since g_t satisfies Löwner's differential equation

$$\partial_t g_t(z) = g_t(z)\frac{\delta_t+g_t(z)}{\delta_t-g_t(z)}, \tag{3.2}$$

we find that

$$dY_t^x = \sqrt{\kappa}\,dB_t + \cot\big(\tfrac{1}{2}Y_t^x\big)\,dt. \tag{3.3}$$

Let

$$\tau^x := \inf\{t\geqslant 0 : Y_t^x\in\{0,2\pi\}\}$$

denote the time at which $\exp(ix)$ is absorbed by $\overline{K}_t$, and define for all $t<\tau^x$

$$\Phi_t^x := |g_t'(\exp(ix))|.$$

On $t\geqslant\tau^x$ set $\Phi_t^x:=0$. Note that on $t<\tau^x$,

$$\Phi_t^x = \partial_x Y_t^x.$$

By differentiating (3.2) with respect to z, we find that for $t<\tau^x$,

$$\partial_t \log\Phi_t^x = -\frac{1}{2\sin^2\big(\frac{1}{2}Y_t^x\big)}, \tag{3.4}$$

and hence (since $\Phi_0^x=1$),

$$(\Phi_t^x)^b = \exp\left(-\frac{1}{2}b\int_0^t \frac{ds}{\sin^2\left(\frac{1}{2}Y_s^x\right)}\right) \tag{3.5}$$

for $t<\tau^x$.

We now show that the right-hand side of (3.5) is 0 when $t=\tau^x$. For all $x\in(0,2\pi)$, choose $m\in\mathbf{N}$ such that $x\in(2^{-m},2\pi-2^{-m})$, and define for all $n\geqslant m$ the stopping times

$$\varrho_n := \inf\{t>0 : Y_t^x \notin (2^{-n},2\pi-2^{-n})\},$$
$$\varrho_n' := \inf\{t>\varrho_n : |Y_t^x - Y_{\varrho_n}^x| \geqslant 2^{-n-1}\},$$

and the event $\mathcal{R}_n := \{\varrho_n' - \varrho_n > 4^{-n}\}$. It is easy to see (for instance by comparing Y^x with two Bessel processes and using their scaling property, or alternatively, by comparing with Brownian motions with constant drifts; see, e.g., [38] for the definition of Bessel processes) that there is some constant $c'>0$ such that for every $n\geqslant m$,

$$\mathbf{P}[\mathcal{R}_n] > c'.$$

The strong Markov property shows that the events $(\mathcal{R}_m, \mathcal{R}_{m+1}, \ldots)$ are independent, so that almost surely there exist infinitely many values of $n\in\mathbf{N}$ such that $\mathcal{R}_n$ holds. For these values of n,

$$\int_{\varrho_n}^{\varrho_{n+1}} \frac{dt}{\sin^2\left(\frac{1}{2}Y_t^x\right)} \geqslant \frac{4^{-n}}{\sin^2(2^{-n})} \geqslant 1.$$

Hence, almost surely,

$$\int_0^{\tau^x} \frac{dt}{\sin^2\left(\frac{1}{2}Y_t^x\right)} = \infty. \tag{3.6}$$

A similar argument shows that

$$\lim_{x\searrow 0} f(x,t) = \lim_{x\nearrow 2\pi} f(x,t) = 0 \tag{3.7}$$

holds for all fixed $t>0$. Suppose, for instance, that $x\leqslant 2^{-n_0}\min\left(\sqrt{t},\frac{1}{2}\pi\right)$, define $\varrho_0''=0$, $x_0=x$, and for all $n\geqslant 1$,

$$\varrho_n'' = \inf\left\{s\geqslant \varrho_{n-1}'' : s=\varrho_{n-1}''+(x_{n-1})^2 \text{ or } |Y_s^x - x_{n-1}| \geqslant \tfrac{1}{2}x_{n-1}\right\}$$

and $x_n = Y_{\varrho_n''}^x$. Clearly, for all $n\leqslant n_0$, $0<x_n<\frac{1}{2}\pi$ and $\varrho_n''<t$. By comparing Y^x with Bessel processes or Brownian motions with constant drift, it is easy to check that for some (sufficiently small) $c>0$ (independent of n and x), if we define the events $\mathcal{R}_n' :=$

$\{\varrho_n''=\varrho_{n-1}''+(x_{n-1})^2\}$, then for all $n\leqslant n_0$, $\mathbf{P}[\mathcal{R}_n'|\mathcal{F}_{n-1}]\geqslant c'$, where $\mathcal{F}_{n-1}$ denotes the σ-field generated by the events $\mathcal{R}_1', \ldots, \mathcal{R}_{n-1}'$. It therefore easily follows that when $x\searrow 0$,

$$\int_0^{\min\{\tau^x,t\}} \frac{ds}{\sin^2\left(\frac{1}{2}Y_s^x\right)} \to \infty$$

in probability, and therefore also that $f(x,t)\to 0$.

Let $F\colon [0,2\pi]\to\mathbf{R}$ be a continuous function with $F(0)=F(2\pi)=0$, which is smooth in $(0,2\pi)$, and set

$$h(x,t)=h_F(x,t):=\mathbf{E}[(\Phi_t^x)^b F(Y_t^x)].$$

By (3.5) and the general theory of diffusion Markov processes (see, e.g., [3]), we know that h is smooth in $(0,2\pi)\times\mathbf{R}_+$. From the Markov property for Y_t^x and (3.5), it follows that $h(Y_t^x,t'-t)(\Phi_t^x)^b$ is a local martingale on $t<\min\{\tau^x,t'\}$. Consequently, the drift term of the stochastic differential $d(h(Y_t^x,t'-t)(\Phi_t^x)^b)$ is zero at $t=0$. By Itô's formula, this means

$$\partial_t h=\Lambda h, \tag{3.8}$$

where

$$\Lambda h:=\tfrac{1}{2}\kappa\partial_x^2 h+\cot\left(\tfrac{1}{2}x\right)\partial_x h-\frac{b}{2\sin^2\left(\frac{1}{2}x\right)}h.$$

It can be verified directly that h^* from the statement of the lemma solves (3.8). We therefore *choose*

$$F(x):=\left(\sin\left(\tfrac{1}{2}x\right)\right)^q,$$

and claim that $h^*=h_F$. Indeed, both satisfy (3.8) on $[0,2\pi]\times[0,\infty)$, and $h^*(x,0)=F(x)=h_F(x,0)$ on $[0,2\pi]$. Moreover, $F\leqslant 1$ implies that $h_F\leqslant f$ everywhere, so $h^*-h_F=0$ on $\{0,2\pi\}\times(0,\infty)$, by (3.7). It is also immediate to verify that $h_F(x,t)\to 0$ as $(x,t)\to(0,0)$ or $(x,t)\to(2\pi,0)$. Set $M=h_F-h^*$. Then M is smooth in $(0,2\pi)\times(0,\infty)$, continuous on $[0,2\pi]\times[0,\infty)$, and satisfies $\partial_t M=\Lambda M$.

The proof that $M=0$ can be viewed as a straightforward application of the maximum principle. Fix some $\varepsilon>0$, and suppose that $M\geqslant\varepsilon$ at some point $(x,t)\in[0,2\pi]\times[0,\infty)$. Among such points, let (x_0,t_0) be a point with t_0 minimal. It is clear that there must be such a minimal point and that $x_0\in(0,2\pi)$, $t_0>0$. At (x_0,t_0) we must have $\partial_t M\geqslant 0$, by minimality of t_0. Similarly, $\partial_x M(x_0,t_0)=0$, $\partial_x^2 M(x_0,t_0)\leqslant 0$ and $M(x_0,t_0)=\varepsilon$. However, this gives $0\leqslant\partial_t M(x_0,t_0)=(\Lambda M)(x_0,t_0)\leqslant -b\varepsilon/2\sin^2\left(\frac{1}{2}x_0\right)$, by the definition of the operator Λ, a contradiction. Since ε was arbitrary, this gives $M\leqslant 0$. The same argument applied to $-M$ shows that $M\geqslant 0$, which verifies (the subscript will henceforth be omitted from h_F)

$$h(x,t)=h^*(x,t)=\exp(-t\nu)\left(\sin\left(\tfrac{1}{2}x\right)\right)^q. \tag{3.9}$$

As mentioned above, $F \leqslant 1$ implies $h(x,t) \leqslant f(x,t)$. Therefore, it remains to prove that for all $t \geqslant 1$ and $x \in (0, 2\pi)$, $f(x,t) \leqslant ch(x,t)$ for some fixed $c > 0$. By the Markov property at time $t-1$, we have for $t > 1$,

$$h(x,t) = \mathbf{E}[(\Phi_{t-1}^x)^b h(Y_{t-1}^x, 1)],$$

and similarly with f replacing h on both sides. Hence, it suffices to prove $ch(x,1) \geqslant f(x,1)$; i.e.,

$$c\mathbf{E}[(\Phi_1^x)^b F(Y_1^x)] \geqslant \mathbf{E}[(\Phi_1^x)^b].$$

Let $\sigma_y = \sigma_y^x$ be the first time $s \geqslant 0$ such that $Y_s^x = y$, and if no such s exists, set $\sigma_y = \infty$. By (3.9), $h(x,t)$ is a decreasing function of t. This and the strong Markov property give

$$\mathbf{E}[F(Y_1^x) 1_{\{\sigma_y < 1\}} (\Phi_1^x)^b] \geqslant h(y,1) \mathbf{E}[1_{\{\sigma_y < 1\}} (\Phi_{\sigma_y}^x)^b]. \tag{3.10}$$

Let

$$a := \min\{y > 0 : y - \cot(\tfrac{1}{2}y) = -2\pi\},$$

and consider the event

$$\mathcal{A} := \bigcap_{s \in [0,1]} \{Y_s^x \in (0, a)\}.$$

From (3.3) it then follows that on the event $\mathcal{A}$, $\sqrt{\kappa}\, B_1 \leqslant a - \cot(\frac{1}{2}a) = -2\pi$. Define the Brownian motion $\widetilde{B}$ on $[0,1]$ by $\widetilde{B}_s := B_s - 2sB_1$, and define $\widetilde{Y}_s^x$ by the equation (3.3), but with $\widetilde{B}$ replacing B. Note that on $\mathcal{A}$ we have $B_1 < 0$, and hence

$$\forall s \in [0,1], \quad \widetilde{Y}_s^x \geqslant Y_s^x. \tag{3.11}$$

Moreover, given $\mathcal{A}$, there is a minimal $s_0 \in [0,1]$ with $\widetilde{Y}_{s_0}^x = \pi$. Since (3.11) holds on $\mathcal{A}$, it follows from (3.5) that

$$\tilde{\Phi}_{s_0}^x \geqslant \Phi_{s_0}^x \geqslant \Phi_1^x,$$

where $\tilde{\Phi}$ is the analogue of Φ for the process $\widetilde{Y}^x$. Since $\widetilde{Y}^x$ has the same law as Y^x, with (3.10) this implies

$$h(x,1) \geqslant h(\pi,1) \mathbf{E}[1_{\{\sigma_\pi < 1\}} (\Phi_{\sigma_\pi}^x)^b] \geqslant h(\pi,1) \mathbf{E}[1_{\mathcal{A}} (\Phi_1^x)^b].$$

The same proof gives this relation with $\mathcal{A}$ replaced by the event

$$\mathcal{A}' := \bigcap_{s \in [0,1]} \{Y_s^x \in (2\pi - a, 2\pi)\}.$$

However, for the event

$$\mathcal{A}'' := \{\exists s \in [0,1] : Y_s^x \in [a, 2\pi - a]\},$$

we have

$$h(x,1) \geqslant h(a,1)\mathbf{E}[1_{\mathcal{A}''}(\Phi_1^x)^b].$$

Since $\mathcal{A}\cup\mathcal{A}'\cup\mathcal{A}''\supset\mathcal{H}(x,1)$ and $h(a,1)\leqslant h(\pi,1)$, we get

$$h(x,1) \geqslant \tfrac{1}{3}h(a,1)f(x,1).$$

This completes the proof of the lemma. □

3.3. Harmonic measure exponents

By conformal invariance, the harmonic measure from 0 of A_t in $D_t=\mathbf{U}\setminus K_t$ is $L_t/2\pi$, where L_t is the length of the arc $g_t(A_t)$.

THEOREM 3.3. *Suppose that $\kappa>0$ and $b\geqslant 1$. Then, when $t\to\infty$,*

$$\mathbf{E}[(L_t)^b] \asymp \exp(-\nu t).$$

Proof. We have to relate the behavior of $|g_t'(e^{ix})|^b$, which we have studied above, to the behavior of

$$(L_t)^b = \left(\int_0^{2\pi} |g_t'(e^{ix})|\,dx\right)^b = \left(\int_0^{2\pi} \Phi_t^x\,dx\right)^b,$$

where we set $\Phi_t^x=0$ if $\tau^x\leqslant t$. By convexity of the function $a\mapsto a^b$, it is clear that

$$\frac{1}{(2\pi)^b}\mathbf{E}[(L_t)^b] = \mathbf{E}\left[\left(\frac{1}{2\pi}\int_0^{2\pi}\Phi_t^x\,dx\right)^b\right] \leqslant \mathbf{E}\left[\frac{1}{2\pi}\int_0^{2\pi}(\Phi_t^x)^b\,dx\right]$$
$$= \frac{1}{2\pi}\int_0^{2\pi} f(x,t)\,dx \leqslant c\exp(-\nu t),$$

where we have used Lemma 3.2 for the last inequality. Consequently, we only need to prove the lower bound for $\mathbf{E}[(L_t)^b]$.

We will find constants $c_1, c_2>0$ and an event $\mathcal{U}_t^*$ such that

$$\mathbf{E}[(\Phi_t^\pi)^b 1_{\mathcal{U}_t^*}] \geqslant c_1 e^{-\nu t},$$

and on the event $\mathcal{U}_t^*$,

$$|\log\Phi_t^\pi - \log\Phi_t^y| \leqslant c_2, \quad \forall y\in[\pi,\pi+c_1].$$

Then

$$\mathbf{E}[(L_t)^b] \geqslant \mathbf{E}\left[1_{\mathcal{U}_t^*}\left(\int_\pi^{\pi+c_1}\Phi_t^x\,dx\right)^b\right] \geqslant (c_1e^{-c_2})^b\mathbf{E}[1_{\mathcal{U}_t^*}(\Phi_t^\pi)^b] \geqslant c_3e^{-\nu t},$$

which will prove the theorem.

Assume (with no loss of generality) that $t>3$, and let t' be the integer in $(t-2, t-1]$. Define the event

$$\mathcal{V}_t = \{\tfrac{1}{2}\pi \leqslant Y_s^x \leqslant \tfrac{3}{2}\pi,\ \forall s \in [t', t]\}.$$

It follows from Lemma 3.2 that there is some constant $c_4>0$ such that

$$\mathbf{E}\big[(\Phi_{t'-1}^x)^b 1_{\{Y_{t'-1}^x \in [c_4, 2\pi - c_4]\}}\big] \geqslant c_4 e^{-\nu t}.$$

This clearly implies

$$\mathbf{E}[(\Phi_t^\pi)^b 1_{\mathcal{V}_t}] \geqslant c_5 e^{-\nu t}$$

for some $c_5>0$.

Let $\mathcal{U}_t = \mathcal{U}_t(\alpha)$ be the event

$$\mathcal{U}_t := \{\alpha e^{-s/8} \leqslant Y_s^\pi \leqslant 2\pi - \alpha e^{-s/8},\ \forall s \in [0, t]\}.$$

We claim that for some $\alpha>0$, and every $t>3$,

$$\mathbf{E}[(\Phi_t^\pi)^b 1_{\mathcal{U}_t} 1_{\mathcal{V}_t}] \geqslant \tfrac{1}{2} c_5 e^{-\nu t}. \tag{3.12}$$

To prove this it suffices to show that for some $\alpha>0$,

$$\mathbf{E}[(\Phi_t^x)^b 1_{\neg \mathcal{U}_t} 1_{\mathcal{V}_t}] \leqslant \tfrac{1}{2} c_5 e^{-\nu t}. \tag{3.13}$$

For $u=0, \ldots, t'-1$, and $\alpha \in (0, \frac{1}{5})$, let $\mathcal{W}_u = \mathcal{W}_u(\alpha)$ be the event

$$\mathcal{W}_u := \{\alpha e^{-u/8} \leqslant Y_s^\pi \leqslant 2\pi - \alpha e^{-u/8},\ \forall s \in [u, u+1]\}.$$

Note that

$$\bigcup_{u=0}^{t'-1} \neg \mathcal{W}_u \supset \mathcal{V}_t \cap \neg \mathcal{U}_t.$$

Hence,

$$\mathbf{E}[(\Phi_t^x)^b 1_{\neg \mathcal{U}_t} 1_{\mathcal{V}_t}] \leqslant \sum_{u=0}^{t'-1} \mathbf{E}[(\Phi_t^x)^b 1_{\neg \mathcal{W}_u}]. \tag{3.14}$$

Note that for $u=0, \ldots, t'-1$, the strong Markov property shows that

$$\mathbf{E}[(\Phi_t^x)^b 1_{\neg \mathcal{W}_u} | \mathcal{F}_u] \leqslant (\Phi_u^x)^b f(\alpha e^{-u/8}, t-u-1)$$

(here $\mathcal{F}_u$ denotes the σ-field generated by $(Y_s, s \leqslant u)$). Hence, by Lemma 3.2,

$$\mathbf{E}[(\Phi_t^x)^b 1_{\neg \mathcal{W}_u}] \leqslant c_6 \alpha^q e^{-uq/8} e^{-\nu t}.$$

Now (3.14) gives

$$\mathbf{E}[(\Phi_t^x)^b 1_{\neg\mathcal{U}_t} 1_{\mathcal{V}_t}] \leqslant c_7 \alpha^q e^{-\nu t},$$

and hence by choosing α sufficiently small, we get (3.13) and therefore (3.12). Fix such an $\alpha \in (0, \frac{1}{5})$, and let $\mathcal{U}_t^* := \mathcal{U}_t \cap \mathcal{V}_t$.

Observe that (3.3) implies that if $0 < x < y < 2\pi$,

$$\partial_t (Y_t^y - Y_t^x) = \cot(\tfrac{1}{2} Y_t^y) - \cot(\tfrac{1}{2} Y_t^x) \leqslant -\tfrac{1}{2}(Y_t^y - Y_t^x)$$

(since $\cot'(u) \leqslant -1$ in the range $u \in (0, \pi)$) as long as $t < \min\{\tau^x, \tau^y\}$, so that

$$|Y_t^y - Y_t^x| \leqslant |x-y| e^{-t/2}. \tag{3.15}$$

Let $y \in (\pi, \pi + \frac{1}{2}\alpha)$. Then

$$0 < Y_s^y - Y_s^\pi \leqslant e^{-s/2}(y-\pi) \leqslant \tfrac{1}{2}\alpha e^{-s/2}, \quad \forall s \leqslant \min\{\tau^\pi, \tau^y\}.$$

On the event $\mathcal{U}_t^*$, we must therefore have $t < \min\{\tau^x, \tau^y\}$ and

$$Y_s^\pi, Y_s^y \in \left[\tfrac{1}{2}\alpha e^{-s/8}, 2\pi - \tfrac{1}{2}\alpha e^{-s/8}\right]$$

for all $s \in [0, t]$. By (3.4), this shows that on the event $\mathcal{U}_t^*$, for all $s \leqslant t$,

$$\begin{aligned} |\partial_s(\log \Phi_s^y - \log \Phi_s^\pi)| &\leqslant |Y_s^y - Y_s^\pi| \max\left\{\tfrac{1}{2}\left|\partial_x\left(\sin^{-2}(\tfrac{1}{2}x)\right)\right| : x \in \left[\tfrac{1}{2}\alpha e^{-s/8}, 2\pi - \tfrac{1}{2}\alpha e^{-s/8}\right]\right\} \\ &\leqslant c_7 |Y_s^y - Y_s^\pi| e^{3s/8}. \end{aligned}$$

Now (3.15) gives

$$|\partial_s(\log \Phi_s^y - \log \Phi_s^\pi)| \leqslant c_8 e^{-s/8}.$$

Therefore, on the event $\mathcal{U}_t^*$,

$$|\log(\Phi_t^y / \Phi_t^\pi)| \leqslant 8 c_8.$$

This completes the proof of the theorem. □

3.4. Extremal distance exponents

Proof of Theorem 3.1. Let $\varrho(t) := \inf\{|z| : z \in K_t\}$. Recall that

$$T(r) = \inf\{t : \varrho(t) = r\},$$

$A_t = \partial\mathbf{U} \setminus \overline{K}_t$ and that L_t is the length of the arc $g_t(A_t)$. Recall that the Schwarz lemma says that if $G: \mathbf{U} \to \mathbf{U}$ is analytic, then $|G'(0)| \leqslant 1$, and the Koebe $\frac{1}{4}$-theorem says that

if $G:\mathbf{U}\to\mathbf{C}$ is conformal with $G(0)=0$, then $G(\mathbf{U})\supset\frac{1}{4}|G'(0)|\mathbf{U}$. (See, e.g., [1].) Since $g_t'(0)=\exp(t)$, applying the former to $z\mapsto g_t(\varrho(t)z)$ and the latter to g_t^{-1} give

$$\tfrac{1}{4}e^{-t}\leqslant\varrho(t)\leqslant e^{-t}.$$

In particular, if we fix the radius $r<\frac{1}{8}$ and define the deterministic times

$$t'=t'(r):=\log\frac{1}{r},\quad t=t(r):=\log\frac{1}{8r},$$

then

$$t<T(r)\leqslant t'\leqslant T\big(\tfrac{1}{4}r\big)$$

and

$$\varrho(t)\geqslant 2r\geqslant r\geqslant\varrho(t')\geqslant\tfrac{1}{4}r\geqslant\tfrac{1}{8}r.$$

In the following lines, $\ell(S,S';U)$ will stand for the π-extremal distance between the sets S and S' in U. Recall that $\mathfrak{L}(r)=\ell(C_r,C_1;\mathbf{U}\setminus K_{T(r)})$ and define $l_r=\ell(C_r,C_1;\mathbf{U}\setminus K_{t(r)})$. It follows from the above that

$$l_r\leqslant\mathfrak{L}(r)\leqslant l_{r/8}. \tag{3.16}$$

Hence, it will be sufficient to study the asymptotic behavior of $\mathbf{E}[\exp(-bl_r)]$.

Note that $g_t:D_t\to\mathbf{U}$ is a conformal map defined on $D_t\supset 2r\mathbf{U}$, that $g_t(0)=0$ and that $g_t'(0)=1/8r$. Hence, it follows immediately from the Schwarz lemma and the Koebe $\frac{1}{4}$-theorem that for any $r<\frac{1}{8}$,

$$2^{-5}\mathbf{U}\subset g_t(r\mathbf{U})\subset\tfrac{1}{2}\mathbf{U}.$$

Hence,

$$\ell(C_{2^{-5}},g_t(A_t);\mathbf{U})\geqslant l_r\geqslant\ell(C_{1/2},g_t(A_t);\mathbf{U}),$$

and this implies easily that

$$\exp(-l_r)\asymp L_{t(r)}.$$

Theorem 3.1 now follows from Theorem 3.3 and (3.16). □

4. Properties of SLE_6

We now turn our attention towards specific properties of SLE_6.

4.1. Equivalence of chordal and radial SLE_6

The following result shows that chordal SLE_6 and radial SLE_6 are nearly the same process. (When $\kappa\neq 6$, a weaker form of equivalence holds.) A consequence of the equivalence

of radial and chordal SLE_6 is that radial SLE_6 satisfies a restriction property, since in [27] a restriction property for chordal SLE_6 has been established. The significance of the restriction property to the Brownian intersection exponents has been evident since [32].

THEOREM 4.1. *Let $\theta\in\partial\mathbf{U}\setminus\{1\}$, and let K_t be the hull of a radial SLE_6-process in the unit disk $\mathbf{U}$ with driving process δ_t satisfying $\delta_0=\theta$. Set*

$$T:=\sup\{t\geqslant 0:1\notin\overline{K}_t\}.$$

Let $\widetilde{K}_u$ be the hull of a chordal SLE_6-process in $\mathbf{U}$ starting also at θ and growing towards 1, and let

$$\widetilde{T}:=\sup\{u\geqslant 0:0\notin\widetilde{K}_u\}.$$

Then, up to a random time change, the process $t\mapsto K_t$ restricted to $[0,T)$ has the same law as the process $u\mapsto\widetilde{K}_u$ restricted to $[0,\widetilde{T})$.

Note that T (resp. $\widetilde{T}$) is the first time where K_t (resp. $\widetilde{K}_u$) disconnects 0 from 1.

Proof. In order to point out where the assumption $\kappa=6$ is important, we let $(K_t, t\geqslant 0)$ and $(\widetilde{K}_u, u\geqslant 0)$ be SLE_κ-processes, without fixing the value of κ for the moment.

Let us first briefly recall (see, e.g., [27]) how $\widetilde{K}_u$ is defined. Let ψ be the Möbius transformation that satisfies $\psi(\mathbf{U})=\mathbf{H}$, $\psi(1)=\infty$, $\psi(-1)=0$ and $\psi(0)=i$; i.e.,

$$\psi(z)=i\frac{1+z}{1-z}.$$

Suppose that $u\mapsto\widetilde{B}_u$ is a real-valued Brownian motion such that $\sqrt{\kappa}\,\widetilde{B}_0=\psi(e^{i\theta})$. For all $z\in\mathbf{U}$, define the function $\tilde{g}_u=\tilde{g}_u(z)$ such that $\tilde{g}_0(z)=\psi(z)$ and

$$\partial_u\tilde{g}_u=\frac{2}{\tilde{g}_u-\sqrt{\kappa}\,\widetilde{B}_u}.$$

This function is defined up to the (possibly infinite) time σ_z where $\tilde{g}_u(z)$ hits $\sqrt{\kappa}\,\widetilde{B}_u$. Then, $\widetilde{K}_u$ is defined by $\widetilde{K}_u=\{z\in\mathbf{U}:\sigma_z\leqslant u\}$, so that $\tilde{g}_u$ is a conformal map from $\mathbf{U}\setminus\widetilde{K}_u$ onto the upper half-plane. This defines the process $(\widetilde{K}_u, u\geqslant 0)$ (the scaling property of Brownian motion shows that the choice of the conformal map ψ only influences the law of $(\widetilde{K}_u)_{u\geqslant 0}$ via a time change).

We are now going to study the radial SLE_κ. Let $g_t:\mathbf{U}\setminus K_t\to\mathbf{U}$ be the conformal map normalized by $g_t(0)=0$ and $g_t'(0)>0$. Recall that

$$\partial_t g=g\frac{\delta+g}{\delta-g}, \tag{4.1}$$

where $\delta_t = \exp(i\sqrt{\kappa}\, B_t)$, and B is Brownian motion on $\mathbf{R}$ with $\exp(iB_0)=\theta$. Let ψ be the Möbius transformation as before, and define

$$\begin{aligned} e_t &:= g_t(1), \\ f_t(z) &:= \psi(g_t(z)/e_t), \\ \gamma_t &:= \psi(\delta_t/e_t). \end{aligned}$$

These are well defined, as long as $t<T$. Note that f_t is a conformal map from $\mathbf{U}\setminus K_t$ onto the upper half-plane, $f_t(1)=\infty$ and $\gamma_t \in \mathbf{R}$. From (4.1) it follows that

$$\partial_t f = -\frac{(1+\gamma^2)(1+f^2)}{2(\gamma - f)}.$$

Let

$$\phi_t(z) = a(t)z + b(t)$$

where

$$a(0)=1, \quad \partial_t a = -\tfrac{1}{2}(1+\gamma^2)a$$

and

$$b(0)=0, \quad \partial_t b = -\tfrac{1}{2}(1+\gamma^2)a\gamma.$$

Set

$$\begin{aligned} h_t &:= \phi_t \circ f_t, \\ \beta_t &:= \phi_t(\gamma(t)). \end{aligned}$$

Then (and this is the reason for the choice of the functions a and b)

$$\partial_t h = -\tfrac{1}{2}a\frac{(1+\gamma^2)^2}{\gamma - f} = -\frac{\tfrac{1}{2}(1+\gamma^2)^2 a^2}{\beta - h}.$$

h_t is also a conformal map from $\mathbf{U}\setminus K_t$ onto the upper half-plane with $h_t(1)=\infty$. Note also that $h_0(z)=\psi(z)$. We introduce a new time parameter $u=u(t)$ by setting

$$\partial_t u = \tfrac{1}{4}(1+\gamma^2)^2 a^2, \quad u(0)=0.$$

Then

$$\frac{\partial h}{\partial u} = \frac{-2}{\beta - h}.$$

Since this is the equation defining the chordal SLE_6-process, it remains to show that $u \mapsto \beta_{t(u)}/\sqrt{\kappa}$ is Brownian motion (stopped at some random time). This is a direct but tedious application of Itô's formula:

$$d\gamma_t = \tfrac{1}{2}(1+\gamma^2)\sqrt{\kappa}\, dB_t + \tfrac{1}{2}\gamma(1+\gamma^2)\left(\tfrac{1}{2}\kappa - 1\right) dt$$

and

$$d\beta_t = \tfrac{1}{2}(1+\gamma^2)a\big(\sqrt{\kappa}\,dB_t + \big(-3+\tfrac{1}{2}\kappa\big)\gamma\,dt\big).$$

This proves the claim, and establishes the theorem. □

Note that when $\kappa \neq 6$, even though $u \mapsto \beta$ is not a local martingale, its law is absolutely continuous with respect to that of $\sqrt{\kappa}$ times a Brownian motion, as long as γ and u remain bounded. More precisely:

PROPOSITION 4.2. *Let $(K_t, t \geqslant 0)$, $(\widetilde{K}_u, u \geqslant 0)$, T and $\widetilde{T}$ be defined just as in Theorem 4.1, except that they are SLE_κ with general $\kappa > 0$. There exist two non-decreasing families of stopping times, $(T_n, n \geqslant 1)$ and $(\widetilde{T}_n, n \geqslant 1)$, such that almost surely $T_n \to T$ and $\widetilde{T}_n \to \widetilde{T}$ when $n \to \infty$, and such that for each $n \geqslant 1$, the laws of $(K_t, t \in [0, T_n])$ and $(\widetilde{K}_u, u \in [0, \widetilde{T}_n])$ are equivalent (in the sense that they have a positive density with respect to each other) modulo increasing time change.*

Proof. It suffices to take

$$T_n = \min\{n, \inf\{t > 0 : |\delta_t - e_t| < 1/n\}\}.$$

Then, it is easy to see that before T_n, $|\gamma|$ remains bounded, a is bounded away from 0 (note also that $a \leqslant 1$ always), so that t/u is bounded and bounded away from 0. Hence, $u(T_n)$ is also bounded (since $T_n \leqslant n$).

It now follows directly from Girsanov's theorem (see, e.g., [38]) that the law of $\big(\beta(u)/\sqrt{\kappa}\big)_{u \leqslant u(T_n)}$ is equivalent to that of Brownian motion up to some (bounded) stopping time, and Proposition 4.2 follows. □

4.2. The crossing exponent for SLE_6

In this section, we are going to study the probability that chordal SLE_6 started at some point on the left-hand side of a rectangle crosses the rectangle from the left to the right without touching the upper and lower boundaries of the rectangle. As we shall see, the estimate obtained is a direct consequence of Cardy's formula for SLE_6 proved in [27].

The notation turns out to be simpler when considering crossings of a quadrilateral in the unit disk, which is equivalent to the rectangle case by conformal invariance. We now describe the setup more precisely. Recall that when $w, w' \in \partial\mathbf{U}$, the counterclockwise

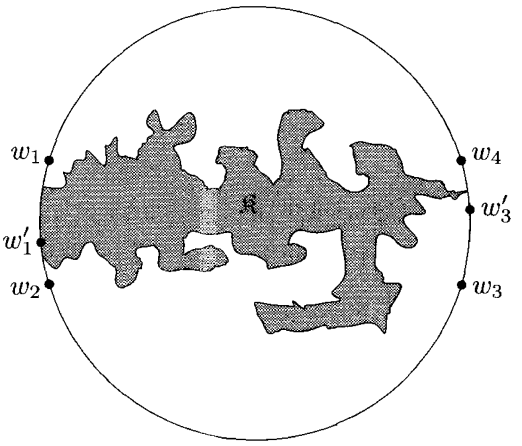


Fig. 4.1. A successful crossing.

arc from w to w' is denoted $a(w,w')$. Let $\theta\in(0,\frac{1}{2}\pi)$, $\alpha\in(-1,1)$, and define the points

$$w_1 := \exp(i(\pi-\theta)),$$
$$w_1' := \exp(i(\pi+\alpha\theta)),$$
$$w_2 := \exp(i(\pi+\theta)),$$
$$w_3 := \exp(-i\theta),$$
$$w_4 := \exp(i\theta),$$

and note that they appear in counterclockwise order on $\partial\mathbf{U}$. See Figure 4.1. Let w_3' be any point in $a(w_3,w_4)$. Consider a chordal SLE_6 in $\mathbf{U}$ started from w_1' and growing towards w_3'. Let K_t be the hull, and let T be the first time t such that $\overline{K}_t$ intersects the arc $a(w_3,w_4)$. Set

$$\mathfrak{K} := \bigcup_{t<T} \overline{K}_t.$$

As shown in [27], the restriction property for SLE_6 shows that up to a monotone time change, the law of the process $(K_t)_{t<T}$ does not depend on the choice of w_3'. Since we use the restriction property, the result derived in this subsection is specific to $\kappa=6$.

We are interested in the event

$$\mathcal{E} := \{\mathfrak{K}\cap(a(w_2,w_3)\cup a(w_4,w_1))=\varnothing\}.$$

LEMMA 4.3. *Suppose that $\alpha_0\in(0,1)$ is fixed. When $\theta\searrow 0$,*

$$\mathbf{P}[\mathcal{E}] \asymp \theta^2,$$

and this estimate is uniform for $\alpha\in(-\alpha_0,\alpha_0)$. Moreover, when $\theta\searrow 0$,

$$\max_{\alpha\in(-1,1)} \mathbf{P}[\mathcal{E}] \asymp \theta^2.$$

Recall also that $\theta^2 \sim \exp(-\ell)$ as $\theta \searrow 0$, where $\ell = \ell(a(w_1,w_2), a(w_3,w_4); \mathbf{U})$ denotes the π-extremal distance between $a(w_1,w_2)$ and $a(w_3,w_4)$ in $\mathbf{U}$.

The lemma is hardly surprising. Suppose for a moment that we take $w_1' = -1$, and let T' be the first time such that $\bar{K}_t$ intersects $a(-i,i)$. Let z be the point in $a(-i,i)$ which is on the boundary of $\bigcup_{t<T'} K_t$. It is then easy to believe that the probability density for the location of z should be bounded away from 0 and infinity in every compact subset of $a(-i,i)$. By conformal invariance, this implies the first part of the lemma for the case $\alpha = 0$. Although it should not be too hard to prove the lemma with some general arguments such as these, we find it easier to rely on the more refined results from [27].

Proof. Define the events

$$\mathcal{E}' = \{\bar{K}_t \text{ hits } a(w_2,w_4) \text{ before } a(w_4,w_1)\},$$
$$\mathcal{E}'' = \{\bar{K}_t \text{ hits } a(w_2,w_3) \text{ before } a(w_3,w_1)\}.$$

Note that $\mathcal{E}'' \subset \mathcal{E}'$, and that $\mathcal{E} = \mathcal{E}' \setminus \mathcal{E}''$. Cardy's formula for SLE_6 derived in [27] (Theorem 3.2 in the case where $b=0$) gives the exact value of $\mathbf{P}(\mathcal{E}')$ and $\mathbf{P}(\mathcal{E}'')$, as follows. Define the cross ratios

$$c' := \frac{(w_1 - w_1')(w_4 - w_2)}{(w_4 - w_1')(w_1 - w_2)}, \quad c'' := \frac{(w_1 - w_1')(w_3 - w_2)}{(w_3 - w_1')(w_1 - w_2)},$$

and set

$$G(x) := {}_2F_1\left(\tfrac{1}{3}, \tfrac{2}{3}, \tfrac{4}{3}; x\right) \frac{\sqrt{\pi}}{2^{1/3}\Gamma\left(\frac{1}{3}\right)\Gamma\left(\frac{7}{6}\right)} x^{1/3}$$

(where ${}_2F_1$ denotes the hypergeometric function). Then

$$\mathbf{P}[\mathcal{E}'] = G(c'), \quad \mathbf{P}[\mathcal{E}''] = G(c''), \quad \mathbf{P}[\mathcal{E}] = G(c') - G(c'').$$

(To compute $\mathbf{P}[\mathcal{E}']$, we view K_t as an SLE_6 from w_1' to w_4, while to compute $\mathbf{P}[\mathcal{E}'']$, we view K_t as an SLE_6 from w_1' to w_3. As remarked above, the choice of $w_3' \in a(w_3,w_4)$ does not matter.) Note that

$$c' = \cot\theta \tan\left(\tfrac{1}{2}(1+\alpha)\theta\right), \quad c'' = \frac{\sin\left(\frac{1}{2}(1+\alpha)\theta\right)}{\sin\theta \cos\left(\frac{1}{2}(1-\alpha)\theta\right)}.$$

Both c' and c'' converge to $\frac{1}{2}(1+\alpha)$ when $\theta \searrow 0$ and $c' - c'' \sim \frac{1}{4}(1-\alpha^2)\theta^2$. Since $G'(x) = \left(\frac{1}{3}\sqrt{\pi}\right) 2^{-1/3} \Gamma\left(\frac{1}{3}\right)^{-1} \Gamma\left(\frac{7}{6}\right)^{-1} ((1-x)x)^{-2/3}$, it follows that

$$\mathbf{P}[\mathcal{E}] = G(c') - G(c'') \sim \left(\tfrac{1}{6}\sqrt{\pi}\right) \Gamma\left(\tfrac{1}{3}\right)^{-1} \Gamma\left(\tfrac{7}{6}\right)^{-1} (1-\alpha^2)^{1/3} \theta^2,$$

as $\theta \searrow 0$, and the lemma follows. □

5. Brownian intersection exponents

5.1. Definitions and statement of results

This section begins with a review of the definitions and some general facts concerning intersection exponents between planar Brownian paths, and proceeds with a statement of some theorems. For more details concerning the background results on intersection exponents, as well as some references to the literature, see [31], [32].

Suppose that $n+m$ independent Brownian motions $B^1, ..., B^n$ and $B'^1, ..., B'^m$ are started from points $B^1(0)=...=B^n(0)=i$ and $B'^1(0)=...=B'^m(0)=1$ in the complex plane, and let $\mathfrak{B}_r^j$, $\mathfrak{B}_r'^k$ denote the traces (i.e., images) of the paths up to the first time they reach the circle C_r. Consider the probability $f_{n,m}(r)$ that the $\mathfrak{B}$-traces do not intersect the $\mathfrak{B}'$-traces, i.e.,

$$f_{n,m}(r) := \mathbf{P}\Big[\Big(\bigcup_{j=1}^{n} \mathfrak{B}_r^j\Big) \cap \Big(\bigcup_{l=1}^{m} \mathfrak{B}_r'^l\Big) = \varnothing\Big].$$

It is easy to see that as $r \to \infty$ this probability decays like a power law; the (n,m)-intersection exponent $\xi(n,m)$ is defined by

$$f_{n,m}(r) = r^{-\xi(n,m)+o(1)}, \quad r \to \infty.$$

We call $\xi(n,m)$ the intersection exponent between one packet of n Brownian motions and one packet of m Brownian motions. It is easy to see (e.g. [31]) that the exponent $\zeta(n,m)$ described in the introduction is equal to $\frac{1}{2}\xi(n,m)$.

By using the conformal map $z \mapsto 1/z$ and invariance of planar Brownian motion under conformal mapping, it is clear that

$$f_{n,m}(r) = f_{n,m}(1/r),$$

so that the exponents also measure the decay of $f_{n,m}$ when $r \searrow 0$.

Similarly, one can define corresponding probabilities for intersection exponents in a half-plane,

$$\tilde{f}_{n,m}(r) := \mathbf{P}\Big[\Big(\bigcup_{j=1}^{n} \mathfrak{B}_r^j\Big) \cap \Big(\bigcup_{l=1}^{m} \mathfrak{B}_r'^l\Big) = \varnothing \text{ and } \Big(\bigcup_{j=1}^{n} \mathfrak{B}_r^j\Big) \cup \Big(\bigcup_{l=1}^{m} \mathfrak{B}_r'^l\Big) \subset H\Big],$$

where H is some half-plane through the origin containing 1 and i. ($\tilde{f}_{n,m}(r)$ does depend on H.) In plain words, we are looking at the probability that all Brownian motions stay in the half-plane H and that all the $\mathfrak{B}$-traces avoid all the $\mathfrak{B}'$-traces. It is also easy to see that there exists a $\tilde{\xi}(n,m)$ (which does not depend on H) such that

$$\tilde{f}_{n,m}(r) = r^{-\tilde{\xi}(n,m)+o(1)}$$

when $r \to \infty$.

These intersection exponents can be generalized in a number of ways. For example, set

$$Z_r := \mathbf{P}\Big[\mathfrak{B}_r'^1 \cap \bigcup_{j=1}^{n} \mathfrak{B}_r^j = \varnothing \,\Big|\, \bigcup_{j=1}^{n} \mathfrak{B}_r^j\Big].$$

Then $f_{n,m}(r)=\mathbf{E}[Z_r^m]$, and define $\xi(n,\lambda)$ for all $\lambda>0$ by the relation

$$\mathbf{E}[Z_r^\lambda] \approx r^{-\xi(n,\lambda)}, \quad r\to\infty.$$

It has been proved by Lawler [24] that in fact

$$\mathbf{E}[Z_r^\lambda] \asymp r^{-\xi(2,\lambda)}.$$

The same proof, with minor notational modifications, applies to the other exponents as well. See also [30] for a new self-contained proof of this result.

One can also define the exponents $\xi(n_1,n_2,...,n_k)$ and $\tilde{\xi}(n_1,n_2,...,n_k)$ describing the probability of non-intersection of k packets of Brownian motions. In fact, [31] proves that there is a natural and rigorous way to generalize these definitions to the case where the numbers n_j are positive reals (i.e., not required to be integers). For the definition of $\xi(\lambda_1,...,\lambda_k)$ one only needs to assume that at least two of the numbers $\lambda_1,...,\lambda_k$ are at least 1, and for $\tilde{\xi}$ even this assumption is not necessary. What makes these extended definitions natural is that they are uniquely determined by certain identities and relations. For one, the functions ξ and $\tilde{\xi}$ are invariant under permutations of their arguments. Moreover, they satisfy the so-called cascade relations [31]: for any $1\leqslant q\leqslant m-1$ and $(\lambda_1,...,\lambda_m)$ such that $\lambda_1\geqslant 1$ and $\max\{\lambda_2,...,\lambda_m\}\geqslant 1$,

$$\xi(\lambda_1,...,\lambda_m)=\xi(\lambda_1,...,\lambda_q,\tilde{\xi}(\lambda_{q+1},...,\lambda_m)). \tag{5.1}$$

In [31] it was also established that the cascade relations imply the existence of a continuous increasing function $\eta\colon[\tilde{\xi}(1,1),\infty)\to[\xi(1,1),\infty)$ such that for all $(\lambda_1,...,\lambda_m)\in\mathbf{R}_+^m$ such that at least two of the λ_j's are at least 1,

$$\xi(\lambda_1,...,\lambda_m)=\eta(\tilde{\xi}(\lambda_1,...,\lambda_m)). \tag{5.2}$$

In [27], we have determined the exact value of exponents $\tilde{\xi}(\lambda_1,...,\lambda_m)$ for a certain class of numbers $(\lambda_1,...,\lambda_m)$; namely, for all $m\geqslant 2$, and all $(\lambda_1,...,\lambda_m)\in\{\frac{1}{6}l(l+1):l\in\mathbf{N}\}^{m-1}\times\mathbf{R}_+$,

$$\tilde{\xi}(\lambda_1,...,\lambda_m)=\tfrac{1}{24}\big(\big(\sqrt{24\lambda_1+1}+...+\sqrt{24\lambda_m+1}-(m-1)\big)^2-1\big). \tag{5.3}$$

In particular,

$$\tilde{\xi}(1,1,\lambda)=\tfrac{1}{24}\big(\big(8+\sqrt{24\lambda+1}\big)^2-1\big), \tag{5.4}$$

$$\tilde{\xi}(1,1,1,\lambda)=\tfrac{1}{24}\big(\big(12+\sqrt{24\lambda+1}\big)^2-1\big). \tag{5.5}$$

In the next section, we will prove the following two results:

THEOREM 5.1.

$$\xi(1,1)=\tfrac{5}{4}. \tag{5.6}$$

THEOREM 5.2. *For every* $\lambda\geqslant 0$,

$$\xi(1,1,1,\lambda)=\tfrac{1}{48}\big(\big(11+\sqrt{24\lambda+1}\big)^2-4\big). \tag{5.7}$$

Let us now see what these two results directly imply:

THEOREM 5.3. *For all* $x\geqslant 7$,

$$\eta(x)=\tfrac{1}{48}\big(\big(\sqrt{24x+1}-1\big)^2-4\big). \tag{5.8}$$

Moreover, for all $m\geqslant 2$ *and for all* $(\lambda_1,\ldots,\lambda_m)\in\{\frac{1}{6}l(l+1):l\in\mathbf{N}\}^{m-1}\times\mathbf{R}_+$,

$$\xi(\lambda_1,\ldots,\lambda_m)=\tfrac{1}{48}\big(\big(\sqrt{24\lambda_1+1}+\ldots+\sqrt{24\lambda_m+1}-m\big)^2-4\big), \tag{5.9}$$

provided that at least two of the numbers $\lambda_1,\ldots,\lambda_m$ *are at least* 1, *and the right-hand side of* (5.9) *is at least* $\frac{35}{12}$.

In particular, for all $\lambda\geqslant\frac{10}{3}$ and $\lambda'\geqslant 2$,

$$\xi(1,\lambda)=\tfrac{1}{48}\big(\big(3+\sqrt{24\lambda+1}\big)^2-4\big), \tag{5.10}$$

$$\xi(2,\lambda')=\tfrac{1}{48}\big(\big(5+\sqrt{24\lambda'+1}\big)^2-4\big), \tag{5.11}$$

and for all integers $m\geqslant 2$ (using (5.6) for the case $m=2$),

$$\xi(1^{\otimes m})=\tfrac{1}{12}(4m^2-1), \tag{5.12}$$

where $1^{\otimes m}:=(1,\ldots,1)\in\mathbf{N}^m$.

Proof of Theorem 5.3 (assuming (5.6) and (5.7)). Combining (5.5), (5.7) and (5.2) gives (5.8). Hence, we get (5.9) from (5.3), (5.2) and (5.8). □

Proof of Theorems 1.1 *and* 1.2 (assuming (5.6) and (5.7)). In view of the fact that the time exponents differ by a factor of 2 from the space exponents, the theorems follow from Theorem 5.3. □

Remark. In [28], we will show that (5.3) holds for all $(\lambda_1,\ldots,\lambda_m)\in\mathbf{R}_+^m$. The proofs in the present paper can then be very easily adapted to show that (5.9) holds for all $(\lambda_1,\ldots,\lambda_m)\in\mathbf{R}_+^m$ such that at least two of the λ_j's are at least 1.

5.2. Excursion measures

In [31], [32] a characterization of the intersection exponents was given in terms of excursions. The Brownian excursion measures are natural and interesting objects. The utility of the excursion measures in [31], [32] arises from the fact that in the context of excursion measures, one can study Brownian motions without specifying the starting point. This significantly simplifies some arguments and estimates.

Roughly speaking, Brownian excursions in a domain D are Brownian motions started on the boundary, conditioned to immediately enter D, and stopped upon leaving. Since Brownian excursions stay in the domain D, conformal transformations of D can be applied to the excursions.

Let us now describe these Brownian excursions more precisely. For any bounded simply-connected open domain D, there exists a Brownian excursion measure μ_D in D. This is an infinite measure on the set of paths $(B(t), t \leqslant \tau)$ in D such that $B(0,\tau) \subset D$ and $B(0), B(\tau) \in \partial D$ (these endpoints can be viewed as prime ends if necessary). $x_s := B(0)$ and $x_e := B(\tau)$ will denote the starting point and terminal point of the excursion. In this discussion, we will identify two paths (two excursions) when one is obtained by an increasing time change of the other.

One possible definition of μ_D is the following. Consider first $D = \mathbf{U}$, the unit disc. For every $s \in (0,1)$ let P^s be the law of a Brownian motion started uniformly on the circle of radius s, and stopped when it exits $\mathbf{U}$ (modulo continuous increasing time change). Since $z \mapsto \log|z|$ is harmonic, for any $r \in (0,s)$,

$$P^s[B \text{ hits } C_r] = \frac{\log(1/s)}{\log(1/r)}.$$

Set

$$\mu_{\mathbf{U}} := \lim_{s \nearrow 1} \frac{2\pi}{\log(1/s)} P^s,$$

as a weak limit. Note that the $\mu_{\mathbf{U}}$-measure of the set of paths that hit the circle C_r is $2\pi/\log(1/r)$.

One can then check that for any Möbius transformation ϕ from $\mathbf{U}$ onto $\mathbf{U}$, $\phi(\mu_{\mathbf{U}}) = \mu_{\mathbf{U}}$. This makes it possible to extend the definition of μ_D to any simply-connected domain D, by conformal invariance. These Brownian excursions also have a "restriction" property [32], which is a result of the fact that the Brownian paths only feel the boundary of D when they hit it (and then stop).

In [27], we made an extensive use of the Brownian excursion measure in rectangles $R_L = (0,L) \times (0,\pi)$. It is easy to see that the measure μ_{R_L}, restricted to those excursions with starting point on the left-hand side of the rectangle $[0, i\pi]$, is obtained as the limit when $s \to 0$ of $\pi s^{-1} P^s_L$, where P^s_L is the law of a Brownian motion with uniform starting

point on $[s, s+i\pi]$ which is stopped when it exits R_L. In particular, this leads to the following result:

LEMMA 5.4. *Let $\mathcal{E}_L$ denote the event that the Brownian excursion B in R_L crosses the rectangle from the left to the right (i.e., $x_s \in (0, i\pi)$ and $x_e \in (L, L+i\pi)$). Then, when $L \to \infty$,*

$$\mu_{R_L}[\mathcal{E}_L] \asymp e^{-L}. \tag{5.13}$$

Proof. Let h_z denote harmonic measure from z on ∂R_L, where $z \in R_L$. Since $\mathrm{Im}(\exp z)$ is a harmonic function, it easily follows that for all $L>1$ and $z \in (1, 1+i\pi)$,

$$\sin\left(\tfrac{1}{4}\pi\right) h_z\left(\left[L+\tfrac{1}{4}i\pi, L+\tfrac{3}{4}i\pi\right]\right) \leqslant e^{-L} \leqslant \frac{h_z([L, L+i\pi])}{\mathrm{Im}(\exp(z))-1}.$$

It is easy to verify (e.g., by conformal invariance or by a reflection argument) that there is a constant c such that $h_z([L, L+i\pi]) \leqslant c h_z\left(\left[L+\tfrac{1}{4}i\pi, L+\tfrac{3}{4}i\pi\right]\right)$ holds for all $L>2$ and $z \in (1, 1+i\pi)$. Hence, for such L and z,

$$c_1 h_z([L, L+i\pi]) \leqslant e^{-L} \leqslant c_2 h_z([L, L+i\pi])\, 1_{\left\{\mathrm{Im}(z) \in \left[\frac{1}{4}\pi, \frac{3}{4}\pi\right]\right\}},$$

with some constants $c_1, c_2 > 0$. Since the μ_{R_L}-measure of the excursions which reach the line $\{\mathrm{Re}(z)=1\}$ does not depend on L, and from these a fixed proportion first hit $\{\mathrm{Re}(z)=1\}$ in $\left[1+\tfrac{1}{4}\pi, 1+\tfrac{3}{4}\pi\right]$, (5.13) now follows from the Markov property, which is valid for the excursion measures. □

Similarly, one can define Brownian excursions in non-simply-connected domains. For instance, consider the annulus $A(r, 1)$ bounded between the circles C_r and C_1 (where $r \in (0, 1)$). Rather than defining the excursion in $A(r, 1)$ directly, we base the definition on the excursions in $\mathbf{U}$. If γ is a path, let $\Psi_r(\gamma)$ be the initial segment of γ, until the first hit of C_r, or all of γ, if γ does not hit C_r. Now set $\mu_1^r := \Psi_r(\mu_{\mathbf{U}})$. This will be called the Brownian excursion measure on $A(r, 1)$ for excursions started on C. It is clear that $\mu_{\mathbf{U}} = \lim_{r \searrow 0} \mu_1^r$.

The measures μ_D and μ_1^r are also related by restriction and conformal invariance. Suppose that O is a simply-connected subset of $A(r, 1)$ such that each of the sets $\overline{O} \cap C_r$ and $\overline{O} \cap C$ is an arc of positive length. Let L denote the π-extremal distance between these two arcs in O, and let ϕ denote the conformal map from O onto $R_L = (0, L) \times (0, \pi)$ such that $\overline{O} \cap C$ corresponds to $(0, i\pi)$ and $\overline{O} \cap C_r$ corresponds to $(L, L+i\pi)$ under ϕ.

Let $\hat{\mathcal{E}}_1$ be the set of paths starting in C that reach C_r without exiting $\overline{O}$. Consider the image under ϕ of the measure μ_1^r restricted to $\hat{\mathcal{E}}_1$. Then (up to time change) the

image measure is exactly μ_{R_L} restricted to the set of excursions that cross the rectangle from the left to the right. (5.13) therefore shows that when $L\to\infty$,

$$\mu_1^r[\hat{\mathcal{E}}_1] \asymp \exp(-L). \tag{5.14}$$

This may also be easily verified directly.

The mapping $z\mapsto r/z$ maps $A(r,1)$ conformally onto itself. Let μ_r^1 be the image of μ_1^r under this map; this is a measure on Brownian excursions started on C_r and stopped upon leaving $A(r,1)$. By symmetry, it follows that $\mu_r^1[\hat{\mathcal{E}}_1]=\mu_1^r[\hat{\mathcal{E}}_1]\asymp\exp(-L)$ when $L\to\infty$.

Although we will not use this here, it is worthwhile to note that the measures μ_r^1 and μ_1^r agree on the set of paths crossing the annulus, up to time reversal of the path.

5.3. Exponents and excursions

We now describe the intersection exponents in terms of excursions (referring to [31], [32] for the proofs). Let $(B(t), t\leqslant\tau)$ be an excursion in R_L, and, as above, let

$$\mathcal{E}_L := \{\mathrm{Re}(x_s)=0 \text{ and } \mathrm{Re}(x_e)=L\}$$

be the event that B crosses R_L from left to right. Let $\mathfrak{B}$ denote the image of B. When $\mathcal{E}_L$ holds, let O_B^+ be the component of $R_L\setminus\mathfrak{B}$ above $\mathfrak{B}$, and let O_B^- be the component of $R_L\setminus\mathfrak{B}$ below $\mathfrak{B}$. Let $\mathfrak{L}_B^-$ (resp. $\mathfrak{L}_B^+$) denote the π-extremal distance between $[0,x_s]$ and $[L,x_e]$ in O_B^- (resp. $[x_s,i\pi]$ and $[x_e,L+i\pi]$ in O_B^+). (We use script fonts for these $\mathfrak{L}$ to indicate that they are random variables.) Then, for any $\alpha\geqslant 0$ and $\alpha'\geqslant 0$, the exponent $\tilde{\xi}(\alpha,1,\alpha')=\tilde{\xi}(1,\tilde{\xi}(\alpha,\alpha'))$ is characterized by

$$\int_{\mathcal{E}_L}\exp(-\alpha\mathfrak{L}_B^+-\alpha'\mathfrak{L}_B^-)\,d\mu_{R_L}(B)=\exp(-\tilde{\xi}(\alpha',1,\alpha)L+o(L)), \tag{5.15}$$

as $L\to\infty$.

Let $\mathcal{E}_L^2$ be the set of pairs of paths $(B,B')\in\mathcal{E}_L\times\mathcal{E}_L$ such that the trace $\mathfrak{B}'$ of B' is contained in O_B^-. It follows from (5.15), Lemma 5.4 and conformal invariance of the excursion measures that

$$\int_{\mathcal{E}_L^2}\exp(-\alpha\mathfrak{L}_B^+)\,d\mu_{R_L}(B)\,d\mu_{R_L}(B')=\exp(-\tilde{\xi}(1,1,\alpha)L+o(L)), \tag{5.16}$$

as $L\to\infty$. On $\mathcal{E}_L^2$, let $\mathfrak{L}_{B'}^B$ be the π-extremal distance from $[0,i\pi]$ to $[L,L+i\pi]$ in the domain between $\mathfrak{B}$ and $\mathfrak{B}'$. Given B', it is clear by conformal invariance and the restriction property of the excursion measure that $1_{\mathcal{E}_L^2}\mathfrak{L}_{B'}^B$ has the same law as $1_{\mathcal{E}_L^2}\mathfrak{L}_B^+$. Consequently, (5.16) gives

$$\int_{\mathcal{E}_L^2}\exp(-\alpha\mathfrak{L}_{B'}^B)\,d\mu_{R_L}(B)\,d\mu_{R_L}(B')=\exp(-\tilde{\xi}(1,1,\alpha)L+o(L)). \tag{5.17}$$

Similarly, one can characterize the exponents ξ in terms of the excursion measure μ_r^1. For any $r<1$, consider two independent excursions B and B' of the annulus $A(r,1)$. Define the events

$$\mathcal{E}=\mathcal{E}(r):=\{\mathfrak{B} \text{ crosses the annulus without separating } C_r \text{ from } C\},$$
$$\tilde{\mathcal{E}}=\tilde{\mathcal{E}}(r):=\{\mathfrak{B} \text{ and } \mathfrak{B}' \text{ are disjoint and both cross the annulus}\}.$$

When $\mathcal{E}$ holds, let $\mathfrak{L}_B$ be the π-extremal distance between C_r and C in $A(r,1)\backslash\mathfrak{B}$. Similarly, when $\tilde{\mathcal{E}}$ is satisfied, let O_1 and O_2 be the two components of $A(r,1)\backslash(\mathfrak{B}\cup\mathfrak{B}')$ which have arcs of C on their boundaries, in such a way that the sequence $\mathfrak{B}, O_1, \mathfrak{B}', O_2$ is in counterclockwise order around $A(r,1)$. Let $\mathfrak{L}_1$ be the π-extremal distance from C_r to C in O_1, and let $\mathfrak{L}_2$ be the corresponding quantity for O_2. Then for $\lambda, \lambda_1, \lambda_2 \in \mathbf{R}_+$, the exponents $\xi(1,\lambda)$ and $\xi(1,\lambda_1,1,\lambda_2)$ can be described [32] as

$$\int_{\mathcal{E}} \exp(-\lambda\mathfrak{L})\, d\mu_r^1(B) \approx r^{\xi(1,\lambda)} \tag{5.18}$$

and

$$\int_{\tilde{\mathcal{E}}} \exp(-\lambda_1\mathfrak{L}_1-\lambda_2\mathfrak{L}_2)\, d\mu_r^1(B)\, d\mu_r^1(B') \approx r^{\xi(1,\lambda_1,1,\lambda_2)}, \tag{5.19}$$

as $r\searrow 0$.

5.4. A useful technical lemma

We now derive a technical refinement of (5.18) and (5.19) (in the case $\lambda_1=1$) that will be useful to identify the Brownian intersection exponents with those computed for SLE_6.

Keep the same notation as above, and on $\mathcal{E}$, let ϕ denote the conformal map that maps O onto $R_{\mathfrak{L}}$ in such a way that the images of $C_1\cap\overline{O}$ and $C_r\cap\overline{O}$ are mapped onto the left and right edges of the rectangle, respectively. Similarly, on $\tilde{\mathcal{E}}$, let ϕ_1 be the corresponding conformal map from O_1 onto $R_{\mathfrak{L}_1}$. For all $\alpha>0$ set

$$\mathcal{H}_\alpha := \mathcal{E}\cap\{i\in\overline{O} \text{ and } \phi(i)\in[i\alpha, i(\pi-\alpha)]\},$$
$$\widetilde{\mathcal{H}}_\alpha := \tilde{\mathcal{E}}\cap\{i\in\overline{O}_1 \text{ and } \phi_1(i)\in[i\alpha, i(\pi-\alpha)]\}.$$

LEMMA 5.5. *Let $\lambda>0$. Then there are sequences $x_n\searrow 0$ and $y_n\searrow 0$, and an $\alpha>0$, such that*

$$\int_{\mathcal{H}_\alpha} \exp(-\lambda\mathfrak{L}_B)\, d\mu_{x_n}^1(B) \approx (x_n)^{\xi(1,\lambda)}, \quad n\to\infty, \tag{5.20}$$

$$\int_{\widetilde{\mathcal{H}}_\alpha} \exp(-\mathfrak{L}_1-\lambda\mathfrak{L}_2)\, d\mu_{y_n}^1(B)\, d\mu_{y_n}^1(B') \approx (y_n)^{\xi(1,1,1,\lambda)}, \quad n\to\infty. \tag{5.21}$$

Actually (see, e.g., [30]), much stronger statements hold: In the above, $\approx$ may be replaced by $\asymp$, and these statements hold for every sequence tending to ∞. But the present statement will be sufficient to determine the values of the Brownian intersection exponents, and it can be easily proved as follows.

Proof. We will only give the detailed proof of (5.21). The proof of (5.20) is easier, follows exactly the same lines, and is safely left to the reader.

Because of (5.19), it suffices to find the lower bound for the left-hand side of (5.21). Let us first introduce some notation. Let $r\in(0,\frac{1}{4})$, and consider the measure $\mu_r^1\times\mu_r^1$ on the space of pairs (B,B'). Let

$$f(r):=\int_{\tilde{\mathcal{E}}(r)}\exp(-\mathfrak{L}_1-\lambda\mathfrak{L}_2)\,d\mu_r^1(B)\,d\mu_r^1(B').$$

Let B^* be the path B stopped when it hits $C_{1/4}$, if it does, and $B^*=B$, if it does not hit $C_{1/4}$. Similarly, define B'^* from B'. Let $\mathfrak{B}$ and $\mathfrak{B}'$ be the traces of B and B', respectively. Let $\tilde{\mathcal{E}}^*$ be the event that the traces of B^* and B'^* are disjoint. On $\tilde{\mathcal{E}}^*$, let O_1^* and O_2^* be the domains defined for B^* and B'^*, as O_1 and O_2 were defined for B and B'. Then $O_j^*\subset O_j$ on $\tilde{\mathcal{E}}$, $j=1,2$. For $j=1,2$, on $\tilde{\mathcal{E}}^*$, let $\mathfrak{L}_j^*$ be the π-extremal distance between C_r and $C_{1/4}$ in O_j^*. Otherwise, set $\mathfrak{L}_j^*=\infty$.

For $a>0$, let $\mathcal{D}_a$ be the event that the distance between $\mathfrak{B}\cap A(\frac{1}{2},1)$ and $\mathfrak{B}'\cap A(\frac{1}{2},1)$ is at least a. Suppose that $a\in(0,\frac{1}{5})$. Observe that for $(B,B')\notin\mathcal{D}_a$, there is for $j=1$ or $j=2$ a path of length at most a in $O_j\cap A(\frac{1}{2},1)$ which separates C_r from C_1 in O_j. It then follows that $\mathfrak{L}_j\geqslant\mathfrak{L}_j^*+c_1\log(1/a)$, for some constant $c_1>0$. Consequently,

$$\begin{aligned}\int_{\tilde{\mathcal{E}}(r)\setminus\mathcal{D}_a}&\exp(-\mathfrak{L}_1-\lambda\mathfrak{L}_2)\,d\mu_r^1(B)\,d\mu_r^1(B')\\&\leqslant a^{c_1\min\{\lambda,1\}}\int_{\tilde{\mathcal{E}}^*}\exp(-\mathfrak{L}_1^*-\lambda\mathfrak{L}_2^*)\,d\mu_r^1(B)\,d\mu_r^1(B')\\&=a^{c_1\min\{\lambda,1\}}f(4r),\end{aligned}\tag{5.22}$$

since the image of μ_r^1 under the map $B\mapsto 4B^*$ is μ_{4r}^1.

By (5.19), we have $f(r)\approx r^{\xi(1,1,\lambda,1)}$ (and f is non-decreasing). Consequently, if a is chosen sufficiently small, there is a sequence $y_n\searrow 0$ (for instance a subsequence of 4^{-n}) such that $f(y_n)\geqslant 2a^{c_1\min\{\lambda,1\}}f(4y_n)$. For these y_n, (5.22) gives

$$\int_{\tilde{\mathcal{E}}(y_n)\cap\mathcal{D}_a}\exp(-\mathfrak{L}_1-\lambda\mathfrak{L}_2)\,d\mu_{y_n}^1(B)\,d\mu_{y_n}^1(B')\geqslant a^{c_1\min\{\lambda,1\}}f(4y_n).\tag{5.23}$$

Fix such an a and such a y_n. Let $\mathcal{I}_a$ be the event that $i\in\bar{O}_1$ and the distance from i to $\mathfrak{B}\cup\mathfrak{B}'$ is at least $\frac{1}{10}a$. Observe that if we apply an independent random uniform rotation

about 0 to a pair $(B,B')\in\mathcal{D}_a\cap\tilde{\mathcal{E}}$, then with probability at least $a/10\pi$ the rotated pair is in $\mathcal{I}_a$. Since the integrand in (5.23) is invariant under rotations, (5.23) and (5.19) give

$$\int_{\tilde{\mathcal{E}}(y_n)\cap\mathcal{I}_a}\exp(-\mathfrak{L}_1-\lambda\mathfrak{L}_2)\,d\mu^1_{y_n}(B)\,d\mu^1_{y_n}(B')\approx(y_n)^{\xi(1,1,\lambda,1)}.$$

It therefore suffices to show that when $\alpha>0$ is small, we have $\mathcal{I}_a\cap\tilde{\mathcal{E}}\subset\widetilde{\mathcal{H}}_\alpha$. To prove this, consider a pair $(B,B')\in\mathcal{I}_a$. Let A be the subarc of $\overline{O}_1\cap C_1$ that has i as one endpoint and the other endpoint is in $\mathfrak{B}$, and let A' be the subarc of $\overline{O}_1\cap C_1$ that has i as one endpoint and the other endpoint is in $\mathfrak{B}'$. Then the extremal distance from A to $\mathfrak{B}'$ in O_1 is bounded from below by a positive constant depending only on a, as is the extremal distance from A' to $\mathfrak{B}$ in O_1. Conformal invariance of extremal distance therefore shows that the distance from $\phi(i)$ to $\{0,i\pi\}$ is bounded from below by a constant depending only on a, which proves that $(B,B')\in\widetilde{\mathcal{H}}_\alpha$, where $\alpha>0$ depends only on a. □

6. The universality argument

We are now ready to combine the results derived so far to prove our main theorems. As in [27], we follow the universality ideas presented in [32]. First, Theorem 5.6 $\left(\xi(1,1)=\frac{5}{4}\right)$ will be proved, followed by Theorem 5.7 (giving $\xi(1,1,1,\lambda)$). As we have seen, Theorems 1.1 and 1.2 are immediate consequences.

6.1. Proof of $\xi(1,1)=\frac{5}{4}$

Let $r>0$, let K_t be a radial SLE_6-process in $\mathbf{U}$ starting at i, and let $T=T(r)$ be the least t such that $\overline{K}_t\cap C_r\neq\varnothing$. Set $\mathfrak{K}:=\overline{K}_T$, let $\mathbf{P}_r$ denote the law of $\mathfrak{K}$, and let $\mathbf{E}_r$ denote expectation with respect to this measure. As before, we let μ^1_r denote the Brownian excursion measure in $A(r,1)$ started from C_r, and let $\mathfrak{B}$ denote the trace of the excursion B. We are interested in the event $\mathcal{E}^*$ in which $\mathfrak{K}\cap\mathfrak{B}=\varnothing$ and $\mathfrak{B}$ crosses $A(r,1)$ (i.e., $\mathfrak{B}\cap C_1\neq\varnothing$).

The proof will proceed by computing $(\mathbf{P}_r\times\mu^1_r)[\mathcal{E}^*]$ in two different ways. In the first computation, we begin by conditioning on $\mathfrak{K}$ and then taking the expectation, while the second computation begins by conditioning on B.

When $\mathfrak{B}$ does not separate C_r from C in $\mathbf{U}$, let O_B denote the connected component of $A(r,1)\setminus\mathfrak{B}$ that touches both circles C_r and C, and $\mathfrak{L}_B$ the corresponding π-extremal distance.

When $\mathfrak{K}$ does not disconnect C_r from C in $\mathbf{U}$, let $O_{\mathfrak{K}}$ denote the connected component of $A(r,1)\setminus\mathfrak{K}$ that touches both circles, and $\mathfrak{L}_{\mathfrak{K}}$ the corresponding π-extremal distance.

Suppose first that $\hat{\mathcal{R}}$ is given and that $\mathfrak{L}_{\hat{\mathcal{R}}}<\infty$. Note that $\mathfrak{L}_{\hat{\mathcal{R}}}\to\infty$ as $r\searrow 0$. By the restriction property and conformal invariance of the Brownian excursion measure (see §5.2) and Lemma 5.4,

$$\mu_r^1[B \text{ crosses the annulus in } O_{\hat{\mathcal{R}}}] = \mu_{R_{\mathfrak{L}_{\hat{\mathcal{R}}}}}[\mathcal{E}_{\mathfrak{L}_{\hat{\mathcal{R}}}}] \asymp \exp(-\mathfrak{L}_{\hat{\mathcal{R}}}), \quad r\searrow 0,$$

where $\mathcal{E}_L$ is as in Lemma 5.4. On the other hand, we know from Theorem 3.1, with $b=1$, $\kappa=6$, that

$$\mathbf{E}_r[\exp(-\mathfrak{L}_{\hat{\mathcal{R}}})] \approx r^{5/4}, \quad r\searrow 0.$$

Combining these two facts implies that

$$(\mathbf{P}_r\times\mu_r^1)[\mathcal{E}^*] \approx r^{5/4}, \quad r\searrow 0. \tag{6.1}$$

Suppose now that B is given and that it does cross the annulus without separating the disk C_r from C. The second part of Lemma 4.3 shows that there exists a constant $c>0$ such that

$$\mathbf{P}_r[\mathcal{E}^*] \leqslant c\exp(-\mathfrak{L}_B).$$

Hence, combining this with (5.18) shows that

$$(\mathbf{P}_r\times\mu_r^1)[\mathcal{E}^*] \leqslant r^{\xi(1,1)+o(1)}. \tag{6.2}$$

On the other hand, suppose now that B is given and that $B\in\mathcal{H}_\alpha$, where $\mathcal{H}_\alpha$ is as in Lemma 5.5. Then by the first part of Lemma 4.3, there exists a constant c' such that

$$\mathbf{P}_r[\mathcal{E}^*] \geqslant c'\exp(-\mathfrak{L}_B).$$

Combining this with Lemma 5.5, we find that

$$(\mathbf{P}_{x_n}\times\mu_{x_n}^1)[\mathcal{E}^*] \geqslant (\mathbf{P}_{x_n}\times\mu_{x_n}^1)[\mathcal{H}_\alpha\cap\mathcal{E}^*] \geqslant (x_n)^{\xi(1,1)+o(1)}, \quad n\to\infty.$$

Comparing with (6.2) gives

$$(\mathbf{P}_{x_n}\times\mu_{x_n}^1)[\mathcal{E}^*] \approx (x_n)^{\xi(1,1)}, \quad n\to\infty.$$

Now, by (6.1),

$$(x_n)^{5/4} \approx (\mathbf{P}_{x_n}\times\mu_{x_n}^1)[\mathcal{E}^*] \approx (x_n)^{\xi(1,1)}$$

when $n\to\infty$, which proves $\xi(1,1)=\frac{5}{4}$.

6.2. The determination of $\xi(1,1,1,\lambda)$

The goal now is to prove (5.7). As $\lambda \mapsto \xi(1,1,1,\lambda) = \xi(1,1,\tilde{\xi}(1,\lambda))$ is continuous on $[0,\infty)$ (see [31]), we can restrict ourselves to the case where $\lambda > 0$. The proof goes along similar lines as the proof of $\xi(1,1) = \frac{5}{4}$. This time, we will consider two independent Brownian excursions B and B' in the annulus $A(r,1)$, and one SLE_6 $\mathfrak{K}$ as before. We are interested in the event that B, B' and $\mathfrak{K}$ all cross the annulus, that they remain disjoint, and that B, $\mathfrak{K}$ and B' are in clockwise order. Let us call this event l. Note that in this case, $\mathfrak{K}$ crosses the annulus in O_1, where we use the notations of §5.3.

We shall compute in two different ways the quantity

$$\int_l \exp(-\lambda \mathfrak{L}_2)\, d\mu_r^1(B)\, d\mu_r^1(B')\, d\mathbf{P}_r(\mathfrak{K}).$$

On the one hand, we know from conformal invariance and the restriction property of Brownian excursions and from (5.17) that when $\mathfrak{K}$ crosses the annulus without separating C_r from C,

$$\int_l \exp(-\lambda \mathfrak{L}_2)\, d\mu_r^1(B)\, d\mu_r^1(B') \approx \exp(-\mathfrak{L}_{\mathfrak{K}} \tilde{\xi}(1,1,\lambda)), \quad r \searrow 0. \tag{6.3}$$

But we know from Theorem 3.1 that

$$\int \exp(-b\mathfrak{L}_{\mathfrak{K}})\, d\mathbf{P}_r \approx r^{\nu}, \tag{6.4}$$

where

$$\nu = \nu(b) = \tfrac{1}{8}\bigl(4b+1+\sqrt{1+24b}\bigr).$$

Note that ν is a continuous increasing function of b on $(0,\infty)$. Consequently, (6.3) and (6.4) give

$$\int_l \exp(-\lambda \mathfrak{L}_2)\, d\mu_r^1(B)\, d\mu_r^1(B')\, d\mathbf{P}_r(\mathfrak{K}) \approx r^{\nu(\tilde{\xi}(1,1,\lambda))}.$$

Combining this with (5.4) shows that

$$\int_l \exp(-\lambda \mathfrak{L}_2)\, d\mu_r^1(B)\, d\mu_r^1(B')\, d\mathbf{P}_r(\mathfrak{K}) \approx r^{a}, \tag{6.5}$$

where

$$a = a(\lambda) := \nu(\tilde{\xi}(1,1,\lambda)) = \tfrac{1}{48}\bigl(\bigl(11+\sqrt{24\lambda+1}\bigr)^2 - 4\bigr).$$

Suppose now that B and B' are given and that $\tilde{\mathcal{E}}$ is satisfied; i.e., that B and B' cross the annulus without intersecting each other. Then Lemma 4.3 shows that

$$\mathbf{P}_r[\mathfrak{K} \subset O_1] \leqslant c \exp(-\mathfrak{L}_1),$$

for some constant c. Hence, combining this with (5.19) shows that

$$\int_l \exp(-\lambda \mathfrak{L}_2)\, d\mathbf{P}_r(\mathfrak{K})\, d\mu_r^1(B)\, d\mu_r^1(B') \leqslant r^{\xi(1,1,1,\lambda)+o(1)}. \tag{6.6}$$

Suppose now that B and B' are given and that $(B, B') \in \widetilde{\mathcal{H}}_\alpha$. Then Lemma 4.3 shows that there exists a constant $c'>0$ such that

$$\mathbf{P}_r[\mathfrak{K} \subset O_1] \geqslant c' \exp(-\mathfrak{L}_1).$$

Combining this with Lemma 5.5 gives

$$\int_l \exp(-\lambda \mathfrak{L}_2)\, d\mathbf{P}_{y_n}(\mathfrak{K})\, d\mu_{y_n}^1(B)\, d\mu_{y_n}^1(B') \geqslant (y_n)^{\xi(1,1,1,\lambda)+o(1)}.$$

Comparing with (6.6) implies that

$$\int_l \exp(-\lambda \mathfrak{L}_2)\, d\mathbf{P}_{y_n}(\mathfrak{K})\, d\mu_{y_n}^1(B)\, d\mu_{y_n}^1(B') \approx (y_n)^{\xi(1,1,1,\lambda)},$$

when $n \to \infty$. Consequently, by (6.5), $\xi(1,1,1,\lambda) = a(\lambda)$. □

References

[1] AHLFORS, L. V., *Conformal Invariants: Topics in Geometric Function Theory*. McGraw-Hill, New York, 1973.

[2] AIZENMAN, M., DUPLANTIER, B. & AHARONY, A., Path crossing exponents and the external perimeter in 2D percolation. *Phys. Rev. Lett.*, 83 (1999), 1359–1362.

[3] AZENCOTT, R., Behaviour of diffusion semi-groups at infinity. *Bull. Soc. Math. France*, 102 (1974), 193–240.

[4] BISHOP, C. J., JONES, P. W., PEMANTLE, R. & PERES, Y., The dimension of the Brownian frontier is greater than 1. *J. Funct. Anal.*, 143 (1997), 309–336.

[5] BURDZY, K. & LAWLER, G. F., Non-intersection exponents for Brownian paths. Part I: Existence and an invariance principle. *Probab. Theory Related Fields*, 84 (1990), 393–410.

[6] — Non-intersection exponents for Brownian paths. Part II: Estimates and applications to a random fractal. *Ann. Probab.*, 18 (1990), 981–1009.

[7] CARDY, J. L., Conformal invariance and surface critical behavior. *Nuclear Phys. B*, 240 (1984), 514–532.

[8] — Critical percolation in finite geometries. *J. Phys. A*, 25 (1992), L201–L206.

[9] — The number of incipient spanning clusters in two-dimensional percolation. *J. Phys. A*, 31 (1998), L105.

[10] CRANSTON, M. & MOUNTFORD, T., An extension of a result of Burdzy and Lawler. *Probab. Theory Related Fields*, 89 (1991), 487–502.

[11] DUPLANTIER, B., Random walks and quantum gravity in two dimensions. *Phys. Rev. Lett.*, 81 (1998), 5489–5492.

[12] — Two-dimensional copolymers and exact conformal multifractality. *Phys. Rev. Lett.*, 82 (1999), 880–883.

[13] — Harmonic measure exponents for two-dimensional percolation. *Phys. Rev. Lett.*, 82 (1999), 3940–3943.

[14] DUPLANTIER, B. & KWON, K.-H., Conformal invariance and intersection of random walks. *Phys. Rev. Lett.*, 61 (1988), 2514–2517.

[15] DUPLANTIER, B. & SALEUR, H., Exact determination of the percolation hull exponent in two dimensions. *Phys. Rev. Lett.*, 58 (1987), 2325–2328.

[16] KENYON, R., Conformal invariance of domino tiling. *Ann. Probab.*, 28 (2000), 759–795.

[17] — Long-range properties of spanning trees. Probabilistic techniques in equilibrium and nonequilibrium statistical physics. *J. Math. Phys.*, 41 (2000), 1338–1363.

[18] — The asymptotic determinant of the discrete Laplacian. *Acta Math.*, 185 (2000), 239–286.

[19] LAWLER, G. F., *Intersections of Random Walks*. Birkhäuser Boston, Boston, MA, 1991.

[20] — Hausdorff dimension of cut points for Brownian motion. *Electron. J. Probab.*, 1:2 (1996), 1–20 (electronic).

[21] — Cut times for simple random walk. *Electron. J. Probab.*, 1:13 (1996), 1–24 (electronic).

[22] — The dimension of the frontier of planar Brownian motion. *Electron. Comm. Probab.*, 1:5 (1996), 29–47 (electronic).

[23] — The frontier of a Brownian path is multifractal. Preprint, 1997.

[24] — Strict concavity of the intersection exponent for Brownian motion in two and three dimensions. *Math. Phys. Electron. J.*, 4:5 (1998), 1–67 (electronic).

[25] — Geometric and fractal properties of Brownian motion and random walk paths in two and three dimensions, in *Random Walks* (Budapest, 1998), pp. 219–258. Bolyai Soc. Math. Stud., 9. János Bolyai Math. Soc., Budapest, 1999.

[26] LAWLER, G. F. & PUCKETTE, E. E., The intersection exponent for simple random walk. *Combin. Probab. Comput.*, 9 (2000), 441–464.

[27] LAWLER, G. F., SCHRAMM, O. & WERNER, W., Values of Brownian intersection exponents, I: Half-plane exponents. *Acta Math.*, 187 (2001), 237–273.
[28] — Values of Brownian intersection exponents, III: Two-sided exponents. To appear in *Ann. Inst. H. Poincaré Probab. Statist.* `http://arxiv.org/abs/math.PR/0005294`.
[29] — Analyticity of intersection exponents for planar Brownian motion. To appear in *Acta Math.*, 188 (2002). `http://arxiv.org/abs/math.PR/0005295`.
[30] — Sharp estimates for Brownian non-intersection probabilities. To appear in *In and Out of Equilibrium. Probability with a Physics Flavor.* Progr. Probab. Birkhäuser Boston, Boston, MA.
[31] LAWLER, G. F. & WERNER, W., Intersection exponents for planar Brownian motion. *Ann. Probab.*, 27 (1999), 1601–1642.
[32] — Universality for conformally invariant intersection exponents. *J. Eur. Math. Soc. (JEMS)*, 2 (2000), 291–328.
[33] LÖWNER, K., Untersuchungen über schlichte konforme Abbildungen des Einheitskreises, I. *Math. Ann.*, 89 (1923), 103–121.
[34] MADRAS, N. & SLADE, G., *The Self-Avoiding Walk.* Birkhäuser Boston, Boston, MA, 1993.
[35] MANDELBROT, B. B., *The Fractal Geometry of Nature.* Freeman, San Francisco, CA, 1982.
[36] NIENHUIS, B., Critical behavior of two-dimensional spin models and charge asymmetry in the Coulomb gas. *J. Statist. Phys.*, 34 (1984), 731–761.
[37] POMMERENKE, CH., On the Loewner differential equation. *Michigan Math. J.*, 13 (1966), 435–443.
[38] REVUZ, D. & YOR, M., *Continuous Martingales and Brownian Motion.* Grundlehren Math. Wiss., 293. Springer-Verlag, Berlin, 1991.
[39] SCHRAMM, O., Scaling limits of loop-erased random walks and uniform spanning trees. *Israel J. Math.*, 118 (2000), 221–288.
[40] WERNER, W., Bounds for disconnection exponents. *Electron. Comm. Probab.*, 1:4 (1996), 19–28 (electronic).

GREGORY F. LAWLER
Department of Mathematics
Duke University
Box 90320
Durham, NC 27708-0320
U.S.A.
jose@math.duke.edu

ODED SCHRAMM
Microsoft Research
1, Microsoft Way
Redmond, WA 98052
U.S.A.
schramm@microsoft.com

WENDELIN WERNER
Département de Mathématiques
Université Paris-Sud
Bât. 425
FR-91405 Orsay Cedex
France
wendelin.werner@math.u-psud.fr

Received May 18, 2000

The Annals of Probability
2004, Vol. 32, No. 1B, 939–995
© Institute of Mathematical Statistics, 2004

CONFORMAL INVARIANCE OF PLANAR LOOP-ERASED RANDOM WALKS AND UNIFORM SPANNING TREES

BY GREGORY F. LAWLER,[1] ODED SCHRAMM AND WENDELIN WERNER

Cornell University, Microsoft Corporation and Université Paris-Sud

This paper proves that the scaling limit of a loop-erased random walk in a simply connected domain $D \subsetneq \mathbb{C}$ is equal to the radial SLE_2 path. In particular, the limit exists and is conformally invariant. It follows that the scaling limit of the uniform spanning tree in a Jordan domain exists and is conformally invariant. Assuming that ∂D is a C^1-simple closed curve, the same method is applied to show that the scaling limit of the uniform spanning tree Peano curve, where the tree is wired along a proper arc $A \subset \partial D$, is the chordal SLE_8 path in $\overline{D}$ joining the endpoints of A. A by-product of this result is that SLE_8 is almost surely generated by a continuous path. The results and proofs are not restricted to a particular choice of lattice.

1. Introduction.

1.1. *Motivation from statistical physics.* One of the main goals of both probability theory and statistical physics is to understand the asymptotic behavior of random systems when the number of microscopic random inputs goes to ∞. These random inputs can be independent, such as a sequence of independent random variables, or dependent, as in the Ising model. Often, one wishes to understand these systems via some relevant "observables" that can be of a geometric or analytic nature. In order to understand this asymptotic behavior, one can attempt to prove convergence toward a suitable continuous model. The simplest and most important example of such random continuous models is Brownian motion, which is the scaling limit of random walks. In particular, a simple random walk on any lattice in $\mathbb{R}^d$ converges to (a linear image of) Brownian motion in the scaling limit.

Physicists and chemists have observed that critical systems (i.e., systems at their phase transition point) can exhibit macroscopic randomness. Hence, various quantities related to the corresponding lattice models should converge as the mesh is refined. In fact, one of the important starting points for theoretical physicists working on two-dimensional critical models is the assumption that the continuous limit is independent of the lattice and, furthermore, displays conformal invariance. This assumption has enabled them to develop and use techniques from conformal field theory to predict the exact values of certain critical exponents. Until very

Received February 2002; revised March 2003.

[1]Supported in part by the NSF and the Mittag-Leffler Institute.

AMS 2000 subject classifications. 82B41.

Key words and phrases. Loop-erased random walk, uniform spanning trees, stochastic Loewner evolution.

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_30, © Springer Science+Business Media, LLC 2011

recently, the existence of the limit, its conformal invariance and the derivation of the exponents assuming conformal invariance remained beyond mathematical justification for the basic lattice models in critical phenomena, such as percolation, the Ising model and random cluster measures. Although there are many interesting questions about higher-dimensional systems, we will limit our discussion to two dimensions where conformal invariance plays an essential role.

1.2. *Recent progress.* In [38], a one-parameter family of random growth processes (loosely speaking, random curves) in two dimensions was introduced. The growth process is based on Loewner's differential equation, where the driving term is time-scaled one-dimensional Brownian motion, and is therefore called stochastic Loewner evolution, or SLE_κ. The parameter $\kappa \geq 0$ of SLE is the time scaling constant for the driving Brownian motion. It was conjectured that the scaling limit of the loop-erased random walk (LERW) is SLE_2, and this conjecture was proved to be equivalent to the conformal invariance of the LERW scaling limit (see [38]). The argument given was quite general and shows that a conformally invariant random path satisfying a mild Markovian property, which will be described below, must be SLE. On this basis, it was also conjectured there that the scaling limits of the critical percolation interface and the uniform spanning tree Peano curve are the paths of SLE_6 and SLE_8, respectively, and it was claimed that conformal invariance is sufficient to establish these conjectures. (For additional conjectures regarding curves tending to SLE, including the interfaces in critical random cluster models—also called FK percolation models—for $q \in [0, 4]$, see [36].)

At some values of the parameter κ, SLE has some remarkable properties. For instance, SLE_6 has a locality property (see [27]) that makes it possible to relate its outer boundary to that of planar Brownian motion. This has led to the proof of conjectures concerning planar Brownian motion and simple random walks (see [27]–[29]).

In [40] and [41], Smirnov recently proved the existence and conformal invariance of the scaling limit of critical site percolation on the two-dimensional triangular lattice: he managed to prove Cardy's formula (see [5]) which is a formula for the limit of the probability of a percolation crossing between two arcs on the boundary of the domain. Combining this information with the independence properties of percolation, Smirnov then showed that the scaling limit of the percolation interface is SLE_6. This has led to the rigorous determination of critical exponents for this percolation model (see [30] and [42]).

1.3. *LERW and UST defined.* The uniform spanning tree (UST), which can be interpreted as the $q = 0$ critical random cluster model (see [15]), is a dependent model that has many remarkable features. In particular, it is very closely related to the loop-erased random walk, whose definition (see [24]) we now briefly recall.

Consider any finite or recurrent connected graph G, a vertex a and a set of vertices V. A *loop-erased random walk* (LERW) from a to V is a random simple curve joining a to V obtained by erasing the loops in chronological order from a simple random walk started at a and stopped upon hitting V. In other words, if $(\Gamma(n), 0 \le n \le T)$ is a simple random walk on G started from a and stopped at its first hitting time T of V, the loop erasure $\beta = (\beta_0, \ldots, \beta_\ell)$ is defined inductively as follows: $\beta_0 = a$; if $\beta_n \in V$, then $n = \ell$; and otherwise $\beta_{n+1} = \Gamma(k)$, where $k = 1 + \max\{m \le T : \Gamma(m) = \beta_n\}$.

A spanning tree T of a connected graph G is a subgraph of G such that for every pair of vertices v, u in G there is a unique simple path (i.e., self-avoiding) in T with these vertices as endpoints. A *uniform spanning tree* (UST) in a finite, connected graph G is a sample from the uniform probability measure on spanning trees of G. It has been shown in [34] that the law of the self-avoiding path with endpoints a and b in the UST is the same as that of the LERW from a to $\{b\}$. See Figure 1.

Wilson [44] established an even stronger connection between LERW and UST by giving an algorithm to generate UST's using LERW. Wilson's algorithm runs as follows. Pick an arbitrary ordering $v_0, v_1, \ldots, v_m$ for the vertices in G. Let $T_0 = \{v_0\}$. Inductively, for $n = 1, 2, \ldots, m$, define T_n to be the union of T_{n-1} and a (conditionally independent) LERW path from v_n to T_{n-1}. (If $v_n \in T_{n-1}$, then $T_n = T_{n-1}$.) Then, regardless of the chosen order of the vertices, T_m is a UST on G.

Wilson's algorithm gives a natural extension of the definition of UST to infinite recurrent graphs. In fact, for transient graphs, there are two natural definitions which often coincide, but this interesting theory is somewhat removed from the topic of this paper. Many striking properties of UST and LERW have been

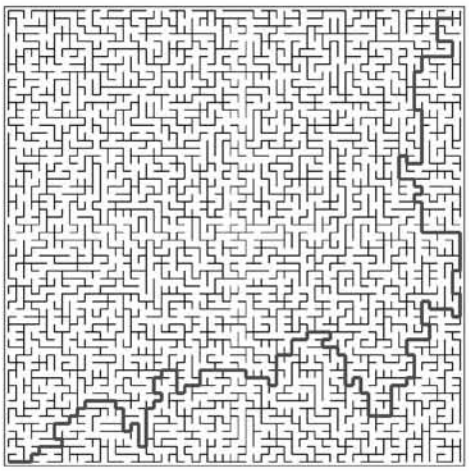


FIG. 1. *The LERW in the UST.*

discovered. See [32] for a survey of UST's and [26] for a survey of properties of LERW in $\mathbb{Z}^d$, $d > 2$.

Exploiting a link with domino tilings and deriving discrete analogs of Cauchy–Riemann equations, Kenyon (see [20] and [21]) rigorously established the values of various critical exponents predicted for the LERW (see [11], [14] and [32]) in two dimensions. In particular, he showed that the expected number of steps of a LERW joining two corners of the $N \times N$ square in the square grid $\mathbb{Z}^2$ is of the order of magnitude of $N^{5/4}$. He also showed conformal invariance for the leading term in the asymptotics of the probability that the LERW contains a given edge. This was the first mathematical evidence for full conformal invariance of the LERW scaling limit.

In [3] and [1] subsequential scaling limits of the UST measures in $\mathbb{Z}^d$ were shown to exist, using a compactness argument. Moreover, these papers prove that all the paths in the scaling limit that intersect a fixed bounded region are uniformly Hölder continuous. In [38] the topology of subsequential scaling limits of the UST on $\mathbb{Z}^2$ was determined. In particular, it was shown that every subsequential scaling limit of the LERW is a simple path.

1.4. *A short description of SLE.* We now briefly describe SLE; precise definitions are deferred to Section 2.1. *Chordal* SLE is a random growing family of compact sets K_t, $t \in [0, \infty)$, in the closure $\overline{\mathbb{H}}$ of the upper half plane $\mathbb{H}$. The evolution of K_t is given by the Loewner differential equation with "driving function" Brownian motion. From [36], it is known that when $\kappa \neq 8$ the process is described by a random curve $\gamma : [0, \infty) \to \overline{\mathbb{H}}$, in the sense that, for every $t \geq 0$, $\mathbb{H} \setminus K_t$ is the unbounded component of $\mathbb{H} \setminus \gamma[0, t]$. A corollary of our results is that this holds for $\kappa = 8$ as well. The curve γ satisfies $\gamma(0) = 0$ and $\lim_{t \to \infty} \gamma(t) = \infty$. If $\kappa \leq 4$, then γ is a simple curve and $K_t = \gamma[0, t]$.

There is another version of SLE called *radial* SLE. Radial SLE also satisfies the description above, except that the upper half plane $\mathbb{H}$ is replaced by the unit disk $\mathbb{U}$, $\gamma(0)$ is on the unit circle $\partial\mathbb{U}$ and $\lim_{t \to \infty} \gamma(t) = 0$.

Both radial and chordal versions of SLE may be defined in an arbitrary simply connected domain $D \subsetneq \mathbb{C}$ by mapping over to D using a fixed conformal map ϕ from $\mathbb{H}$ or $\mathbb{U}$ to D.

1.5. *The main results of the paper.* Let $D \subsetneq \mathbb{C}$ be a simply connected domain with $0 \in D$. For $\delta > 0$, let μ_δ be the law of the loop erasure of a simple random walk on the grid $\delta\mathbb{Z}^2$, started at 0 and stopped when it hits ∂D. See Figure 2. Let ν be the law of the image of the radial SLE_2 path under a conformal map from the unit disk $\mathbb{U}$ to D fixing 0. When the boundary of D is very rough, the conformal map from $\mathbb{U}$ to D might not extend continuously to the boundary, but the proof of the following theorem in fact shows that even in this case the image of the SLE_2 path has a unique endpoint on ∂D.

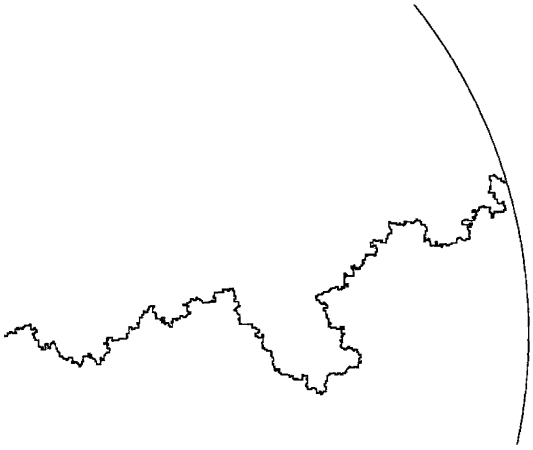


FIG. 2. *A sample of the loop-erased random walk; proved to converge to radial* SLE_2.

On the space of unparameterized paths in $\mathbb{C}$, consider the metric $\rho(\beta, \gamma) = \inf \sup_{t\in[0,1]} |\hat{\beta}(t) - \hat{\gamma}(t)|$, where the infimum is over all choices of parameterizations $\hat{\beta}$ and $\hat{\gamma}$ in $[0, 1]$ of β and γ.

THEOREM 1.1 (LERW scaling limit). *The measures* μ_δ *converge weakly to* ν *as* $\delta \to 0$ *with respect to the metric* ρ *on the space of curves.*

Since SLE is conformally invariant by definition, this theorem implies conformal invariance of the LERW. The theorem and proof apply also to some other walks on lattices in the plane where the scaling limit of the walk is isotropic Brownian motion. It even applies in the nonreversible setting. See Section 6 for further details.

There are two distinct definitions for the UST corresponding to a domain $D \subsetneqq \mathbb{C}$, as follows. Let $G_F(D)$ denote the subgraph of $\mathbb{Z}^2$ consisting of all the edges and vertices which are contained in D. If $G_F(D)$ is connected, then we refer to the UST on $G_F(D)$ as the UST on D with *free* boundary conditions. Let $G_W(D)$ denote the graph obtained from $\mathbb{Z}^2$ by contracting all the vertices outside of D to a single vertex (and removing edges which become loops). Then the UST on $G_W(D)$ is the UST on D with *wired* boundary conditions.

Since the UST is built from the LERW via Wilson's algorithm, it is not surprising that conformal invariance of the UST scaling limit should follow from that of the LERW scaling limit. In fact, [38], Theorem 11.3, says just that.

COROLLARY 1.2 (UST scaling limit). *The wired and free UST scaling limits (as defined in* [38]*) in a simply connected domain* $D \subset \mathbb{C}$ *whose boundary is a* C^1*-smooth simple closed curve exist and are conformally invariant.*

One can easily show, using [38], Theorem 11.1(i), that the wired tree depends continuously on the domain, and hence for that case D may be an arbitrary simply connected domain. However, some regularity assumption is needed for the free UST scaling limit: conformal invariance fails for the domain whose boundary contains the topologist's sine curve (the closure of $\{x + i\sin(1/x) : x \in (0, 1]\}$).

The UST Peano curve is an entirely different curve derived from the UST in two dimensions. The curve is rather remarkable, as it is a natural random path visiting every vertex in an appropriate graph or lattice. We now roughly describe two natural definitions of this curve; further details appear in Section 4.

Let G be a finite planar graph, with a particular embedding in the plane, and let $G^\dagger$ denote its planar dual, again with a particular embedding. Then there is a bijection $e \leftrightarrow e^\dagger$ between the edges of G and those of $G^\dagger$ such that, for every edge e in G, $e \cap e^\dagger$ is a single point, and e does not intersect any other edge of $G^\dagger$. Given a spanning tree T of G, let $T^\dagger$ denote the graph whose vertices are the vertices of $G^\dagger$ and whose edges are those edges $e^\dagger$ such that $e \notin T$. It is then easy to verify that $T^\dagger$ is a spanning tree for $G^\dagger$. Therefore, if T is a UST on G, then $T^\dagger$ is a UST on $G^\dagger$.

The UST Peano curve is a curve that winds between T and $T^\dagger$ and separates them. More precisely, consider the graph $\hat{G}$ drawn in the plane by taking the union of G and $G^\dagger$, where each edge e or $e^\dagger$ is subdivided into two edges by introducing a vertex at $e \cap e^\dagger$. The subgraph of the planar dual $\hat{G}^\dagger$ of $\hat{G}$ containing all edges which do not intersect $T \cup T^\dagger$ is a simple closed path—the *UST Peano path*. See Figure 3.

Some properties of the UST Peano path on $\mathbb{Z}^2$ have been studied in the physics literature; see, for example, [10] and [18]. There, it has been called the Hamiltonian path on the Manhattan lattice. The reason for this name is as follows. On $\mathbb{Z}^2$, say, orient each horizontal edge whose y-coordinate is even to the right and each horizontal edge whose y-coordinate is odd to the left. Similarly, orient down each vertical edge whose x-coordinate is even and orient up each vertical edge whose

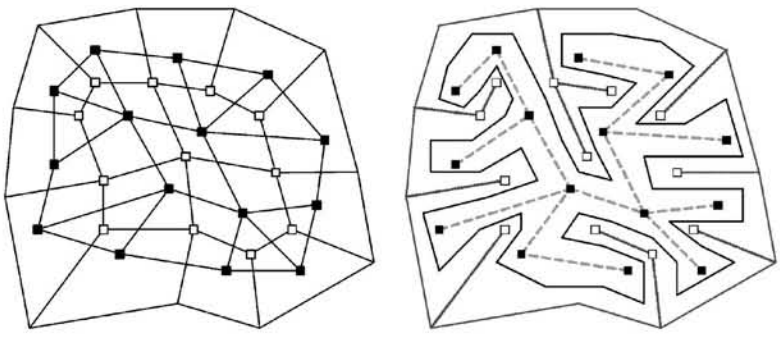


FIG. 3. *The graph, dual graph, tree, dual tree and Peano curve. The vertex of the dual graph corresponding to the unbounded face is drawn as a cycle.*

x-coordinate is odd. Now rescale the resulting oriented graph by $1/2$ and translate it by $(1/4, 1/4)$. It is easy to check that a Hamiltonian path (a path visiting every vertex exactly once) respecting the orientation on the resulting oriented graph is the same as the UST Peano path of $\mathbb{Z}^2$. It should be expected that the uniform measure on Hamiltonian paths in $\mathbb{Z}^2$ has the same scaling limit as that of the UST Peano path.

Given a domain D, one can consider the UST Peano curve for the wired or for the free UST (which is essentially the same as the wired, by duality). However, the conjecture from [38] regarding the convergence to chordal SLE pertains to the UST Peano curve associated with the tree with mixed wired and free conditions.

Let $D \subset \mathbb{C}$ be a domain whose boundary is a C^1-smooth simple closed curve and let $a, b \in \partial D$ be distinct boundary points. Let α and β denote the two complementary arcs of ∂D whose endpoints are a and b. For all $\delta > 0$, consider an approximation G_δ of the domain D in the grid $\delta\mathbb{Z}^2$. (A precise statement of what it means for G_δ to be an approximation of D will be given in Section 4.) Let γ_δ denote the Peano curve associated to the UST on G_δ with wired boundary near α and free boundary near β. Then γ_δ may be considered as a path in D from a point near a to a point near b.

THEOREM 1.3 (UST Peano path scaling limit). *The UST Peano curve scaling limit in D with wired boundary on α and free boundary on β exists and is equal to the image of the chordal* SLE_8 *path under any conformal map from $\mathbb{H}$ to D mapping* 0 *to a and ∞ to b.*

Again, the convergence is weak convergence of measures with respect to the metric ρ. Figure 4 shows a sample of the UST Peano path on a fine grid.

As explained above, it was proved in [36] that each SLE_κ is generated by a path, except for $\kappa = 8$. In Section 4.4, the remaining case $\kappa = 8$ is proved, using the convergence of the Peano curve.

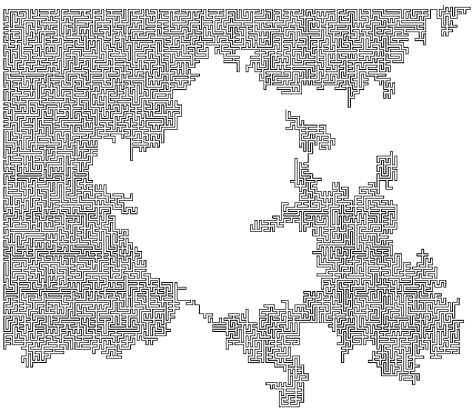


FIG. 4. *An arc from a sample of the UST Peano path; proved to converge to chordal* SLE_8.

Corollary 1.2 and Theorem 1.3 (and their proofs) apply to other reversible walks on planar lattices (the self-duality of $\mathbb{Z}^2$ does not play an important role); see Section 6.

To add perspective, we note that the convergence to SLE of the LERW and the UST Peano curve are two boundary cases of the conjectured convergence in [36] of the critical FK random cluster measures with parameter $q \in (0, 4)$. For these parameter values, the scaling limit of the interface of a critical cluster with mixed boundary values is conjectured to converge to chordal $\mathrm{SLE}_\kappa(q)$, where $\kappa(q) = 4\pi/\cos^{-1}(-\sqrt{q}/2)$. The boundary case $\kappa(0) = 8$ corresponds to the convergence of the UST Peano path to SLE_8.

The outer boundary of the scaling limit of a macroscopic critical cluster is not the same as the scaling limit of a critical cluster outer boundary, because of "fjords" which are pinched off in the limit. The former is conjectured to "look like" $\mathrm{SLE}_{16/\kappa(q)}$, but a precise form of this conjecture is not yet known. In the case $q = 0$, however, such a correspondence is easy to explain. In $\mathbb{Z}^2$, an arc of the Peano curve is surrounded on one side by a simple path in the tree, and on the other side by a simple path in the dual tree. Both these paths are LERWs. Similar correspondences exist for the UST in a subdomain of $\mathbb{R}^2$, but one has to set appropriate boundary conditions. Thus, the convergence of LERW to SLE_2 also corresponds to the case $q = 0$, as $16/\kappa(0) = 2$.

Suppose that $0 \in D$ and $\alpha, \beta \subset \partial D$, as before. Consider the simple random walk on $\delta\mathbb{Z}^2$ which is reflected off β and stopped when it hits α. Using an analogous method to that of the present paper, one could handle the scaling limit of the loop erasure of this walk. It is described by a variant of SLE_2 where the driving term is Brownian motion with time scaled by 2, but having an additional drift. The drift is not constant, but can be explicitly computed.

The identification of the scaling limit as one of the SLEs should facilitate the derivation of critical exponents and also the asymptotic probabilities of various events, including some results which have not been predicted by arguments from physics. This was the case for critical site percolation on the triangular grid; see [30], [39], [40] and [42].

1.6. *Some comments about the proof.* Since a loop-erased random walk is obtained in a deterministic way from a simple random walk (by erasing its loops) and since a simple random walk converges to Brownian motion in the scaling limit, it is natural to think that the scaling limit of the LERW should simply be the process obtained by erasing the loops from a planar Brownian motion. The problem with this approach is that planar Brownian motion has loops at every scale, so that there is no simple algorithm to erase loops. In particular, there is no "first" loop. Our proof does use the relation between the LERW and simple random walks, combined with the fact that quantities related to simple random walks, such as hitting probabilities, converge to their continuous conformally invariant counterparts.

The proof of each of our main theorems is naturally divided into two parts. The first part establishes the convergence to SLE with respect to a weaker topology than the topology induced by the metric ρ of paths; namely, we show that the Loewner driving process for the discrete random path converges to a Brownian motion. This part of the proof, which we consider to be the more important one, is essentially self-contained. The second part uses some regularity properties of the discrete processes from [38] to prove convergence with respect to the stronger topology.

The method for the first part can be considered as a rather general method for identifying the scaling limit of a dependent system that is conjectured to be conformally invariant. It requires having some "observable" quantity that can be estimated well and a mild Markovian property, which we now describe. Suppose that to every simply connected domain D containing 0 there is associated a random path γ from ∂D to 0 (e.g., the orientation reversal of the LERW). The required property is that if β is an arc with one endpoint in ∂D and we condition on $\beta \subset \gamma$ (assuming this has positive probability, say), then the conditioned distribution of $\gamma \setminus \beta$ is the same as the random path in the domain $D \setminus \beta$ conditioned to start at the other endpoint q of β. [Thus, $(D \setminus \beta, q)$ is the state of a Markov chain whose transitions correspond to adding edges from γ to β and modifying q appropriately.] Interestingly, among the discrete processes conjectured to converge to SLE, the LERW is the only one where the verification of this property is not completely trivial. [For the LERW it is not trivial, but not difficult; see part (iii) of Lemma 3.2.] The statement of this property for the UST Peano curve is given in Lemma 4.1. The fact that SLE satisfies this property follows from the Markovian property of its driving Brownian motion.

The particular choice of observable is not so important. What is essential is that one can conveniently calculate the asymptotics of the observable for appropriate large-scale configurations. The particular observable that we have chosen for the LERW convergence is the expected number of visits to a vertex v by the simple random walk generating the LERW. Conformal invariance is not assumed but comes out of the calculation—hitting probabilities for random walks are discrete harmonic functions, which converge to continuous harmonic functions. One technical issue is to establish this convergence without any boundary smoothness assumption. Once the observable has been approximated, the conditional expectation and variance of increments of the Loewner driving function for the discrete process can be estimated, and standard techniques (the Skorohod embedding) can be used to show that this random function approaches the appropriate Brownian motion.

Although Theorem 1.3 can probably be derived with some work from Corollary 1.2, instead, to illustrate our method, we prove it by applying again the same general strategy of the proof of Theorem 1.1, with the choice of a different observable.

Actually, it is easier to explain the main ideas behind the proof of Theorem 1.3. Fix some vertex v in D and a subarc $\alpha_1 \subset \alpha$. Let $\mathcal{A}$ be the event that the UST path (not the Peano path, but the path contained in the UST) from v to α hits α_1. By Wilson's algorithm the probability of $\mathcal{A}$ is the same as the probability that a simple random walk started at v reflected off β first hits α in α_1. The latter probability can be estimated directly. If $\gamma[0,n]$ denotes the restriction of the Peano path to its first n steps, then $\mathbf{P}[\mathcal{A}|\gamma[0,n]]$, the probability of $\mathcal{A}$ conditioned on $\gamma[0,n]$, is clearly a martingale with respect to n. But, by the Markovian property discussed above, the value of $\mathbf{P}[\mathcal{A}|\gamma[0,n]]$ may be estimated in precisely the same way that $\mathbf{P}[\mathcal{A}]$ is estimated. The estimate turns out to be a function of the conformal geometry of the configuration $(v, D \setminus \gamma[0,n], \gamma(n), \alpha_1, \beta)$. Knowing that this is a martingale for two appropriately chosen vertices v is sufficient to characterize the large-scale behavior of γ.

As mentioned above, in the case of the LERW, the observable we chose to look at is the expected number of visits to a fixed vertex v by the simple random walk Γ generating the LERW γ. The walk Γ can be considered as the union of γ with a sequence of loops Γ^j based at vertices of γ. We look at the conditioned expectation of the number of visits of Γ to v given an arc $\tilde{\gamma}$ of γ adjacent to the boundary of the domain. This is clearly a martingale with respect to the filtration obtained by taking larger and larger arcs $\tilde{\gamma} \subset \gamma$. This quantity falls into two parts: the visits to v in the loops Γ^j based at $\tilde{\gamma}$, and those that are not. Each of these two parts can be well estimated by random walk calculations. Translating the fact that this is a martingale to information about the Loewner driving process for γ inevitably leads to the identification of this driving process as appropriately scaled Brownian motion.

Actually, we first had a longer proof of convergence of the LERW to the SLE_2, based on the fact that it is possible to construct the hull of a Brownian motion by adding Brownian loops to SLE_2. This can be viewed as a particular case of the restriction properties of SLE_κ with Brownian loops added, which we study in [31]. Let us also mention the following related open question. Consider a sequence of simple random walks $S^k(n)$ on a lattice with lattice spacing $\delta_k \to 0$, from $S^k(0) = 0$ to $\partial\mathbb{U}$, and let γ^k denote the corresponding loop-erased paths. Theorem 1.1 shows that one can find a subsequence such that the law of the pair (γ^k, S^k) converges to a coupling of SLE_2 with Brownian motion [i.e., a law for a pair (X, Y), where X has the same distribution as the SLE_2 path and Y has the same distribution as Brownian motion]. The question is whether, in this coupling, SLE_2 is a deterministic function of the Brownian motion. In other words, is it possible to show that this is not a deterministic procedure to erase loops from a Brownian motion?

2. Preliminaries. The reading of this paper requires some background knowledge in several different fields. Some background about Loewner's equation and SLE is reviewed in the next section. It is assumed that the reader is familiar

with some of the basic properties of Brownian motion (definition, strong Markov property, etc.). Some of the basic properties of conformal maps (Riemann's mapping theorem, compactness, Koebe distortion) are also needed for the proof. This material may be learned from the first two chapters of [35], for example. In terms of the theory of conformal mappings, this suffices for understanding the argument showing that the driving process of the LERW converges to Brownian motion. For improving the topology of convergence, some familiarity with the notion of extremal length (a.k.a. extremal distance) is also required. A possible source for that is [2]. The reader also needs to know some of the very basic properties of harmonic measure.

2.1. *Loewner's equation and SLE.* We now review some facts concerning Loewner's equations and stochastic Loewner evolutions. For more details, see, for example, [27], [28], [36] and [38].

Suppose that $D \subsetneq \mathbb{C}$ is a simply connected domain with $0 \in D$. Then there is a unique conformal homeomorphism $\psi = \psi_D : D \to \mathbb{U}$ which is onto the unit disk $\mathbb{U} = \{z \in \mathbb{C} : |z| < 1\}$ such that $\psi_D(0) = 0$ and $\psi_D'(0)$ is a positive real. If $D \subset \mathbb{U}$, then $\psi_D'(0) \geq 1$, and $\log \psi_D'(0)$ is called the *capacity* of $\overline{\mathbb{U}} \setminus D$ from 0.

Now suppose that $\eta : [0, \infty] \to \overline{\mathbb{U}}$ is a continuous simple curve in the unit disk with $\eta(0) \in \partial\mathbb{U}$, $\eta(\infty) = 0$ and $\eta(0, \infty] \subset \mathbb{U}$. For each $t \geq 0$, set $K_t := \eta[0, t]$, $U_t := \mathbb{U} \setminus K_t$ and $g_t := \psi_{U_t}$. Since $t \mapsto g_t'(0)$ is increasing (by the Schwarz lemma, say), one can reparameterize the path in such a way that $g_t'(0) = \exp(t)$. If that is the case, we say that η is parameterized by *capacity* from 0. By standard properties of conformal maps (see [35], Proposition 2.5), for each $t \in [0, \infty)$ the limit

$$W(t) := \lim_{z \to \eta(t)} g_t(z),$$

where z tends to $\eta(t)$ from within $\mathbb{U} \setminus \eta[0, t]$, exists. One can also verify that

$$W : [0, \infty) \to \partial\mathbb{U}$$

is continuous. Assuming the parameterization by capacity, Loewner's theorem states that g_t satisfies the differential equation

$$\partial_t g_t(z) = -g_t(z) \frac{g_t(z) + W(t)}{g_t(z) - W(t)}. \tag{2.1}$$

It is also clear that

$$\forall z \in \mathbb{U}, \qquad g_0(z) = z. \tag{2.2}$$

We call $(W(t), t \geq 0)$ the *driving function* of the curve η.

The driving function W is sufficient to recover the two-dimensional path η, because the procedure may be reversed, as follows. Suppose that $W : [0, \infty) \to \partial\mathbb{U}$ is continuous. Then for every $z \in \overline{\mathbb{U}}$ there is a solution $g_t(z)$ of the ODE (2.1) with initial value $g_0(z) = z$ up to some time $\tau(z) \in (0, \infty]$, beyond which the solution

does not exist. In fact, if $\tau(z) < \infty$ and $z \neq W(0)$, then we have $\lim_{t \uparrow \tau(z)} g_t(z) - W(t) = 0$, since this is the only possible reason the ODE cannot be solved beyond time $\tau(z)$. Then one defines $K_t := \{z \in \overline{\mathbb{U}} : \tau(z) \leq t\}$ and $D_t := \mathbb{U} \setminus K_t$ is the domain of definition of g_t. The set K_t is called the *hull* at time t. If W arises from a simple path η as described in the previous section, then we can recover η from W by using $\eta(t) = g_t^{-1}(W(t))$. However, if $W : [0, \infty) \to \mathbb{U}$ is an arbitrary continuous driving function, then, in general, K_t need not be a path, and even if it is a path, it does not have to be a simple path.

Radial SLE_κ is the process $(K_t, t \geq 0)$, where the driving function $W(t)$ is set to be $W(t) := \exp(iB_{\kappa t})$, where $B : [0, \infty) \to \mathbb{R}$ is Brownian motion. Often, one takes the starting point B_0 to be random uniform in $[0, 2\pi]$. It has been shown in [36] that the hull K_t is a.s. a simple curve for every $t > 0$ if $\kappa \leq 4$ and that, a.s. for every $t > 0$, K_t is not a simple curve if $\kappa > 4$. For every $\kappa \geq 0$, there is a.s. some random continuous path $\eta : [0, \infty) \to \overline{\mathbb{U}}$ such that, for all $t > 0$, D_t is the component of $\mathbb{U} \setminus \eta[0, t]$ containing 0. When $\kappa \neq 8$, this was proved in [36], while for $\kappa = 8$ this will be proven in the current paper. This path is called the radial SLE path.

Suppose that D is a simply connected domain containing 0. If γ is a continuous simple curve joining ∂D to 0 with only an endpoint in ∂D, one can reparameterize the path $\eta := \psi \circ \gamma$ according to capacity and find its driving function W, as before. The conformal map

$$\hat{g}_t = \psi_{D \setminus \gamma[0,t]} : D \setminus \gamma[0, t] \to \mathbb{U}$$

still satisfies (2.1), but this time, $\hat{g}_0 = \psi_D$. (Here, the parameterization chosen for γ is according to the capacity of $\psi \circ \gamma[0, t]$.) Radial SLE in D is then simply the image under ψ_D^{-1} of radial SLE in the unit disk.

Similarly, one can encode continuous simple curves η from 0 to ∞ in the closed upper half plane $\overline{\mathbb{H}}$ via a variant of Loewner's equation. For each time $t \geq 0$, there is a unique conformal map g_t from $H_t := \mathbb{H} \setminus \eta[0, t]$ onto $\mathbb{H}$ satisfying the so-called *hydrodynamic normalization*

$$\lim_{z \to \infty} g_t(z) - z = 0, \tag{2.3}$$

where $z \to \infty$ in $\mathbb{H}$. If we write $g_t(z) = z + a(t)\, z^{-1} + o(z^{-1})$ near ∞, it turns out that $a(t)$ is monotone. Consequently, one can reparameterize η in such a way that $a(t) = 2t$, that is, $g_t(z) = z + 2tz^{-1} + o(z^{-1})$ when $z \to \infty$. This parameterization of η is called the parameterization by *capacity* from ∞. (This notion of capacity is analogous to the notion of capacity in the radial setting; however, these are two distinct notions and should not be confused.) If $g : \mathbb{H} \setminus K \to \mathbb{H}$ is the conformal homeomorphism satisfying the hydrodynamic normalization, then $\lim_{z \to \infty} (g_t(z) - z)\, z/2$ is called the *capacity* of K from ∞. Assuming that η is parameterized by capacity, the following analog of Loewner's

equation holds:

$$\forall t > 0,\ \forall z \in H_t, \qquad \partial_t g_t(z) = \frac{2}{g_t(z) - W(t)}, \tag{2.4}$$

where the driving function W is again defined by $W(t) := g_t(\eta(t))$. As above, η is determined by W.

Conversely, suppose that W is a real-valued continuous function. For $z \in \overline{\mathbb{H}}$, one can solve the differential equation (2.4) starting with $g_0(z) = z$, up to the first time $\tau(z)$ where $g_t(z)$ and $W(z)$ collide [possibly, $\tau(z) = \infty$]. Let the hull be defined by $K_t := \{z \in \overline{\mathbb{H}} : \tau(z) \le t\}$. Then $g_t : \mathbb{H} \setminus K_t \to \mathbb{H}$ is a conformal map onto $\mathbb{H}$, and $g_0(z) = z$. In general, K_t is not necessarily a simple curve. If $W(t) = B_{\kappa t}$, then $(K_t, t \ge 0)$ is called *chordal* SLE_κ.

It turns out (see [28], Section 4.1) that the local properties of chordal SLE_κ and of radial SLE_κ are essentially the same. [That is the reason the normalization $a(t) = 2t$ was chosen over the seemingly more natural $a(t) = t$.] In particular, for every κ, chordal SLE_κ is generated by a random continuous path, called the chordal SLE_κ path.

At some points in our proofs, we will need the following simple observation.

LEMMA 2.1 (Diameter bounds on K_t). *There is a constant $C > 0$ such that the following always holds. Let $W : [0, \infty) \to \mathbb{R}$ be continuous and let $(K_t, t \ge 0)$ be the corresponding hull for Loewner's chordal equation* (2.4) *with driving function W. Set*

$$k(t) := \sqrt{t} + \max\{|W(s) - W(0)| : s \in [0, t]\}.$$

Then

$$\forall t \ge 0, \qquad C^{-1} k(t) \le \operatorname{diam} K_t \le C\, k(t).$$

Similarly, when $K_t \subset \overline{\mathbb{U}}$ is the radial hull for a continuous driving function $W : [0, \infty) \to \partial\mathbb{U}$, then

$$\forall t \ge 0, \qquad C^{-1} \min\{k(t), 1\} \le \operatorname{diam} K_t \le C\, k(t).$$

PROOF. This lemma can be derived by various means. We will only give a detailed argument in the radial case. The chordal case is actually easier and can be derived using the same methods. It can also be seen as a consequence of the result in the radial setting (because chordal Loewner equations can be interpreted as scaling limits of radial Loewner equations).

We start by proving the upper bound on $\operatorname{diam} K_t$. Let $\delta \ge \max\{|W(s) - W(0)| : s \in [0, t]\}$. Then, as long as $|g_t(z) - W(0)| \ge 3\delta$, we have $|\partial_t g_t(z)| \le 1/\delta$. Hence, if $|z - W(0)| \ge 4\delta$, then, for all $t \le \delta^2$, $|g_t(z) - z| \le \delta$ and therefore $z \notin K_t$. Hence, $\operatorname{diam} K_t \le 8k(t)$.

In order to derive the lower bound, we will compare capacity with harmonic measure. It is sufficient to consider the case where $\operatorname{diam} K_t < 1/10$. Let μ denote the harmonic measure on $K_t \cup \partial\mathbb{U}$ from 0. Because K_t is contained in the disk of radius $\operatorname{diam} K_t$ with center $W(0) \in K_t \cap \partial\mathbb{U}$, there is a universal constant c such that $\mu(K_t) \le c \operatorname{diam} K_t$. Hence, it suffices to give a lower bound for $\mu(K_t)$.

Since $g_t(z)/z$ is analytic and nonzero in a neighborhood of 0, the function $h(z) = \log|g_t(z)| - \log|z|$ is harmonic in $U_t := \mathbb{U} \setminus K_t$. Note that $h(0) = t$. Because $|g_t(z)| \to 1$ as z tends to the boundary of U_t, the mean value property of $h \circ g_t^{-1}$ implies the following relation between harmonic measure and capacity: $t = h(0) = \int \log(1/|z|)\, d\mu(z)$. Since K_t contains points in $\partial\mathbb{U}$ and $\operatorname{diam} K_t \le 1/10$, we have $\log(1/|z|) \le c' \operatorname{diam} K_t$ for all $z \in K_t$. Therefore, $t \le c'\mu(K_t) \operatorname{diam} K_t \le c''(\operatorname{diam} K_t)^2$.

It now remains to compare $\mu(K_t)$ and $|W(t) - W(0)|$. We still assume that $\operatorname{diam} K_t < 1/10$. Let $A_t := \partial\mathbb{U} \setminus g_t(\overline{\mathbb{U}} \setminus K_t)$. If $z \in \partial\mathbb{U} \setminus K_s$ and $s \le t$, then (2.1) shows that $\partial_u |g_{s+u}(z) - W(s)| \ge 0$ at $u = 0$. This implies that $(A_s, s \le t)$ is nondecreasing. Hence, for all $s \le t$, we have $W(0) \in A_0 \subset A_t$ and $W(s) \in A_s \subset A_t$ so that $|W(s) - W(0)|$ is bounded by the length of A_t, which is equal to $2\pi\mu(K_t)$. This completes the proof of the lemma. □

2.2. *A discrete harmonic measure estimate.* In this section, we introduce some notation and state an estimate relating discrete harmonic measure and continuous harmonic measure in domains in the plane. In order to get more quickly to the core of our method in Section 3.2, we postpone the proof of the harmonic measure estimate to Section 5.

A *grid domain* D is a domain whose boundary consists of edges of the grid $\mathbb{Z}^2$. For an arbitrary domain $D \subset \mathbb{C}$, and $p \in D$ define the inner radius of D with respect to p,

$$\operatorname{rad}_p(D) := \inf\{|z - p| : z \notin D\}.$$

Let $\mathfrak{D}$ denote the set of all simply connected grid domains such that $0 < \operatorname{rad}_0(D) < \infty$ (i.e., $D \ne \mathbb{C}$ and $0 \in D$).

Points in $\mathbb{R}^2 = \mathbb{C}$ with integer coordinates will be called *vertices*, or *lattice points*. Let $V(D) := D \cap \mathbb{Z}^2$ denote the lattice points in D.

Let $D \in \mathfrak{D}$ and let v be a vertex in ∂D. If ∂D contains more than one edge incident with v, then it may happen that the intersection of D with a small disk centered at v will not be connected. Hence, as viewed from D, v appears as more than one vertex. In particular, $\psi = \psi_D$ does not extend continuously to v. This is a standard issue in conformal mapping theory, which is often resolved by introducing the notion of prime ends. But in the present case, there is a simpler solution which suffices for our purposes. Suppose $v \in \mathbb{Z}^2 \cap \partial D$ and e is an edge incident with v that intersects D. The set of such pairs $w = (v, e)$ will be denoted $V_\partial(D)$. If $\psi : D \to \mathbb{U}$ is conformal, then $\psi(w)$ will be shorthand for the

limit of $\psi(z)$ as $z \to v$ along e (which always exists, by [35], Proposition 2.14). Similarly, if a random walk first exits D at v, we say that it exited D at w if the edge e was used when first hitting v. A reader of this paper who chooses to be sloppy and not distinguish between v and w will not lose anything in the way of substance. We will not always be so careful to make this distinction.

If $a \in V(D)$ and $b \in V(D) \cup V_\partial(D)$, define $H(a,b) = H_D(a,b)$ as the probability that the simple random walk started from a and stopped at its first exit time of D visits b.

For any $w \in D$ and $u \in V_\partial(D)$, we define

$$\lambda = \lambda(w,u;D) := \frac{1-|\psi(w)|^2}{|\psi(w)-\psi(u)|^2} = \mathrm{Re}\left(\frac{\psi(u)+\psi(w)}{\psi(u)-\psi(w)}\right). \tag{2.5}$$

Note that λ is also equal to the imaginary part of the image of w by the conformal map from D onto the upper half plane that maps 0 onto i and u to ∞. It is also the limit when $\epsilon \to 0$ of the ratio between the harmonic measure in D of the ϵ neighborhood of u in ∂D, taken, respectively, at w and at 0 (i.e., it corresponds to the Poisson kernel). Therefore, λ can be viewed as the continuous analog of $H(w,u)/H(0,u)$. Note that the function $h(w) = H(w,u)/H(0,u)$ is *discrete harmonic* on $V(D)$, which means that $h(w)$ is equal to the average of h on the neighbors of w when $w \in V(D)$.

PROPOSITION 2.2 (Hitting probability). *For every $\epsilon > 0$ there is some $r_0 > 0$ such that the following holds. Let $D \in \mathfrak{D}$ satisfy $\mathrm{rad}_0(D) > r_0$ and let $u \in V_\partial(D)$ and $w \in V(D)$. Suppose $|\psi_D(w)| \le 1-\epsilon$ and $H(0,u) \neq 0$. Then*

$$\left|\frac{H(w,u)}{H(0,u)} - \lambda(w,u;D)\right| < \epsilon. \tag{2.6}$$

The proof is given in Section 5.

3. Conformal invariance of the LERW.

3.1. *Loop-erased random walk background.* We now recall some well-known facts concerning loop-erased random walks.

LEMMA 3.1 (LERW reversal). *Let $D \in \mathfrak{D}$ and let Γ be a simple random walk from 0 stopped when it hits ∂D. Let β be the loop erasure of Γ and let γ be the loop erasure of the time reversal of Γ. Then γ has the same distribution as the time reversal of β.*

See [25]. A simpler proof follows immediately from the symmetry of (12.2.3) in [26]. This result (and the proofs) also holds if we condition Γ to exit ∂D at a prescribed $u \in V_\partial(D)$, which corresponds to the event $\{\gamma \cap \partial D = \{u\}\} = \{\beta \cap \partial D = \{u\}\}$ (assuming this has positive probability).

Throughout our proof, we will use the simple random walk Γ and the loop erasure $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_\ell)$ of its time reversal (so that $\gamma_0 \in \partial D$ and $\gamma_\ell = 0$). We use D_j to denote the grid domains $D_j := D \setminus \bigcup_{i=0}^{j-1} [\gamma_i, \gamma_{i+1}]$. Define, for $j \in \{0, 1, \ldots, \ell\}$,

$$n_j := \min\{n \geq 0 : \Gamma(n) = \gamma_j\}$$

and note that $n_{j+1} < n_j$ for $j = 0, 1, \ldots, \ell - 1$ by the definition of γ. Also set

$$\Gamma^{j+1} := \Gamma[n_{j+1}, n_j].$$

More precisely, consider Γ^j as the grid path given by

$$\Gamma^j(m) := \Gamma(m + n_j), \qquad m = 0, 1, \ldots, n_{j-1} - n_j.$$

LEMMA 3.2 (Markovian property). *Let $j \in \mathbb{N}$ and let $u_0, \ldots, u_j \in \mathbb{Z}^2$. Suppose that the probability of the event $(\gamma_0, \ldots, \gamma_j) = (u_0, \ldots, u_j)$ is positive. Conditioned on this event, the following hold*:

(i) *The paths $\Gamma^1, \ldots, \Gamma^j$ and $\Gamma[0, n_j]$ are conditionally independent.*

(ii) *For $k \in \{1, \ldots, j\}$, the conditional law of Γ^k is that of a simple random walk in D_{k-1} started from u_k and conditioned to leave D_{k-1} through the edge $[u_k, u_{k-1}]$.*

(iii) *The conditional law of $\Gamma[0, n_j]$ is that of a simple random walk started from 0 conditioned to leave D_j at u_j, and $\gamma[j, \ell]$ is the loop erasure of the time reversal of $\Gamma[0, n_j]$.*

PROOF. Since γ is the loop erasure of the reversal of Γ, the event $(\gamma_0, \ldots, \gamma_j) = (u_0, \ldots, u_j)$ is equivalent to the statement that, for each $k = 0, 1, \ldots, j-1$, the first hit of Γ to $\{u_0, \ldots, u_k\} \cup \partial D$ is through the edge $[u_{k+1}, u_k]$. Let $\tau_k := \min\{n : \Gamma(n) \in \{u_0, \ldots, u_k\} \cup \partial D\}$, $k = 0, \ldots, j$. The strong Markov property of Γ with the stopping times τ_k now implies the lemma. $\square$

The following simple lemma will also be needed.

LEMMA 3.3 (Expected visits). *Suppose that $v \in V(D)$ and that u_0 and u_1 are two vertices satisfying $\mathbf{P}[\gamma_0 = u_0, \gamma_1 = u_1] > 0$. Conditioned on $\gamma_0 = u_0$ and $\gamma_1 = u_1$, the expected number of visits to v by Γ^1 is $G(u_1, v)H(v, u_1)$.*

Here $G(u, v)$ denotes the discrete Green's function, that is, the expected number of visits to v by a simple random walk started at u, which is stopped on exiting D.

PROOF OF LEMMA 3.3. Let X be a simple random walk from u_1 stopped on exiting D and let k be the last time such that $X(k) = u_1$. Then Γ^1 conditioned on $\gamma_0 = u_0$ and $\gamma_1 = u_1$ has the same distribution as X conditioned on $X(k+1) = u_0$. But the path $j \mapsto X(k + j)$ is independent from $X[0, k]$. Consequently, the expected number of visits of X to v conditioned on $X(k + 1) = u_0$ is equal to the expected number of visits to v of $X[0, k]$. The lemma follows. $\square$

3.2. *The core argument.* We keep the previous notation and also use the conformal maps $\psi_j : D_j \to \mathbb{U}$ satisfying $\psi_j(0) = 0$ and $\psi_j'(0) > 0$. Set $U_j := \psi_j(\gamma_j)$ and $U := U_0$. Note that γ can also be viewed as a continuously growing simple curve from ∂D to 0 and therefore can be represented by Loewner's equation. Let $W : [0, \infty) \to \partial\mathbb{U}$ denote the (unique) continuous function such that solving the radial Loewner equation with driving function $W(t)$ gives the path γ. Note that $U_j = W(t_j)$, where t_j is the continuous capacity of $\gamma[0, j]$ from 0 in D (i.e., the capacity of $\psi(\gamma[0, j])$ from 0 in $\mathbb{U}$). We denote by $(\vartheta(t), t \geq 0)$ the continuous real-valued function with $\vartheta(0) = 0$ such that $W(t) = W(0)\exp(i\vartheta(t))$. We also define $\Delta_j = \vartheta(t_j)$, so that $U_j = U\exp(i\Delta_j)$.

PROPOSITION 3.4 (The key estimate). *There exists a positive constant C such that, for all small positive δ, there exists $r_0 = r_0(\delta)$ such that the following holds. Let $D \in \mathfrak{D}$ satisfy* $\mathrm{rad}_0(D) > r_0$. *For every $u_0 \in V_\partial(D)$ with $\mathbf{P}[\gamma_0 = u_0] > 0$, let γ denote the random path from u_0 to 0 obtained by loop erasure of the time reversal of a simple random walk from* 0 *to ∂D conditioned to hit ∂D in u_0. Let*

$$m := \min\{j \geq 1 : t_j \geq \delta^2 \text{ or } |\Delta_j| \geq \delta\},$$

where Δ_j and t_j are as described above. Then

$$|\mathbf{E}[\Delta_m]| \leq C\delta^3 \tag{3.1}$$

and

$$|\mathbf{E}[\Delta_m^2] - 2\mathbf{E}[t_m]| \leq C\delta^3. \tag{3.2}$$

Recall that Lemma 3.1 says that γ has the same distribution as the *chronological* loop erasure of a random walk from 0 to ∂D conditioned to hit ∂D at u_0.

Here is a rough sketch of the proof. Let $v \in V(D)$ satisfy

$$\mathrm{rad}_0(D)/200 < |v| < \mathrm{rad}_0(D)/5. \tag{3.3}$$

Let h_0^+ denote the number of visits to v by Γ. (This is the quantity which we referred to in Section 1 as the "observable.") The proof is based on estimating the two sides of the equality

$$\mathbf{E}[h_0^+] = \mathbf{E}[\mathbf{E}[h_0^+ | \gamma[0, m]]]. \tag{3.4}$$

The estimate for the right-hand side will involve the distribution of t_m and Δ_m. We get the two relations (3.1) and (3.2) by considering two different choices for such a v.

The estimates for the two sides of (3.4) are rather straightforward. Basically, each side is translated into expressions involving the Green's functions G_j and the hitting probabilities H_j. These are then translated into analytic quantities using (2.6). Earlier versions of the proof required other estimates, somewhat more delicate, in addition to (2.6). Fortunately, it turned out that (2.6) is sufficient. Since

we came across several different variants for the proof, based on choosing different observables, it may be said that the proof is inevitable, rather than accidental (and this also applies to Theorem 1.3). Basically, the reason the proof works is that the expected number of visits to v in $\bigcup_{j=1}^{m} \Gamma^j$ given $\gamma[0,m]$ can be estimated rather well given the rough geometry of $\gamma[0,m]$ in a scale much coarser than the scale of the grid. Similarly, it is important that $\mathbf{E}[h_0^+]$ can be estimated given the rough geometry of D, but this fact is not surprising.

In the following, we abbreviate the Green's functions and hitting probabilities in D_j by $G_j := G_{D_j}$ and $H_j := H_{D_j}$. The following lemma will be needed.

LEMMA 3.5 (Green's function bounds). *There is a constant $C > 0$ such that, for every $D \in \mathfrak{D}$ and $v \in V(D)$ satisfying* (3.3),

$$1/C \le G_D(0,v) \le C \tag{3.5}$$

holds. Also, given $\delta > 0$, there is an $r = r(\delta)$ such that, if $\mathrm{rad}_0(D) > r$, *then, with the notation of Proposition* 3.4,

$$G_0(0,v) - G_m(0,v) \le C\delta^2. \tag{3.6}$$

PROOF OF PROPOSITION 3.4. Since $t_{m-1} < \delta^2$, it follows from the Koebe $1/4$ theorem that $\mathrm{rad}_0(D_{m-1}) > \mathrm{rad}_0(D)/5 \ge r_0/5$ if δ is small. (Apply [35], Corollary 1.4, with $z = 0$ to ψ_0^{-1} and ψ_{m-1}^{-1}.) Moreover, the continuous harmonic measure in D at 0 of any edge e with a vertex on ∂D can be made arbitrarily small by requiring $\mathrm{rad}_0(D)$ to be large. [A Brownian motion started at 0 has probability going to 1 to surround the disk $\mathrm{rad}_0(D)\mathbb{U}$ before hitting e, as $\mathrm{rad}_0(D) \to \infty$.] By the conformal invariance of the harmonic measure, this implies that the diameter of $\psi(e)$ can be made arbitrarily small. Applying this to the domains D_j and using Lemma 2.1, we see that we may take r_0 large enough so that, for all $j < m$, for all $t \in [t_j, t_{j+1}]$, $|\vartheta(t) - \vartheta(t_j)| \le \delta^3$ and $|t_{j+1} - t_j| \le \delta^3$. In particular, $t_m \le \delta^2 + \delta^3$ and $|\Delta_m| \le \delta + \delta^3$. We also require $r_0/8$ to be larger than the $r(\delta)$ of Lemma 3.5.

Suppose $v \in V(D)$ satisfies (3.3). Set $Z_j := \psi_j(v)$ and $Z := Z_0$. For each $j \in \{1, \dots, \ell\}$, let h_j denote the number of visits to v by Γ^j. Also let

$$h_j^+ := \sum_{k=j+1}^{\ell} h_k,$$

which is the number of visits of v by $\Gamma[0, n_j]$. Let $\lambda_j := \lambda(v, \gamma_j; D_j)$, where $\lambda(v, v'; D_j)$ is as in (2.5). Since, conditionally on $\gamma[0,j]$, $\Gamma[0,n_j]$ is a random walk in D_j conditioned to leave D_j at γ_j,

$$\mathbf{E}\big[h_j^+ | \gamma[0,j]\big] = \frac{G_j(0,v) H_j(v, \gamma_j)}{H_j(0, \gamma_j)}$$

and Proposition 2.2 [together with (3.5)] implies that if $r_0(\delta)$ is sufficiently large, then, for every $j \in \{0, 1, \ldots, m\}$,

$$\mathbf{E}\big[h_j^+ | \gamma[0, j]\big] = G_j(0, v)\lambda_j + O(\delta^3).$$

[This $O(\cdot)$ notation is shorthand for the statement that there is an absolute constant C such that $|\mathbf{E}[h_j^+ | \gamma[0, j]] - G_j(0, v)\lambda_j| \le C\delta^3$. We freely use this shorthand below.] In particular,

$$\mathbf{E}\left[\sum_{j=1}^{m} h_j\right] = \mathbf{E}[h_0^+ - h_m^+] = \mathbf{E}[G_0(0, v)\lambda_0 - G_m(0, v)\lambda_m] + O(\delta^3). \tag{3.7}$$

We will now get a different approximation for the left-hand side. Applying Lemma 3.3 to the domain D_{j-1} gives

$$\mathbf{E}\big[h_j | \gamma[0, j]\big] = G_{j-1}(\gamma_j, v)H_{j-1}(v, \gamma_j).$$

Proposition 2.2 implies that, for $r_0(\delta)$ large enough,

$$\mathbf{E}\big[h_j | \gamma[0, j]\big] = \big(\lambda_{j-1} + O(\delta)\big)G_{j-1}(\gamma_j, v)H_{j-1}(0, \gamma_j). \tag{3.8}$$

Considering the same simple random walk starting at 0 and stopped when it exits D_j or D_{j-1} shows that

$$G_{j-1}(0, v) - G_j(0, v) = H_{j-1}(0, \gamma_j)G_{j-1}(\gamma_j, v). \tag{3.9}$$

We now derive an a priori bound on $\max\{|\lambda_j - \lambda_m| : j \le m\}$. Recall that

$$\lambda_j - \lambda_0 = \operatorname{Re}\left(\frac{U_j + Z_j}{U_j - Z_j} - \frac{U + Z}{U - Z}\right). \tag{3.10}$$

But $|U_j - U| \le O(\delta)$ for $j \le m$, and Loewner's equation shows that

$$\begin{aligned} \forall\, j \le m, \qquad Z_j &= Z + t_j Z\frac{U + Z}{U - Z} + t_j O(\delta) \\ &= Z + t_j Z\frac{U + Z}{U - Z} + O(\delta^3), \end{aligned} \tag{3.11}$$

and, in particular, $Z_j = Z + O(\delta^2)$. (The equation blows up when $|U - Z|$ is small, and such estimates would not be valid in such a situation. However, this is not a problem here. First, $\psi_0'(0) \le 1/\operatorname{rad}_0(D)$ by the Schwarz lemma applied to the restriction of ψ_0 to $\operatorname{rad}_0(D)\mathbb{U}$. Now the Koebe 1/4 theorem (the case $z = 0$ in the left-hand inequality in [35], Corollary 1.4) gives $\psi_0^{-1}((4/5)\mathbb{U}) \supset (1/4)|\psi_0'(0)|^{-1}(4/5)\mathbb{U} \supset (\operatorname{rad}_0(D)/5)\mathbb{U}$. In particular, $|Z| = |\psi_0(v)| \le 4/5$ by (3.3). Since $t_m = O(\delta^2)$, it is clear that if δ is small and one starts flowing from Z according to Loewner's equation, it is impossible for Z to get close to $\partial\mathbb{U}$ up to time t_m.) Thus, we get our bound,

$$\forall\, j \le m, \qquad |\lambda_j - \lambda_m| \le O(\delta).$$

Using (3.8), this implies

$$\mathbf{E}[h_j|\gamma[0,j]] = (\lambda_m + O(\delta))G_{j-1}(\gamma_j, v)H_{j-1}(0,\gamma_j).$$

Now applying (3.9) yields

$$\mathbf{E}\left[\sum_{j=1}^{m} h_j\right] = \mathbf{E}\big[(\lambda_m + O(\delta))\big(G_0(0,v) - G_m(0,v)\big)\big],$$

and hence (3.6) implies

$$\mathbf{E}\left[\sum_{j=1}^{m} h_j\right] = \mathbf{E}\big[\lambda_m\big(G_0(0,v) - G_m(0,v)\big)\big] + O(\delta^3).$$

Comparing with (3.7) gives $G_0(0,v)\mathbf{E}[\lambda_m - \lambda_0] = O(\delta^3)$, and hence (3.5) implies

$$\mathbf{E}[\lambda_m - \lambda_0] = O(\delta^3). \tag{3.12}$$

[The reader may wonder about the apparent miracle happening here, namely, that λ_j turns out to be "almost" a martingale. In fact, this is not important for identifying the scaling limit. If the right-hand side in (3.12) turned out to be any other explicit quantity, up to δ^3 error terms, the proof would still work, but give a different limiting process. In Remark 3.6 below, we give a short proof of (3.12) and further comments.]

Recall that this equation is valid uniformly over all choices of v. We now Taylor-expand $\lambda_m - \lambda_0$ with respect to $U_m - U$ and $Z_m - Z$, up to $O(\delta^3)$ error terms. As we have seen, $U_m - U = O(\delta)$ and $Z_m - Z = O(\delta^2)$, and hence only the first-order derivative with respect to $Z_m - Z$ and the first two derivatives with respect to

$$U_m - U = (e^{i\Delta_m} - 1)U = iU\Delta_m - U\Delta_m^2/2 + O(\delta^3)$$

come into play (the mixed derivatives can be ignored). Using (3.10) and (3.11), we get

$$\lambda_m - \lambda_0 = \Delta_m \operatorname{Im}\left(\frac{2ZU}{(U-Z)^2}\right) + (2t_m - \Delta_m^2)\operatorname{Re}\left(\frac{ZU(U+Z)}{(U-Z)^3}\right) + O(\delta^3),$$

and therefore (3.12) gives

$$\operatorname{Im}\left(\frac{2ZU}{(U-Z)^2}\right)\mathbf{E}[\Delta_m] + \operatorname{Re}\left(\frac{ZU(U+Z)}{(U-Z)^3}\right)\mathbf{E}[2t_m - \Delta_m^2] = O(\delta^3). \tag{3.13}$$

We claim that when $r_0(\delta)$ is large enough we may find $v_1, v_2 \in V(D)$ in the range (3.3) satisfying $|\psi(v_1) - U/30| < \delta^3$ and $|\psi(v_2) - iU/30| < \delta^3$. Indeed, by Theorem 1.3 and Corollary 1.4 from [35], for every $R \in (0,1)$, there is a $C = C(R) < \infty$ such that $|\psi'(z)| \le C/\operatorname{rad}_z(D)$ and $\operatorname{rad}_z(D) \ge C^{-1}\operatorname{rad}_0(D)$ hold for all $z \in \psi^{-1}(R\mathbb{U})$. Let v_1 be a vertex closest to $\psi^{-1}(U/30)$. By integrating

the above bound on ψ' along the line segment from $\psi^{-1}(U/30)$ to v_1 (whose length is less than 1), we get $|\psi(v_1) - U/30| < \delta^3$ if $\mathrm{rad}_0(D)$ is large enough. Another application of Theorem 1.3 and Corollary 1.4 from [35] now shows that v_1 satisfies (3.3). An entirely similar argument produces v_2.

Consequently, (3.13) holds with $Z \in \{U/30, iU/30\}$. Plugging in these two values for Z produces two linearly independent equations in the variables $\mathbf{E}[2t_m - \Delta_m^2]$ and $\mathbf{E}[\Delta_m]$ and thereby proves (3.1) and (3.2). □

REMARK 3.6. Here is another proof of (3.12). Given a vertex $v \in V(D)$, let $\beta = (\beta_0, \beta_1, \dots)$ denote the loop erasure of the reversal of the simple random walk Γ_v started from v and stopped on exiting D (i.e., the analog of γ, but starting from v instead of 0). Abbreviate $\gamma^n := (\gamma_0, \dots, \gamma_n)$ and similarly $\beta^n := (\beta_0, \dots, \beta_n)$. For a sequence of vertices $u = (u_0, u_1, \dots, u_n)$, let $a_n(u) := \mathbf{P}[\gamma^n = u]$ and $b_n(u) := \mathbf{P}[\beta^n = u]$. Set $M_n := b_n(\gamma^n)/a_n(\gamma^n)$. (In other words, M_n is the Radon–Nikodym derivative of the law of β^n with respect to the law of γ^n.) It is easy to verify that M_n is a martingale:

$$\mathbf{E}[M_{n+1}|\gamma^n] = \sum_w \frac{b_{n+1}(\gamma^n w)}{a_{n+1}(\gamma^n w)} \frac{a_{n+1}(\gamma^n w)}{a_n(\gamma^n)} = \frac{\sum_w b_{n+1}(\gamma^n w)}{a_n(\gamma^n)} = M_n.$$

Lemma 3.2 implies that $M_n = H_n(v, \gamma_n)/H_n(0, \gamma_n)$, since, on the event that Γ_v and Γ first hit $\{u_0, \dots, u_n\} \cup \partial D$ at u_n, we may couple them to agree after that first visit to u_n. Now (2.6) implies (3.12).

Although this proof is shorter than the first proof of (3.12), it is harder to motivate and less natural. For this reason, we chose to stress the first proof.

Let us finally note that (as opposed to the martingale that shows up in the analysis of the UST Peano curve), the quantity corresponding to this martingale in the scaling limit is unbounded and converges almost surely to 0 (it is not uniformly integrable), so that it cannot be interpreted as a conditional probability or a conditional expectation. Correspondingly, in the discrete setting, M_n is very large when the path hits v (if it does) and M_n is very small when the path hits 0.

3.3. *Recognizing the driving process.* The objective in this section is to show that W of the previous section is close to a time-scaled Brownian motion on the unit circle.

THEOREM 3.7 (Driving process convergence). *For every $T > 0$ and $\epsilon > 0$, there is an $r_1 = r_1(\epsilon, T) > 0$ such that, for all $D \in \mathfrak{D}$ with $\mathrm{rad}_0(D) > r_1$, there is a coupling of γ with Brownian motion $B(t)$ starting at a random uniform point in $[0, 2\pi]$ such that*

$$\mathbf{P}\big[\sup\{|\vartheta(t) - B(2t)| : t \in [0, T]\} > \epsilon\big] < \epsilon.$$

Recall that a coupling of two random variables (or random processes) A and B is a probability space with two random variables A' and B', where A' has the same distribution as A and B' has the same distribution as B. In the above statement (as is customary), we do not distinguish between A and A' and between B and B'.

In order to deduce this theorem from Proposition 3.4, we will use the Skorohod embedding theorem, which is one of the standard tools for proving convergence to Brownian motion (one could work out a more direct proof but the following proof seems cleaner).

LEMMA 3.8 (Skorohod embedding). *If $(M_n)_{n\le N}$ is an $(\mathcal{F}_n)_{n\le N}$ martingale, with $\|M_n - M_{n-1}\|_\infty \le 2\delta$ and $M_0 = 0$ a.s., then there are stopping times $0 = \tau_0 \le \tau_1 \le \cdots \le \tau_N$ for standard Brownian motion $(B_t, t \ge 0)$ such that $(M_0, M_1, \ldots, M_N)$ and $(B_{\tau_0}, B_{\tau_1}, \ldots, B_{\tau_N})$ have the same law. Moreover, one can impose, for $n = 0, 1, \ldots, N-1$,*

$$\mathbf{E}\big[\tau_{n+1} - \tau_n | B[0,\tau_n]\big] = \mathbf{E}\big[(B_{\tau_{n+1}} - B_{\tau_n})^2 | B[0,\tau_n]\big] \tag{3.14}$$

and

$$\tau_{n+1} \le \inf\{t \ge \tau_n : |B_t - B_{\tau_n}| \ge 2\delta\}. \tag{3.15}$$

The proof can be found in many probability textbooks, including [9] and [37]. Often, it is stated for just one random variable M_1; for a statement in terms of martingales, see, for instance, [8] and [43]. The relation (3.15) is not stated explicitly in these references (since the assumption that the increments of M_n are bounded is weakened), but is a consequence of the proof. It can also be derived a posteriori from $\mathbf{E}[\tau_{n+1} - \tau_n] = \mathbf{E}[(M_{n+1} - M_n)^2] < \infty$, since the expected time for Brownian motion started outside an interval to hit the interval is infinite.

PROOF OF THEOREM 3.7. Since the hitting measure of a simple random walk from 0 is close to the hitting measure for Brownian motion when $\mathrm{rad}_0(D)$ is large (see, e.g., Section 5), it is clear that $W(0)$ is nearly uniform in $\partial\mathbb{U}$. It is therefore enough to show that $\vartheta(t/2)$ is close to standard Brownian motion.

Assume, without loss of generality, that $T \ge 1$. Pick $\delta = \delta(\epsilon, T) > 0$ small. Let r_0 be as in Proposition 3.4 and take $r_1 := 8\exp(20T)r_0$. Let γ^t denote the initial segment of γ such that $\psi_D(\gamma^t)$ has capacity t from 0. By the Schwarz lemma, $\psi_D'(0) \le \mathrm{rad}_0(D)^{-1}$. Therefore, the Koebe 1/4 theorem implies $\mathrm{rad}_0(D \setminus \gamma^t) \ge \exp(-t)\,\mathrm{rad}_0(D)/4$. Hence, if $\mathrm{rad}_0(D) \ge r_1$, Proposition 3.4 is valid not only for the initial domain D, but also for the domain D slitted by subarcs of γ, up to capacity $20T$.

As in Proposition 3.4, define m to be the first $j = 1, 2, \ldots$ such that $|\Delta_j| \ge \delta$ or $t_j \ge \delta^2$. Set $m_0 := 0$, $m_1 := m$ and inductively let m_{n+1} be the first $j \ge m_n + 1$ such that $|\Delta_j - \Delta_{m_n}| \ge \delta$ or $t_j - t_{m_n} \ge \delta^2$, whichever happens first. Let $\mathcal{F}_n$ denote the σ-field generated by $\gamma[0, m_n]$. Set

$$N := \lceil 10T\delta^{-2} \rceil.$$

Our choice of r_1 ensures that $t_{j+1} - t_j \le 2\delta^2$ for all $j < N$ and that $t_N \le 20T$. Hence, Proposition 3.4 holds for all domains D_{m_n} with $n < N$. Applying clause (iii) of Lemma 3.2 therefore gives

$$\mathbf{E}\big[\Delta_{m_{n+1}} - \Delta_{m_n} \big| \mathcal{F}_n\big] = O(\delta^3) \tag{3.16}$$

and

$$\mathbf{E}\big[(\Delta_{m_{n+1}} - \Delta_{m_n})^2 \big| \mathcal{F}_n\big] = 2\mathbf{E}\big[t_{m_{n+1}} - t_{m_n} \big| \mathcal{F}_n\big] + O(\delta^3). \tag{3.17}$$

For $n \le N$, set

$$M_n := \sum_{j=0}^{n-1} \big(\Delta_{m_{j+1}} - \Delta_{m_j} - \mathbf{E}\big[\Delta_{m_{j+1}} - \Delta_{m_j} \big| \mathcal{F}_j\big]\big). \tag{3.18}$$

Clearly, $M_0, \dots, M_N$ is a martingale for $\mathcal{F}_0, \dots, \mathcal{F}_N$. The definition of m_n and the choice of r_1 imply that $\|M_{n+1} - M_n\|_\infty \le 2\delta$.

By Lemma 3.8, we may couple $(M_0, \dots, M_N)$ with a standard Brownian motion with stopping times $\tau_0 \le \tau_1 \le \cdots \le \tau_N$ such that $B_{\tau_n} = M_n$ and (3.14) hold. Extend the coupling to include γ (this clearly can be done).

Note that the definition of t_{m_n} and (3.15) ensure that, for all $n < N$,

$$\begin{aligned} \sup\big\{\big|B_t - B_{\tau_n}\big| : t \in [\tau_n, \tau_{n+1}]\big\} &\le 2\delta, \\ \sup\big\{\big|\vartheta(t) - \Delta_{t_{m_n}}\big| : t \in \big[t_{m_n}, t_{m_{n+1}}\big]\big\} &\le 2\delta \end{aligned} \tag{3.19}$$

and (3.16) shows that

$$\sup\big\{\big|\Delta_{t_{m_n}} - M_n\big| : n \le N\big\} = O(\delta^3 N) = O(\delta T). \tag{3.20}$$

Hence, as $M_n = B_{\tau_n}$ and B_t is a.s. continuous, it remains to relate the capacities t_{m_n} with the stopping times τ_n and verify that $t_{m_N} > T$ with high probability. For this purpose, define

$$Y_n = \sum_{j=0}^{n-1} (M_{j+1} - M_j)^2.$$

We first show that Y_n is close to $2t_{m_n}$. Let $Z_n := Y_n - 2t_{m_n}$. By (3.18) and (3.16), we have, for $n < N$, $|M_{n+1} - M_n - \Delta_{t_{m_{n+1}}} + \Delta_{t_{m_n}}| = O(\delta^3)$. This implies $|M_{n+1} - M_n| = O(\delta)$ and hence also

$$Y_{n+1} - Y_n = (M_{n+1} - M_n)^2 = \big(\Delta_{t_{m_{n+1}}} - \Delta_{t_{m_n}}\big)^2 + O(\delta^4).$$

Consequently, (3.17) gives

$$\mathbf{E}[Z_{n+1} - Z_n | \mathcal{F}_n] \le O(\delta^3).$$

From the fact that the increments of t_{m_n} and those of Y_n are bounded by $O(\delta^2)$,

we also have $\mathbf{E}[(Z_{n+1} - Z_n)^2 | \mathcal{F}_n] \le O(\delta^4)$. Set $Z'_n := Z_n - \sum_{j=1}^{n} \mathbf{E}[Z_j - Z_{j-1} | \mathcal{F}_{j-1}]$. Since this is an $\mathcal{F}_n$-martingale, we have $\mathbf{E}[{Z'_N}^2] = \sum_{j=1}^{N} \mathbf{E}[(Z'_j - Z'_{j-1})^2]$ and the above estimates give $\mathbf{E}[{Z'_N}^2] = O(N\delta^4)$. Assuming $N\delta^3 < \delta^{1/2}/2$, without loss of generality, and applying Doob's maximal inequality (see [37], II.1.7) for L^2-martingales to Z'_n, we get

$$\mathbb{P}\Big[\max_{n\le N} |Y_n - 2t_{m_n}| > \delta^{1/2}\Big] = O(N\delta^3) = O(T\delta). \tag{3.21}$$

By the definition of the t_{m_n}, we have $Y_{n+1} - Y_n + t_{m_{n+1}} - t_{m_n} \ge \delta^2$. Summing gives $Y_N + t_{m_N} \ge N\delta^2 \ge 10T$. Therefore, (3.21) implies

$$\mathbf{P}\big[t_{m_N} < 2T\big] = O(T\delta). \tag{3.22}$$

We now show that with high probability τ_n is also close to Y_n for every $n \le N$. By (3.15), it is clear that $\mathbf{E}[(\tau_{n+1} - \tau_n)^2 | B[0, \tau_n]] = O(\delta^4)$ and therefore

$$\mathbf{E}\big[\big((\tau_{n+1} - Y_{n+1}) - (\tau_n - Y_n)\big)^2 | B[0, \tau_n]\big] = O(\delta^4).$$

Also, (3.14) gives

$$\mathbf{E}\big[(\tau_{n+1} - Y_{n+1}) - (\tau_n - Y_n) | B[0, \tau_n]\big] = 0.$$

Doob's inequality therefore implies

$$\mathbf{P}\Big[\max_{n\le N} |\tau_n - Y_n| > \delta^{1/2}\Big] = O(T\delta).$$

Combining this with (3.21) leads to

$$\mathbf{P}\Big[\max_{n\le N} |\tau_n - 2t_{m_n}| > \delta^{1/2}\Big] = O(T\delta).$$

Since B_t is a.s. continuous, together with (3.22), (3.19) and (3.20), this completes the proof. □

3.4. *Convergence with respect to a stronger topology.* Theorem 3.7 provides a kind of convergence of the loop-erased random walk to SLE_2. As we will see in this section, this kind of convergence suffices, for example, to show that the scaling limit with respect to the Hausdorff metric of the union of $\partial\mathbb{U}$ and the LERW in $\mathbb{U}$ is the union of $\partial\mathbb{U}$ and the SLE_2 path.

Let $\alpha : [0, 1] \to \mathbb{C}$ and $\beta : [0, 1] \to \mathbb{C}$ be two continuous paths. Define

$$\rho(\alpha, \beta) := \inf_{\phi\in\Phi} \sup_{t\in[0,1]} |\alpha(t) - \beta \circ \phi(t)|,$$

where Φ is the collection of all monotone nondecreasing continuous maps from $[0, 1]$ onto $[0, 1]$. It is an easy well-known fact that ρ is a metric on equivalence classes of paths, where two paths α and β are equivalent if $\alpha \circ \phi_1 = \beta \circ \phi_2$, where

$\phi_1, \phi_2 \in \Phi$. Since $\rho(\alpha, \beta)$ does not depend on the particular parameterization of α or β, the metric ρ is also defined for paths on intervals other than $[0, 1]$.

To explain our present goal, let us point out that there is a sequence of paths α_n from 1 to 0 in $\overline{\mathbb{U}}$ such that their Loewner driving functions $W_n(t)$ converge uniformly to the constant 1 but α_n does not converge to the path $\alpha(t) = 1 - t$, $t \in [0, 1]$, in the metric ρ, although the driving function for α (reparameterized by capacity) is the constant 1. For example, we may take α_n as the polygonal path through the points $a_1, b_1 + in^{-2}, a_2, b_2 - in^{-2}, a_3, b_3 + in^{-2}, \ldots, a_{\lfloor n/2 \rfloor}, 0$, where $a_j := 1 - n^{-1} + (jn)^{-1}$ and $b_j := 1 - j/n$.

THEOREM 3.9 (LERW image in $\overline{\mathbb{U}}$ converges). *For any sequence $D_n \in \mathfrak{D}$ with $\operatorname{rad}_0(D_n) \to \infty$, if μ_n denotes the law of $\tilde{\gamma}^n := \psi_{D_n} \circ \gamma^n$, where γ^n is the time reversal of the LERW from 0 to ∂D_n, then μ_n converges weakly (with respect to the metric ρ) to the law of the radial* SLE_2 *path started uniformly on the unit circle.*

The outline of the proof is as follows. We define a suitable family of compact subsets of the space of simple paths from $\partial\mathbb{U}$ to 0 in $\overline{\mathbb{U}}$, which we can use to show that the sequence μ_n is tight. (See, e.g., [9] for background on weak convergence and the notion of tightness.) This implies that a subsequence of μ_n converges weakly to some probability measure. Theorem 3.9 then shows that the law of SLE_2 is the unique possible subsequential limit.

In order to prove tightness, we will use properties of the loop-erased random walk proved in [38]. The actual details will require some background in the geometric theory of conformal maps. In particular, some properties of extremal distance (a.k.a. extremal length) will be used. See, for example, [2] for background. The basic ideas that are used in the proof are taken from [3] and [38].

For a simply connected $D \subsetneq \mathbb{C}$ containing 0, let $\mathcal{X}_0(D)$ denote the space of all simple paths $\gamma : [0, \infty] \to \overline{D}$ from ∂D to 0 in $\overline{D}$, which intersect ∂D only at the starting point. Given a monotone nondecreasing function $\Upsilon : (0, \infty) \to (0, 1]$, let $\mathcal{X}_\Upsilon(D) \subset \mathcal{X}_0(D)$ denote the space of all simple paths $\gamma \in \mathcal{X}_0(D)$ such that, for every $0 \le s_1 < s_2$,

$$\operatorname{dist}(\gamma[0, s_1] \cup \partial D, \gamma[s_2, \infty]) / \operatorname{rad}_0(D) \ge \Upsilon\big(\operatorname{diam}(\gamma[s_1, s_2]) / \operatorname{rad}_0(D)\big).$$

Note that whether $\gamma \in \mathcal{X}_\Upsilon(D)$ does not depend on the parameterization of γ and is scaling invariant.

LEMMA 3.10 (Compactness). *Let $\Upsilon : (0, \infty) \to (0, 1]$ be monotone nondecreasing. Then $\mathcal{X}_\Upsilon(\mathbb{U})$ is compact in the topology of convergence with respect to ρ.*

PROOF. We use an idea from [3]. For all $n \in \mathbb{N}$, let Z_n be a finite collection of points such that the open balls $\mathcal{B}(z, 2^{-n})$, $z \in Z_n$, cover $\mathbb{U}$. Given a set $K \subset \overline{\mathbb{U}}$

and a point $z \in Z_n$, let $s(K, z, n)$ denote the diameter of $K \cap \mathcal{B}(z, 2^{1-n})$ in the metric obtained from the Euclidean metric on the disk $\overline{\mathcal{B}(z, 2^{1-n})}$ by collapsing the boundary $\partial\mathcal{B}(z, 2^{1-n})$ to a single point. [In other words, this metric $d(x, y)$ is defined as $d(x, y) = \min\{|x - y|, \operatorname{dist}(x, \partial\mathcal{B}) + \operatorname{dist}(y, \partial\mathcal{B})\}$, where $\mathcal{B} = \mathcal{B}(z, 2^{1-n})$.]

Fix $\gamma \in \mathcal{X}_\Upsilon(\mathbb{U})$. Given $t \geq 0$, let

$$s(t) = s_\gamma(t) := \sum_{n\in\mathbb{N}} \sum_{z\in Z_n} \frac{s(\gamma[0, t], z, n)}{|Z_n|}.$$

Clearly, $s_\gamma(t) \leq \sum_{n\geq 0} 2^{2-n} = 8$, and $s : [0, \infty] \to [0, \infty)$ is continuous and strictly monotone increasing. (To verify that s is strictly monotone increasing, note that, if $t_2 > t_1 \geq 0$, then there is some $n \in \mathbb{N}$ such that $\operatorname{dist}(\gamma(t_2), \gamma[0, t_1]) \geq 2^{2-n}$, and so $s(\gamma[0, t_2], z, n) \geq s(\gamma[0, t_1], z, n) + 2^{-n}$ if $z \in Z_n$ satisfies $\gamma(t_2) \in \mathcal{B}(z, 2^{-n})$.) Let $\hat{\gamma}(s)$ be γ parameterized by s; that is, $\hat{\gamma} = \gamma \circ s^{-1}$. Let $s_1 < s_2$ and set $\epsilon := \operatorname{diam} \hat{\gamma}[s_1, s_2] > 0$. Then $\operatorname{dist}(\hat{\gamma}(s_2), \hat{\gamma}[0, s_1] \cup \partial\mathbb{U}) \geq \Upsilon(\epsilon)$. By the argument for strict monotonicity given above, this shows that $s_2 - s_1 \geq 2^{-n}/|Z_n|$, where $n := \min\{k \in \mathbb{N} : 2^{2-k} \leq \Upsilon(\epsilon)\}$. Therefore, $\hat{\gamma}$ satisfies an equicontinuity estimate. By the Arzela–Ascoli theorem, it follows that the closure of $\mathcal{X}_\Upsilon(\mathbb{U})$ is compact in the ρ metric. It is also clear that $\mathcal{X}_\Upsilon(\mathbb{U})$ is closed. □

Our next goal is to use these compact sets to prove tightness, and we start by observing that the diameter is tight.

LEMMA 3.11 (Diameter is tight). *There are constants $c, C > 0$ such that for every $D \in \mathfrak{D}$ and every $r \geq 1$ the simple random walk Γ starting from 0 and stopped on hitting ∂D satisfies*

$$\mathbf{P}[\operatorname{diam}(\Gamma) \geq r \operatorname{rad}_0(D)] \leq Cr^{-c}.$$

Consequently, the same estimate holds for the loop erasure γ.

The first statement is an easy well-known fact. Since the complement of D is connected and unbounded, if the random walk makes a loop separating the circle $\operatorname{rad}_0(D)\partial\mathbb{U}$ from the circle $(r/2)\operatorname{rad}_0(D)\partial\mathbb{U}$ before hitting the latter circle, then it must hit ∂D before $(r/2)\operatorname{rad}_0(D)\partial\mathbb{U}$. Thus, the lemma is easily proved directly and also follows from the convergence of a simple random walk to Brownian motion. A rather precise form of this estimate for the random walk, where $c = 1/2$, is known as the discrete Beurling theorem (see [25], Theorem 2.5.2).

LEMMA 3.12 (Tameness). *For every $\epsilon > 0$, there is some monotone nondecreasing $\Upsilon : (0, \infty) \to (0, 1]$ and some $r_0 > 0$ such that for every $D \in \mathfrak{D}$ with $\operatorname{rad}_0(D) \geq r_0$ its time-reversed loop-erased walk $\gamma = \gamma_D$ satisfies*

$$\mathbf{P}[\gamma \in \mathcal{X}_\Upsilon(D)] \geq 1 - \epsilon.$$

PROOF. The proof is essentially contained in the proof of [38], Theorem 1.1, where it is established that every subsequential scaling limit of LERW is a.s. a simple path. We will not repeat the complete proof from [38] here, but indicate how it may be adapted to yield the statement of the lemma.

Let $\epsilon > 0$. Clearly, $\gamma \in \mathcal{X}_0(D)$. If $\gamma \notin \mathcal{X}_\Upsilon(D)$, then there are $0 \le s_1 < s_2 < \infty$ such that the distance between $\gamma[0, s_1] \cup \partial D$ and $\gamma[s_2, \infty]$ is smaller than $\operatorname{rad}_0(D)\Upsilon(\operatorname{diam}\gamma[s_1, s_2]/\operatorname{rad}_0(D))$. Let us first deal with the case where the distance between $\gamma[s_2, \infty]$ and ∂D is small. Let Γ be the walk generating the time reversal of γ, and let t_n be the first time t where the distance from $\Gamma(t)$ to ∂D is smaller than $2^{-n}\operatorname{rad}_0(D)$ and let $\tau = \inf\{t : \Gamma(t) \in \partial D\}$. By the Markov property of Γ at time t_n and Lemma 3.11,

$$\mathbf{P}\big[\operatorname{diam}\Gamma[t_n, \tau] > 2^{-n/2}\operatorname{rad}_0(D)\big] \le C2^{-cn/2}.$$

Consequently, there is an $N = N(\epsilon)$ such that with probability $1 - \epsilon/2$ for every integer $n \ge N$ we have $\operatorname{diam}\Gamma[t_n, \tau] \le 2^{-n/2}\operatorname{rad}_0(D)$. In this case, if $\operatorname{diam}\gamma[0, s_2] > 2^{-n/2}\operatorname{rad}_0(D)$, where $n > N$, then $\gamma[0, s_2]$ is not contained in $\Gamma[t_n, \tau]$, which implies that $\gamma[s_2, \infty] \subset \Gamma[0, t_n]$ and gives $\operatorname{dist}(\gamma[s_2, \infty], \partial D) \ge 2^{-n}\operatorname{rad}_0(D)$. In other words, if Υ satisfies

$$\Upsilon(t) < \min\{t^2, 2^{-2N}\}/4, \tag{3.23}$$

then with probability at least $1 - \epsilon/2$, for every $s_1, s_2 \in [0, \infty]$,

$$\operatorname{dist}(\partial D, \gamma[s_2, \infty]) \ge \operatorname{rad}_0(D)\Upsilon\big(\operatorname{diam}\gamma[s_1, s_2]/\operatorname{rad}_0(D)\big). \tag{3.24}$$

We now focus on the case where the distance between $\gamma[0, s_1]$ and $\gamma[s_2, \infty)$ is small. We shall say that γ has a (β, α)-quasi-loop if there are $0 < s_1 < s_2 < \infty$ such that $|\gamma(s_1) - \gamma(s_2)| \le \alpha\operatorname{rad}_0(D)$ but $\operatorname{diam}\gamma[s_1, s_2] \ge \beta\operatorname{rad}_0(D)$. Note that if there are $0 < s_1 < s_2 < \infty$ such that $\operatorname{dist}(\gamma[0, s_1], \gamma[s_2, \infty]) < \alpha\operatorname{rad}_0(D)$ and $\operatorname{diam}\gamma[s_1, s_2] \ge \beta\operatorname{rad}_0(D)$, then γ has a (β, α)-quasi-loop. Let $\mathcal{A}(\beta, \alpha)$ denote the event that γ has a (β, α)-quasi-loop. Assume, for the moment, that, for all $n \ge 0$,

$$\lim_{\alpha \searrow 0} \mathbf{P}[\mathcal{A}(2^{-n}, \alpha)] = 0, \tag{3.25}$$

uniformly in D. Then we may take a decreasing sequence $\alpha_n \searrow 0$ such that $\sum_{n=1}^{\infty}\mathbf{P}[\mathcal{A}(2^{-n}, \alpha_n)] < \epsilon/2$ holds for every $D \in \mathfrak{D}$. Then with probability at least $1 - \epsilon/2$, γ has no $(2^{-n}, \alpha_n)$-quasi-loop for any $n = 1, 2, \ldots$. Assuming that $\Upsilon(t) < \alpha_n$ holds whenever $t \le 2^{1-n}$, $n \in \mathbb{N}$, and $\Upsilon(t) < \alpha_1$ for all t, on this event we also have

$$\operatorname{dist}(\gamma[0, s_1], \gamma[s_2, \infty]) \le \operatorname{rad}_0(D)\Upsilon\big(\operatorname{diam}\gamma[s_1, s_2]/\operatorname{rad}_0(D)\big)$$

for all $0 < s_1 < s_2 < \infty$. If we also assume (3.23), then together with (3.24) we get $\mathbf{P}[\gamma \in \mathcal{X}_\Upsilon(D)] \ge 1 - \epsilon$, completing the proof of the lemma. Thus, it remains to verify (3.25).

Let $\mathcal{A}(z_0, \beta, \alpha)$ denote the event that there are $0 < s_1 < s_2 < \infty$ such that $|\gamma(s_1) - \gamma(s_2)| \le \alpha \operatorname{rad}_0(D)$, $\gamma(s_1), \gamma(s_2) \in \mathcal{B}(z_0, \beta \operatorname{rad}_0(D)/4)$ and $\operatorname{diam}(\gamma[s_1, s_2]) \ge \beta \operatorname{rad}_0(D)$. In particular, this implies that $\gamma[s_1, s_2]$ is not contained in the interior of $\mathcal{B}(z_0, \beta \operatorname{rad}_0(D)/2)$. Assume that $8\alpha < \beta$. By Lemma 3.11, there is an $R = R(\epsilon) > 0$ such that with probability at least $1 - \epsilon/2$ we have $\gamma[0, \infty] \subset \mathcal{B}(0, R \operatorname{rad}_0(D))$. There is a collection $\{z_1, z_2, \ldots, z_k\}$ of points such that every disk of radius $2\alpha \operatorname{rad}_0(D)$ with center in $\mathcal{B}(0, R \operatorname{rad}_0(D))$ is contained in one of the k balls $\mathcal{B}(z_j, \beta \operatorname{rad}_0(D)/2)$, $j = 1, 2, \ldots, k$, and we may take $k < c((R/\beta)^2 + 1)$, where c is an absolute constant. On the event $\gamma[0, \infty] \subset \mathcal{B}(0, R \operatorname{rad}_0(D))$, we have $\mathcal{A}(\beta, \alpha) \subset \bigcup_{j=1}^{k} \mathcal{A}(z_j, \beta, \alpha)$. Since $\epsilon > 0$ was arbitrary and $\mathbf{P}[\gamma[0, \infty] \subset \mathcal{B}(0, R \operatorname{rad}_0(D))] \ge 1 - \epsilon/2$, it is therefore sufficient to show that $\mathbf{P}[\mathcal{A}(z_j, \beta, \alpha)] \to 0$ as $\alpha \to 0$ uniformly in D. The proof of this statement is given (with minor changes in the setup) in [38], Theorem 1.1. $\square$

Let $\mathcal{X}^r_\Upsilon(D)$ denote the set of paths $\gamma \in \mathcal{X}_\Upsilon(D)$ that are contained in the ball of radius $r \operatorname{rad}_0(D)$ about 0. Given $\gamma \in \mathcal{X}_0(D)$, let $\gamma^* : [0, \infty) \to \overline{\mathbb{U}}$ denote the path $\psi_D \circ \gamma$, parameterized by capacity.

LEMMA 3.13 (Tameness invariance). *For every monotone nondecreasing* $\Upsilon : (0, \infty) \to (0, 1]$ *and every* $r > 1$, *there is a monotone nondecreasing* $\Upsilon^* : (0, \infty) \to (0, 1]$ *such that, for all* $D \in \mathfrak{D}$ *and* $\gamma \in \mathcal{X}^r_\Upsilon(D)$, $\gamma^* \in \mathcal{X}_{\Upsilon^*}(\mathbb{U})$.

PROOF. Let $D \in \mathfrak{D}$, $\gamma \in \mathcal{X}^r_\Upsilon(D)$ and $0 \le s_1' < s_2' \le \infty$. Note that there exist s_1 and s_2 satisfying $s_1' \le s_1 \le s_2 \le s_2'$ such that

$$\operatorname{diam}(\gamma^*[s_1, s_2]) \ge \operatorname{diam}(\gamma^*[s_1', s_2'])/4$$

and

$$\operatorname{dist}(0, \gamma^*[s_1, s_2]) \ge \operatorname{diam}(\gamma^*[s_1, s_2]). \tag{3.26}$$

Since

$$\operatorname{dist}\big(\gamma^*[0, s_1'] \cup \partial\mathbb{U}, \gamma^*[s_2', \infty)\big) \ge \operatorname{dist}\big(\gamma^*[0, s_1] \cup \partial\mathbb{U}, \gamma^*[s_2, \infty)\big), \tag{3.27}$$

it is sufficient to give a lower bound of the right-hand side of (3.27) in terms of $\epsilon := \operatorname{diam}(\gamma^*[s_1, s_2])$.

The Schwarz lemma gives $\psi_D'(0) \le 1/\operatorname{rad}_0(D)$. Therefore, by the Koebe 1/4 theorem (applied to the restriction of ψ_D^{-1} to $\epsilon\mathbb{U}$) and (3.26), $\operatorname{dist}(0, \gamma[s_1, s_2]) > c_1 \operatorname{rad}_0(D)$, where $c_1 = \epsilon/4$. On the other hand, the harmonic measure in $\mathbb{U}$ from 0 of $\gamma^*[s_1, s_2]$ is at least c_2, where $c_2 = c_2(\epsilon) > 0$, so that the harmonic measure in D from 0 of $\gamma[s_1, s_2]$ is at least c_2. Hence,

$$\operatorname{diam} \gamma[s_1, s_2] \ge c_3 \operatorname{rad}_0(D), \tag{3.28}$$

where $c_3 = c_3(\epsilon)$.

Also set $\delta := \operatorname{dist}(\gamma^*[0, s_1] \cup \partial\mathbb{U}, \gamma^*[s_2, \infty])$. Since

$$\operatorname{diam} \gamma^*[s_2, \infty] \geq \operatorname{dist}(0, \gamma^*[s_1, s_2]) \geq \epsilon,$$

the extremal distance between $\gamma^*[0, s_1] \cup \partial\mathbb{U}$ and $\gamma^*[s_2, \infty]$ is at most $\phi_1(\delta, \epsilon) > 0$, where ϕ_1 is some function satisfying $\phi_1(\delta, \epsilon) \to 0$ as $\delta \downarrow 0$. By the conformal invariance of the extremal distance, this implies that the extremal distance between $\gamma[0, s_1] \cup \partial D$ and $\gamma[s_2, \infty]$ is at most $\phi_1(\delta, \epsilon)$. Because γ is contained in the disk of radius $r \operatorname{rad}_0(D)$ about 0, this implies that

$$\operatorname{dist}(\gamma[0, s_1] \cup \partial D, \gamma[s_2, \infty]) \leq \phi_2(\delta, \epsilon) r \operatorname{rad}_0(D),$$

where $\phi_2 \to 0$ as $\delta \downarrow 0$. Because $\gamma \in \mathcal{X}_\Upsilon(D)$, (3.28) and this together imply

$$\phi_2(\delta, \epsilon) r \geq \Upsilon(c_3(\epsilon)),$$

which gives a positive lower bound for $\delta = \operatorname{dist}(\gamma^*[0, s_1] \cup \partial\mathbb{U}, \gamma^*[s_2, \infty])$ in terms of Υ, r and $\epsilon = \operatorname{diam}(\gamma^*[s_1, s_2])$. This completes the proof. □

LEMMA 3.14 (Convergence relations). *Suppose W^n, W are continuous functions from $[0, \infty)$ to $\partial\mathbb{U}$ such that $W^n \to W$ locally uniformly. Let g_t^n, g_t be the corresponding solutions to Loewner's radial equation and set $f_t^n = (g_t^n)^{-1}$, $f_t = g_t^{-1}$. Then $f_t^n \to f_t$ locally uniformly on $[0, \infty) \times \mathbb{U}$. If there are continuous curves $\gamma^n : [0, \infty) \to \overline{\mathbb{U}}$ such that, for all $t \geq 0$, the image of f_t^n is the component of 0 in $\mathbb{U} \setminus \gamma^n[0, t]$ and there is a $\gamma : [0, \infty) \to \overline{\mathbb{U}}$ such that $\gamma^n \to \gamma$ locally uniformly on $[0, \infty)$, then for all $t \geq 0$ the image of f_t is the component of 0 in $\mathbb{U} \setminus \gamma[0, t]$.*

PROOF. Since g_t is obtained by flowing along a vector field depending on W, the inverse f_t is obtained by flowing along the opposite field, with the time reversed. Hence, the first statement is an immediate consequence of the principle that solutions of ODE depend continuously on the parameters of the ODE. The second statement is an immediate consequence of the Carathéodory kernel theorem (see [35], Theorem 1.8). □

PROOF OF THEOREM 3.9. Let W^n denote the Loewner parameter of $\tilde{\gamma}^n$ and let $\hat{\mu}_n$ denote the law of the pair $(\tilde{\gamma}^n, W^n)$. By Theorem 3.7, we know that the law of W^n tends weakly to the law of Brownian motion. Lemmas 3.10–3.13 show that the set of measures $\{\mu_n\}$ is tight with respect to the metric ρ. Consequently, the sequence $\hat{\mu}_n$ is also tight. Prokhorov's theorem (e.g., [9] and [37]) implies that there is a subsequence such that $\hat{\mu}_n$ converges weakly along the subsequence. Let $\hat{\mu}$ be any subsequential weak limit and let $(\tilde{\gamma}, W)$ be a sample from $\hat{\mu}$. The lemmas show that $\tilde{\gamma}$ is a.s. a simple path, and Theorem 3.7 shows that W is Brownian motion (with time scaled). By the properties of weak convergence, we may couple the subsequence of pairs $(\tilde{\gamma}^n, W^n)$ and $(\tilde{\gamma}, W)$ so that a.s. $\rho(\tilde{\gamma}^n, \tilde{\gamma}) \to 0$ and $W^n \to W$ locally uniformly.

Recall that the capacity is continuous with respect to the metric ρ; that is, if $\beta, \beta_n : [0,1] \to \overline{\mathbb{U}} \setminus \{0\}$ and $\rho(\beta_n, \beta) \to 0$, then the capacity of $\beta_n[0,1]$ tends to the capacity of $\beta[0,1]$. (In fact, it is enough that $\beta_n[0,1]$ tends to $\beta[0,1]$ in the Hausdorff metric.) Indeed, this follows immediately from Carathéodory's kernel theorem (see [35], Theorem 1.8) and the fact that the local uniform convergence of conformal maps implies the convergence of the derivatives (by Cauchy's formula for the derivative).

Since $\tilde{\gamma}$ is almost surely a simple path, the capacity of $\tilde{\gamma}$ increases strictly, and one can parameterize the path continuously by its capacity. We also parameterize the paths $\tilde{\gamma}^n$ by capacity. The next goal is to show that $\tilde{\gamma}^n \to \tilde{\gamma}$ locally uniformly on $[0,\infty)$. Since $\rho(\tilde{\gamma}, \tilde{\gamma}^n) \to 0$, there are strictly monotone continuous onto maps $\varepsilon_n : [0,\infty) \to [0,\infty)$ so that $\tilde{\gamma}^n \circ \varepsilon_n \to \tilde{\gamma}$ locally uniformly. If $t_n \in [0,\infty)$ and $t_n \to t \in [0,\infty)$, then it follows from the continuity of capacity with respect to ρ that $\varepsilon_n(t_n) \to t$ [because if s is a subsequential limit of $\varepsilon_n(t_n)$, then the capacity of $\tilde{\gamma}(s)$ must be t; i.e., $s = t$]. This implies that ε_n converges to the identity map $t \mapsto t$, locally uniformly. By the continuity of $\tilde{\gamma}$, it follows that $\tilde{\gamma} \circ \varepsilon_n^{-1} \to \tilde{\gamma}$ locally uniformly. This gives $\tilde{\gamma}^n \to \tilde{\gamma}$ locally uniformly.

We can now finally apply Lemma 3.14 to show that $\tilde{\gamma}$ is the SLE_2 path. As the law of the limit $\tilde{\gamma}$ does not depend on the subsequence, the theorem follows. □

In the following proof of Theorem 1.1, the main technical point is that we do not make any smoothness assumptions on ∂D. If ∂D is a simple closed path, the theorem follows easily from Theorem 3.9, because the suitably normalized conformal maps from $\mathbb{U}$ to the discrete approximations of D converge uniformly to the conformal map onto D.

PROOF OF THEOREM 1.1. Let D_δ be the component of 0 in the complement of all the closed square faces of the grid $\delta\mathbb{Z}^2$ intersecting ∂D. Let γ_δ be the time reversal of the loop-erased random walk from 0 to ∂D_δ and let β be the radial SLE_2 path in $\overline{\mathbb{U}}$. Let $\phi_\delta : \mathbb{U} \to D_\delta$ be the conformal map satisfying $\phi_\delta(0) = 0$ and $\phi_\delta'(0) > 0$ and let $\phi : \mathbb{U} \to D$ be the conformal map satisfying $\phi(0) = 0$, $\phi'(0) > 0$. Theorem 3.9 tells us that we may couple β with each of the paths γ_δ such that $\rho(\phi_\delta^{-1} \circ \gamma_\delta, \beta) \to 0$ in probability as $\delta \downarrow 0$. Moreover, the proof shows that if we use the capacity parameterization for both, then, in probability,

$$\sup\{|\phi_\delta^{-1} \circ \gamma_\delta(t) - \beta(t)| : t \geq 0\} \to 0.$$

(There is no problem with convergence in a neighborhood of $t = \infty$, because we know that the weak limit of $\phi_\delta^{-1} \circ \gamma_\delta$ with respect to ρ is a simple path tending to 0 as $t \to \infty$.)

The Carathéodory kernel theorem (see [35], Theorem 1.8) implies that $\phi_\delta \to \phi$ uniformly on compact subsets of $\mathbb{U}$ as $\delta \searrow 0$. Consequently, the above gives

$$\forall\, t_0 > 0, \qquad \sup\{|\gamma_\delta(t) - \phi \circ \beta(t)| : t \geq t_0\} \to 0, \tag{3.29}$$

in probability. Let $\epsilon > 0$ be small. Then, by Lemma 3.11, there is an $\epsilon' > 0$ such that for every $D' \in \mathfrak{D}$ the probability that a simple random walk from 0 gets to distance $\mathrm{rad}_0(D')/(2\epsilon')$ before hitting $\partial D'$ is less than $\epsilon/2$. Let A be the connected component of 0 in the set of points in $D \cap (\mathrm{rad}_0(D)/\epsilon')\mathbb{U}$ having distance at least $\epsilon\epsilon' \mathrm{rad}_0(D)$ from ∂D. By considering the first point where the random walk generating γ_δ exits A, it follows that, with probability at least $1 - \epsilon$, the diameter of $\gamma_\delta[0, \infty] \setminus A$ is at most $\epsilon \,\mathrm{rad}_0(D) + \delta$. Now note that there is a compact $A' \subset \mathbb{U}$ such that $\phi_\delta^{-1}(A) \subset A'$ for all sufficiently small δ, since $\phi_\delta \to \phi$ uniformly on compacts. Therefore, there is some $t_1 > 0$ such that $\gamma_\delta[0, t_1] \cap A = \varnothing$ a.s. for all sufficiently small $\delta > 0$. In particular,

$$\mathbf{P}\big[\mathrm{diam}\,\gamma_\delta[0, t_1] > \epsilon \,\mathrm{rad}_0(D) + \delta\big] < \epsilon. \tag{3.30}$$

If we take $t_2 \in (0, t_1)$, then taking $\delta \searrow 0$ in (3.29) implies

$$\mathbf{P}\big[\mathrm{diam}\,\phi \circ \beta[t_2, t_1] > 2\epsilon \,\mathrm{rad}_0(D)\big] < \epsilon.$$

Since this holds for every t_2, it follows that

$$\mathbf{P}\big[\mathrm{diam}\,\phi \circ \beta(0, t_1] > 2\epsilon \,\mathrm{rad}_0(D)\big] < \epsilon.$$

Using this with (3.30) and choosing $t_0 = t_1$ in (3.29) gives

$$\mathbf{P}\big[\sup\{|\gamma_\delta(t) - \phi \circ \beta(t)| : t > 0\} < 3\epsilon \,\mathrm{rad}_0(D)\big] \to 1.$$

Since this holds for every $\epsilon > 0$, the theorem follows. □

4. The UST Peano curve.

4.1. *Setup.* The UST Peano curve is obtained as the interface between the UST and the dual UST. The setup which corresponds to chordal SLE_8 is where there is symmetry between the UST and the dual UST. Loosely speaking, the UST is the uniform spanning tree on the grid inside a domain D but with an entire arc $\alpha \subset \partial D$ on the boundary identified (wired) as a single vertex, and the dual UST also has an arc $\beta \subset \partial D$ on the boundary which is identified. The arcs α and β are essentially complementary arcs. See Figure 5, where D is approximately a rectangle. As mentioned in Section 1, it was conjectured in [36] that, for an analogous setup, the interface defined for the critical random cluster models with $q \in (0, 4]$ converges to SLE_κ, where $\kappa = \kappa(q) \in [4, 8)$.

A combinatorial framework is necessary in order to be more precise. There are several different possible setups that would work, and the following is somewhat arbitrary.

If a tree T lies in the grid $\mathbb{Z}^2$, then its dual tree $T^\dagger$ will lie in the dual grid $(\mathbb{Z} + 1/2)^2$, and the Peano path γ will lie in the graph G whose vertices are $(1/4 + \mathbb{Z}/2)^2$ and where v, u neighbor iff $|v - u| = 1/2$. We have three kinds of vertices: elements of $\mathbb{Z}^2$ are the *primal* vertices, elements of $(1/2 + \mathbb{Z})^2$ are the *dual*

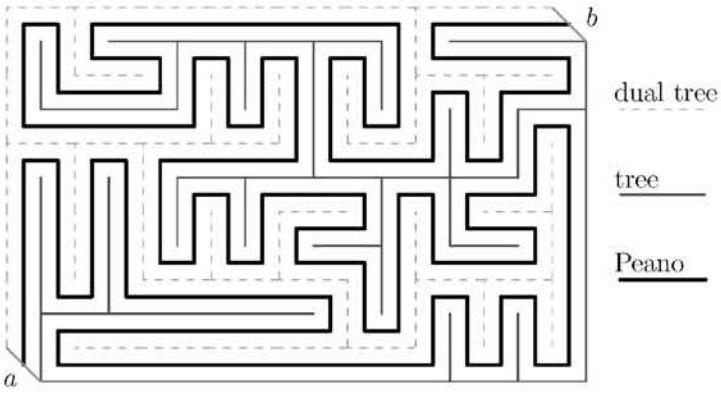


FIG. 5. *The tree, dual tree and Peano UST path* γ.

vertices and elements of $(1/4+\mathbb{Z}/2)^2$ are the *Peano* vertices. If $w \neq v$ are vertices of any kind, not necessarily the same, we say that they are *adjacent* if the distance between them is as small as it can be for distinct vertices of these particular kinds. In other words, if they are of the same kind, this means that they are neighbors, if $v \in (1/4+\mathbb{Z}/2)^2$ and $w \in \mathbb{Z}^2 \cup (1/2+\mathbb{Z})^2$, this means $\|v-w\|_\infty = 1/4$, while if $v \in \mathbb{Z}^2$ and $w \in (1/2+\mathbb{Z})^2$, this means $\|v-w\|_\infty = 1/2$.

Since there is no added complication, we consider a more general case where α and β are trees, rather than arcs. Let α be some finite tree in the primal grid $\mathbb{Z}^2$ and let β be a finite tree in the dual grid $(1/2+\mathbb{Z})^2$. Suppose that no edge of α intersects an edge of β. Further suppose that there are two Peano vertices $a, b \in (1/4+\mathbb{Z}/2)^2$ such that a is adjacent to both a primal vertex $\alpha_a \in \alpha$ and a dual vertex $\beta_a \in \beta$, and b is adjacent to both a primal vertex $\alpha_b \in \alpha$ and a dual vertex $\beta_b \in \beta$. See Figure 6. Note that the line segment $[\alpha_a, \beta_a]$ has a as its midpoint, and the line segment $[\alpha_b, \beta_b]$ has b as its midpoint. Let $D = D(\alpha, \beta, a, b)$ be the (unique) bounded

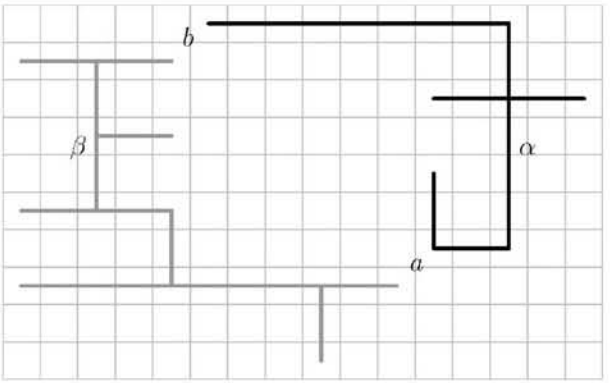


FIG. 6. *The boundary data and the Peano grid.*

connected component of $\mathbb{C} \setminus (\alpha \cup [\alpha_b, \beta_b] \cup \beta \cup [\beta_a, \alpha_a])$. Let $V_P = V_P(D)$ denote the collection of all Peano vertices in $\overline{D}$, and, as before, $V(D)$ denotes the collection of all primal vertices in D. Let $\ell = \ell(D)$ denote the cardinality of $V_P \setminus \{a, b\}$. By switching the role of a and b, if necessary, assume that D lies to the immediate right of the oriented segment $[\alpha_a, \beta_a]$. Let $\mathfrak{D}^*$ denote the collection of all domains obtained in this way.

Let $H = H(D)$ denote the subgraph of $\mathbb{Z}^2$ whose vertices are the vertices of α and $V(D)$ and whose edges are those edges on this set of vertices which do not intersect β. Since β is a tree, H is connected.

Since H is connected, there is at least one spanning tree T of H which contains α. If we replace α by T and apply the dual argument, it follows that there is also a tree $T^\dagger$ in the dual grid $(1/2 + \mathbb{Z})^2$, which is disjoint from T, contains β and whose vertices are the dual vertices in β and the dual vertices in D. In fact, $T^\dagger$ contains every dual edge lying in $\overline{D} \setminus T$.

We now need to give an orientation to the Peano grid G. Every edge in G is either on the boundary of a square face of G centered on a primal vertex or on the boundary of a square face of G centered on a dual vertex, and these two possibilities are exclusive. We orient the edges of G by specifying that the square faces of G containing a primal vertex are oriented clockwise, while those containing a dual vertex are oriented counterclockwise. When we want to emphasize the orientation of the edges, we write $G_\rightarrow$ instead of G. Note that the edges of G contained in a horizontal or vertical line all get the same direction in $G_\rightarrow$, and consecutive parallel lines get opposite orientations. For this reason, $G_\rightarrow$ is often called the *Manhattan lattice*.

Let $\gamma = \gamma(T)$ denote the set of all edges of $G_\rightarrow$ which do not intersect $T \cup T^\dagger$ and which have at least one endpoint in D. Let $v \in V_P \setminus \{a, b\}$ be some Peano vertex in D. Note that there are precisely two oriented edges of $G_\rightarrow$ with initial point v, say e_1 and e_2, where one of these, say e_1, intersects an edge f_1 of the primal grid $\mathbb{Z}^2$, and the other intersects an edge f_2 of the dual grid $(1/2 + \mathbb{Z})^2$. Note also that $f_1 \cap f_2 \neq \varnothing$. It therefore follows that exactly one of the edges f_1, f_2 is in $T \cup T^\dagger$. Consequently, exactly one of the edges e_1, e_2 is in γ. This shows that γ has out-degree 1 at every $v \in V_P \setminus \{a, b\}$. An entirely similar argument shows that γ has in-degree 1 at every such v. In particular, this shows that γ does not contain the entire boundary of a square face of G that does not contain a primal or dual vertex. If γ had a cycle, the cycle therefore would have to surround some primal or dual vertex. But as T and $T^\dagger$ are connected and disjoint from γ, this is impossible. It therefore follows that γ is an oriented simple path (i.e., self-avoiding path), and the endpoints of γ are a and b. Since we are assuming that D lies to the right of $[\alpha_a, \beta_a]$, the initial point of γ is a and the terminal point is b.

Conversely, suppose that $\gamma^* = (\gamma_0^*, \ldots, \gamma_{\ell+1}^*)$ is any oriented simple path in $G_\rightarrow$, respecting the orientation of $G_\rightarrow$, from a to b, whose vertices are V_P. For $n \in \{0, \ldots, \ell + 1\}$, let v_n be the (unique) primal vertex adjacent to γ_n^* and

let $v_n^\dagger$ be the dual vertex adjacent to γ_n^*. Note that v_n and v_{n+1} are either the same vertex or adjacent vertices when $n = \{0, \dots, \ell\}$. Let $\alpha_n = \alpha_n(\gamma^*)$ denote the union of α with the collection of all edges $[v_k, v_{k+1}]$ for $k < n$ such that $v_k \neq v_{k+1}$ and, similarly, let $\beta = \beta_n(\gamma^*)$ denote the union of β with the collection of all dual edges $[v_k^\dagger, v_{k+1}^\dagger]$ for $k < n$ such that $v_k^\dagger \neq v_{k+1}^\dagger$. Then $T(\gamma^*) := \alpha_{\ell+1}(\gamma^*)$ and $T^\dagger(\gamma^*) := \beta_{\ell+1}(\gamma^*)$ are obviously connected, and there are no edges in $T(\gamma^*)$ intersecting edges in $T^\dagger(\gamma^*)$. Now, $T(\gamma^*)$ cannot contain a cycle, for such a cycle would have to separate $T^\dagger(\gamma^*)$. Hence, $T(\gamma^*)$ is a spanning tree of H containing α. It is also clear that $\gamma^* = \gamma(T(\gamma^*))$. That is, $T \mapsto \gamma(T)$ is a bijection between the set of spanning trees of H containing α and the set of oriented paths in $G_\rightarrow \cap \overline{D}$ from a to b containing V_P. Hence, when T is the UST on H conditioned to contain α, γ is uniformly distributed among such Peano paths; it is the *UST Peano path* associated with (α, β, a, b).

Let $(a = w_0, w_1, \dots, w_{\ell+1} = b)$ be the order of the vertices in the UST Peano path γ. For $n \in \{0, 1, \dots, \ell\}$, let $\gamma[0, n]$ denote the initial arc of γ from w_0 to w_n. Since γ is uniformly distributed among simple oriented paths in $G_\rightarrow$ from a to b which contain V_P, we immediately get the following Markov property.

LEMMA 4.1 (Markovian property). *Fix any $n \in \{1, 2, \dots, \ell\}$. Conditioned on $\gamma[0, n]$, the distribution of $(\gamma \setminus \gamma[0, n]) \cup \{w_n\}$ is the same as that of the UST Peano curve associated with $(\alpha_n(\gamma), \beta_n(\gamma), w_n, b)$.*

This lemma will play the same role in the proof as Lemma 3.2 in the case of the LERW. We will also use the convergence of certain discrete harmonic functions toward their continuous counterparts. To facilitate this, we have to set the combinatorial notation for the discrete Dirichlet–Neumann problem.

Let H be a finite nonempty connected subgraph of $\mathbb{Z}^2$ with vertices V_H and let E_∂ denote the set of oriented edges in $\mathbb{Z}^2$ whose initial endpoint is in V_H, but whose unoriented version is not in H. Suppose $E_\partial = E_0 \cup E_1 \cup E_2$ is a disjoint union, where $E_0 \cup E_1 \neq \varnothing$. Suppose also that $\hat{h} : V_H \to [0, 1]$ is some function. For $v \in V_H$, set $\Delta_{H,E_0,E_1,E_2}\hat{h}(v) := \sum d\hat{h}[v, u]$, where the sum is over all neighbors u of v in $\mathbb{Z}^2$, and $d\hat{h}[v, u] := \hat{h}(u) - \hat{h}(v)$ when $[v, u] \notin E_\partial$, $d\hat{h}[v, u] := 0 - \hat{h}(v)$ when $[v, u] \in E_0$, $d\hat{h}[v, u] := 1 - \hat{h}(v)$ when $[v, u] \in E_1$ and $d\hat{h}[v, u] := 0$ when $[v, u] \in E_2$. Note that there is a unique $\hat{h} : V_H \to [0, 1]$ such that $\Delta_{H,E_0,E_1,E_2}\hat{h}(v) = 0$ in V_H: $\hat{h}(v)$ is the probability that a simple random walk on $H \cup E_0 \cup E_1$ started from v will use an edge in E_1 before using an edge of E_0. This $\hat{h}$ will be called the Δ_{H,E_0,E_1,E_2}-harmonic function.

PROPOSITION 4.2 (Dirichlet–Neumann approximation). *For every $\epsilon > 0$, there is an $r_0 = r_0(\epsilon)$ such that the following holds. Let $D \subset \mathbb{C}$ be a simply connected domain satisfying $\mathrm{rad}_0(D) \geq r_0$. Let $A_0, A_1 \subset \partial\mathbb{U}$ be two disjoint arcs, each of length at least ϵ, and set $A_2 := \partial\mathbb{U} \setminus (A_0 \cup A_1)$. Let $\eta \subset D$ be a simple*

closed path which surrounds 0, *such that each point of* η *is within distance* 5 *from* ∂D. *Suppose that* $A_0', A_1' \subset \eta$ *are two disjoint arcs,* $A_2' := \eta \setminus (A_0' \cup A_1')$ *and the triple* (A_0', A_1', A_2') *corresponds to* (A_0, A_1, A_2) *under* ψ_D, *in the sense that for each* $j = 0, 1, 2$ *and each* $p \in A_j'$ *there is a continuous path* $\sigma : [0, 1) \to D$ *satisfying* $\operatorname{diam} \sigma[0, 1) \le 5$, $\sigma(0) = p$, *and* $\lim_{s \uparrow 1} \psi_D \circ \sigma(s)$ *exists and is in* A_j.

Let H *be the component of* 0 *in the set of edges of* $\mathbb{Z}^2$ *that do not intersect* η. *For* $j = 0, 1, 2$, *let* E_j *denote the set of oriented edges* $[v, u]$ *intersecting* η, *where* v *is in* H, *and the first point of intersection from the direction of* v *is in* A_j'. *Let* $\hat{h}$ *denote the* $\Delta_{H, E_0, E_1, E_2}$*-harmonic function. Let* $h : \mathbb{U} \to [0, 1]$ *be the continuous harmonic function which has boundary value* 0 *on* A_0, 1 *on* A_1, *and satisfies the Neumann boundary condition on* A_2. *Then* $|\hat{h}(0) - h(0)| < \epsilon$.

The proof will be given in Section 5.4.

4.2. *Driving process convergence.* Let α, β, a, b and $D = D(\alpha, \beta, a, b)$ be as above and suppose now that $0 \in D$. As before, let ℓ denote the number of Peano vertices in D and let $\gamma = (\gamma(0), \dots, \gamma(\ell + 1))$ be the UST Peano path from a to b in $G_{\to} \cap \overline{D}$. For each $n \le \ell$, there are two domains that are naturally associated to $\gamma[0, n]$. The first one (as in Lemma 4.1) is $\tilde{D}_n := D(\alpha_n, \beta_n, \gamma(n), b)$ (see Figure 7). But $\tilde{D}_n$ is not so useful if we want to make estimates using Loewner's equation. We therefore also define $D_n := D \setminus \gamma[0, n]$. Let $\phi_0 = \phi : D \to \mathbb{H}$ be the conformal map which takes D to $\mathbb{H}$, takes b to ∞, takes a to 0 and satisfies $|\phi(0)| = 1$. Let $\phi_n : D_n \to \mathbb{H}$ be the conformal maps satisfying $\phi_n(z) - \phi_0(z) \to 0$ as $z \to b$ within D_n. Define $W_n := \phi_n(\gamma(n)) \in \mathbb{R}$. Also let t_n denote the capacity from ∞ in $\mathbb{H}$ of $\phi_0 \circ \gamma[0, n]$, so that $\phi_n \circ \phi_0^{-1}(z) = z + 2t_n/z + o(1/z)$ when $z \to \infty$ in $\mathbb{H}$.

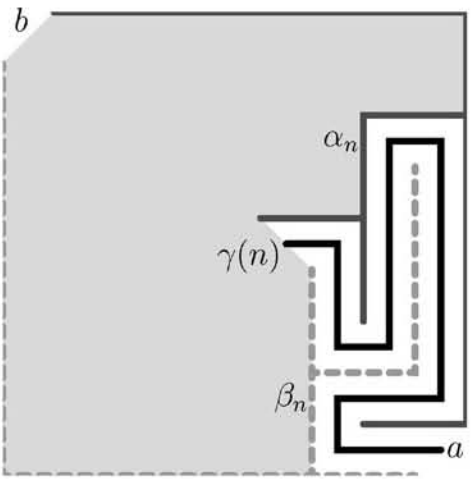


FIG. 7. *The domain* $\tilde{D}_n$ *is shaded.*

We now prove the analog of Proposition 3.4 for the UST Peano curve. Let $\mathfrak{H}_D(z, A)$ denote the continuous harmonic measure of A from z in the domain $D \setminus A$.

PROPOSITION 4.3 (The key estimate). *For every sufficiently small $\delta, \epsilon > 0$, there is some $r_0 = r_0(\delta, \epsilon)$ such that the following holds. Let γ, D_n, ϕ_n, W_n and t_n be as above, let $k \in \mathbb{N}$ and let m be the first $n \geq k$ such that $|W_n - W_k| \geq \delta$ or $t_n - t_k \geq \delta^2$. Then*

$$\mathbf{E}[W_m - W_k | D_k] = O(\delta^3) \tag{4.1}$$

and

$$\mathbf{E}[(W_m - W_k)^2 | D_k] = 8\mathbf{E}[t_m - t_k | D_k] + O(\delta^3), \tag{4.2}$$

provided that $\mathrm{rad}_0(D_k) \geq r_0$ *and* $\mathfrak{H}_{D_k}(0, \alpha_k(\gamma)) \in [\epsilon, 1 - \epsilon]$.

PROOF. Assume first $k = 0$. Let $v_0 \in V(D)$ be some vertex such that $|\phi(v_0)| \leq 2$ and $\mathrm{Im}\,\phi(v_0) \geq 1/2$, say. [As we have seen in Section 3.2, there is such a v_0 when $\mathrm{rad}_0(D)$ is large.] If $Q = [q, q']$ is a line segment where $q \in D$ is a dual vertex and $q' \in \alpha$ is the midpoint of a dual edge containing q, then let $\phi^*(Q) \in \mathbb{R}_+$ denote the limit of $\phi(z)$ as z tends to ∂D along Q (which always exists by [35], Proposition 2.14). Fix such a Q_0 satisfying $U := \phi^*(Q_0) \in [1/2, 2]$; there clearly is such Q_0 when r_0 is large, because the harmonic measure from 0 of any square of the dual grid adjacent to the boundary of D is small. Let $\eta \subset D$ be the set of points within distance $1/10$ from ∂D. Then η is a simple closed path. Consider it as oriented counterclockwise around the bounded domain of $\mathbb{C} \setminus \eta$. Let p_0 be the point of η closest to a, p_1 the point in $\eta \cap Q_0$ and p_2 the point of η closest to b. Let A_0' be the positively oriented subarc of η from p_0 and p_1, A_1' the positively oriented arc from p_1 to p_2 and A_2' the positively oriented arc from p_2 to p_0.

Let $\mathcal{A}$ be the event that the path in the tree $T(\gamma) = \alpha_{\ell+1}(\gamma)$ from v_0 to α hits A_1'. We will now estimate both sides of the identity

$$\mathbf{P}[\mathcal{A}] = \mathbf{E}[\mathbf{P}[\mathcal{A} | D_m]] \tag{4.3}$$

using Proposition 4.2. By Wilson's algorithm, $\mathbf{P}[\mathcal{A}]$ is the probability that a simple random walk on the graph $H(D)$ started at v_0 stopped on hitting α will cross A_1'. This is exactly $\hat{h}(v_0)$, where the function $\hat{h}$ is as defined in Proposition 4.2. Set

$$h(z) := \frac{1}{\pi} \cot^{-1}\left(\frac{1 - |z|}{2\,\mathrm{Im}\,\sqrt{z}}\right) = \frac{1}{\pi} \cot^{-1}\left(\frac{1 - r}{2\sqrt{r}\sin(\theta/2)}\right),$$

where $z = re^{i\theta}$ and we take the value of $\cot^{-1}$ between 0 and π. Note that h is harmonic in $\mathbb{H}$, is equal to 0 on $(0, 1)$, is equal to 1 on $(1, \infty)$, and $\partial_y h = 0$ on $(-\infty, 0)$. (Of course, we found the map h satisfying these boundary conditions by reflecting the domain along the negative real axis, mapping this larger domain to $\mathbb{H}$

with $z \mapsto \sqrt{z}$ and then using a conformal map from $\mathbb{H}$ to $\mathbb{U}$ to calculate the hitting probabilities.) Consequently, Proposition 4.2 shows that if r_0 is sufficiently large, then

$$\mathbf{P}[\mathcal{A}] = h\big(\phi(v_0)/U\big) + O(\delta^3). \tag{4.4}$$

Set $V_j := \phi_j(v_0)$ and $U_j := \phi_j \circ \phi_0^{-1}(U)$. By the chordal version of Loewner's equation and the definition of m, we have

$$V_m = V_0 + \frac{2t_m}{V_0} + O(\delta^3), \qquad U_m = U_0 + \frac{2\,t_m}{U_0} + O(\delta^3). \tag{4.5}$$

Note also that $\mathrm{rad}_0(D_m) > \mathrm{rad}_0(D)/2$, $v_0 \in D_m$ and $U_m \in [1/4, 4]$ provided that δ is small enough.

We now employ a similar argument to estimate $\mathbf{P}[\mathcal{A}|D_m]$. Recall that $\tilde{D}_n = D(\alpha_n, \beta_n, \gamma(n), b)$. Assume that Q_0 intersects $\tilde{D}_n$, which will be the case if $U_n > 1/4$, say. Let η_n be the set of points in $\tilde{D}_n$ at distance $1/10$ from $\partial\tilde{D}_n$. Again, η_n is a simple closed path, and we write $\eta_n = A_0'(n) \cup A_1'(n) \cup A_2'(n)$, where $A_0'(n)$ is the arc of η_n from the closest point to $\gamma(n)$ to the point of intersection of Q_0 with η_n, $A_1'(n)$ is the arc of η_n from the point in $Q_0 \cap \eta_n$ to the point of η_n closest to b and $A_2'(n)$ is the remaining part of η_n. By Lemma 4.1, $\mathbf{P}[\mathcal{A}|D_n]$ is the same as the quantity $\hat{h}_n(v_0)$, where $\hat{h}_n$ is the function $\hat{h}$ defined in Proposition 4.2, but with $A_0'(n)$, $A_1'(n)$, $A_2'(n)$ and η_n replacing A_0', A_1', A_2' and η and D_n replacing D. (It is D_n replacing D, not $\tilde{D}_n$. The conditions of Proposition 4.2 hold for either of these, but the conformal map we consider is defined on D_n.) Proposition 4.2 therefore gives

$$\mathbf{P}[\mathcal{A}|D_m] = h\left(\frac{V_m - W_m}{U_m - W_m}\right) + O(\delta^3). \tag{4.6}$$

Write $f(U, V, W) := h((V - W)/(U - W))$. We Taylor-expand the right-hand side in (4.6) to second order in W_m and to first order in $V_m - V_0$ and $U_m - U_0$. Together with (4.3)–(4.5) this gives

$$\begin{aligned} 0 &= \mathbf{E}\big[\mathbf{P}[\mathcal{A}|D_m]\big] - \mathbf{P}[\mathcal{A}] \\ &= \frac{1}{2}\partial_W^2 f\,\mathbf{E}[W_m^2] + \partial_W f\,\mathbf{E}[W_m] + \partial_V f\,\frac{2\mathbf{E}[t_m]}{V_0} + \partial_U f\,\frac{2\mathbf{E}[t_m]}{U_0} + O(\delta^3). \end{aligned}$$

Here, the derivatives of f are evaluated at (V_0, W_0, U_0). (Note that V is complex valued, and we interpret $\partial_V f$ as an $\mathbb{R}$-linear map from $\mathbb{C}$ to $\mathbb{R}$.) If we plug in $V_0 = i + O(\delta^3)$ and $U_0 = 1 + O(\delta^3)$ [as we have seen in Section 3.2, one can certainly find v_0 and u_0 satisfying $\phi(v_0) = i + O(\delta^3)$ and $\phi(u_0) = 1 + O(\delta^3)$ if r_0 is large], then after some tedious but straightforward computations the above equality simplifies to

$$\mathbf{E}[W_m^2] + 2\mathbf{E}[W_m] - 8\mathbf{E}[t_m] = O(\delta^3),$$

while $V_0 = 2i + O(\delta^3)$ and $U_0 = 1 + O(\delta^3)$ give

$$3\mathbf{E}[W_m^2] + 8\mathbf{E}[W_m] - 24\mathbf{E}[t_m] = O\delta^3).$$

Combining these two relations implies (4.1) and (4.2) in the case $k = 0$. For $k > 0$, the proof is basically the same; the only essential difference is that one must use η_k in place of η. □

THEOREM 4.4 (Driving process convergence). *For every positive* ϵ_1,ϵ_2, ϵ_3 *and* $\bar{t}$, *there is some positive* $r_1 = r_1(\epsilon_1, \epsilon_2, \epsilon_3, \bar{t})$ *such that the following holds. Let* $D = D(\alpha, \beta, a, b) \in \mathfrak{D}^*$ *satisfy* $\mathrm{rad}_0(D) > r_1$ *and* $\mathfrak{H}_D(0, \alpha) \in [\epsilon_1, 1 - \epsilon_1]$. *Let* γ *be the corresponding UST Peano path, let* $\phi : D \to \mathbb{H}$ *denote the conformal map which takes a to* 0, b *to* ∞ *and satisfies* $|\phi(0)| = 1$, *let* $\hat{\gamma} := \phi \circ \gamma$, *parameterized according to capacity from* ∞, *and let* $W(t)$ *denote the Loewner driving process for* $\hat{\gamma}$. *Then there is a coupling of standard Brownian motion* $B : [0, \bar{t}\,] \to \mathbb{R}$ *and* W *such that*

$$\mathbf{P}\big[\sup\{|W(t) - B(8t)| : t \in [0, \bar{t}\,]\} > \epsilon_2\big] < \epsilon_3.$$

PROOF. The proof is almost identical to the proof of Theorem 3.7, where we used Skorohod's embedding, but one has to be a little careful because it may happen that 0 is "swallowed" before time $\bar{t}$.

Let us first assume that $\phi(0)$ is close to i, say $|\phi(0) - i| < 1/100$, and that $\bar{t}$ is small enough so that

$$\bar{t} \le 1/100 \quad \text{and} \quad \mathbf{P}\big[B[0, \bar{t}\,] \subset [-1/10, 1/10]\big] > 1 - \epsilon_3/3, \tag{4.7}$$

where B is standard Brownian motion. Take $\delta = \delta(\epsilon_1, \epsilon_2, \epsilon_3) > 0$ small and $\epsilon = 1/10$. Define $r_0(\epsilon, \delta)$ as in Proposition 4.3. Let $k \in \mathbb{N}$ be the first integer where $\mathrm{rad}_0(D_k) \le r_0$ or $\mathfrak{H}_{D_k}(0, \alpha_k) \notin [\epsilon, 1 - \epsilon]$ and define $\bar{t}_0 := \min\{\bar{t}, t_k\}$, where t_n is as in the proposition. Exactly as in the proof of Theorem 3.7, Proposition 4.3 implies that we may couple W with a Brownian motion B in such a way that

$$\mathbf{P}\big[\sup\{|W(t) - B(8t)| : t \in [0, \bar{t}_0]\} > \epsilon_2/3\big] < \epsilon_3/3,$$

if $\mathrm{rad}_0(D) \ge r_1$ and r_1 is large enough. By our assumptions regarding $\bar{t}$, we have with high probability that, for all $t \in [0, \bar{t}_0]$, $W(t) \in [-1/5, 1/5]$. If we choose r_1 large enough, this guarantees that $\mathbf{P}[\bar{t}_0 \ne \bar{t}\,] < \epsilon_3/3$ and proves the theorem when (4.7) is satisfied and $\phi(0)$ is close to i.

Consider now a general $\bar{t} < \infty$. Let $\bar{t}_1 > 0$ be some constant satisfying (4.7) and let $z_0 := \phi^{-1}(2i\bar{t}^2/\bar{t}_1^2)$. From the Koebe distortion theorem, it follows that there is a constant $c = c(\bar{t}, \epsilon_1)$ such that $\mathrm{rad}_{z_0}(D) \ge c\,\mathrm{rad}_0(D)$. (See, e.g., Theorem 1.3 and Corollary 1.4 in [35].) Consequently, by choosing r_1 appropriately larger, we may invoke the above argument with the basepoint moved from 0 to a vertex near z_0 and with a smaller ϵ_2. Rescaling now completes the proof of the theorem. □

4.3. *Uniform continuity.* In order to prove convergence with respect to a stronger topology, tightness will be needed, and we therefore derive in this section some regularity estimates for UST Peano curves with respect to the capacity parameterization. Some results from [38] will be used.

Let $D \subset \mathbb{C}$ be a simply connected domain containing 0, whose boundary is a C^1-simple closed path. Let a and b be two distinct points on ∂D. In this section, we consider for large R the UST Peano curve from a point near Ra to a point near Rb on a grid approximation of RD. One reason not to consider arbitrary domains is that we need to partially adapt to the framework of [38] in order to quote results from there. Also, it is natural (since the UST Peano curve is asymptotically space filling) to impose regularity conditions on ∂D in order to get uniform regularity estimates for the UST Peano curve.

Let α_D and β_D be, respectively, the clockwise and anticlockwise arcs of ∂D from b to a. Given R large, let $D^R = D(\alpha^R, \beta^R, a^R, b^R) \in \mathfrak{D}^*$ be an approximation of $(RD, R\alpha_D, R\beta_D)$ in the following sense. Fix some sufficiently large constant $C > 0$; for example, $C = 10$ would do. We require α^R to be a simple path in $\mathbb{Z}^2$ satisfying $\rho(\alpha^R, R\alpha_D) \le C$ and require β^R to be a simple path in the dual grid $(1/2 + \mathbb{Z})^2$ satisfying $\rho(\beta^R, R\beta_D) \le C$. We also require that $\beta^R \cap \alpha^R = \varnothing$, of course, and that each of a^R, b^R is a Peano vertex adjacent to an endpoint of α^R and an endpoint of β^R.

Let $\gamma = \gamma^R$ be the UST Peano path in D^R. Let $\phi : D \to \mathbb{H}$ be the conformal homeomorphism satisfying $\phi(a) = 0$, $\phi(b) = \infty$ and $|\phi(0)| = 1$. Let $\phi_R : D^R \to \mathbb{H}$ be the conformal homeomorphism satisfying $|\phi_R(0)| = 1$, taking a^R to 0 and b^R to ∞. Then $\lim_{R\to\infty} R^{-1}\phi_R^{-1}(z) = \phi^{-1}(z)$ uniformly in $\mathbb{H}$. (This follows, e.g., from Corollary 2.4 in [35].) Let $\hat{\gamma} := \phi_R \circ \gamma$, parameterized according to capacity from ∞. Let $g_t : \mathbb{H} \setminus \hat{\gamma}[0, t] \to \mathbb{H}$ be the conformal map with the usual normalization $g_t(z) - z \to 0$ when $|z| \to \infty$.

PROPOSITION 4.5 (Uniform continuity estimate). *For every $\epsilon > 0$ and $\bar{t} > 0$, there are some positive $R_0 = R_0(D, \bar{t}, \epsilon)$ and $\delta = \delta(D, \bar{t}, \epsilon)$ such that, for all $R > R_0$,*

$$\mathbf{P}\big[\sup\{|\hat{\gamma}(t_2) - \hat{\gamma}(t_1)| : t_1, t_2 \in [0, \bar{t}\,], |t_2 - t_1| \le \delta\} > \epsilon\big] < \epsilon.$$

We first prove a slightly modified version of this proposition.

LEMMA 4.6. *For $0 < t_1 < t_2 < \infty$, let $Y(t_1, t_2) := \operatorname{diam}(g_{t_1} \circ \hat{\gamma}[t_1, t_2])$. For every $\epsilon > 0$, there is a $\delta = \delta(D, \epsilon) > 0$ and an $R_0 = R_0(D, \epsilon) > 0$ such that, for all $R \ge R_0$,*

$$\mathbf{P}\big[\sup\{|\hat{\gamma}(t_2) - \hat{\gamma}(t_1)| : 0 \le t_1 \le t_2 \le \tau, Y(t_1, t_2) \le \delta\} \ge \epsilon\big] < \epsilon, \tag{4.8}$$

where $\tau := \inf\{t \ge 0 : |\hat{\gamma}(t)| = \epsilon^{-1}\}$.

The proof will use Theorems 10.7 and 11.1(ii) of [38]. As explained there, the proofs of these theorems are now easier, because we have established the conformal invariance of the UST; see Corollary 1.2.

PROOF OF LEMMA 4.6. Let δ_R be a positive function of R such that $\lim_{R\to\infty}\delta_R = 0$. It suffices to show that (4.8) holds for all sufficiently large R with δ_R in place of δ. Let Z denote the semicircle $2\epsilon^{-1}\partial\mathbb{U}\cap\mathbb{H}$, say. For R large, let t_1 and t_2 be such that $|\hat\gamma(t_2)-\hat\gamma(t_1)|$ is maximal subject to the constraints $0\le t_1\le t_2\le\tau$ and $Y(t_1,t_2)\le\delta_R$. Note that $\min_{t\le\tau}\operatorname{dist}(g_t(Z), g_t\circ\hat\gamma[t,\tau])$ is bounded from below, as $\hat\gamma[0,\tau]\subset\epsilon^{-1}\overline{\mathbb{U}}$. [Considering the harmonic measure from a point near ∞, one deduces that the diameter of $g_t((1/2)Z)$ is bounded below. The extremal distance between $g_t(Z)$ and $g_t((1/2)Z)$ is the same as the extremal distance between Z and $(1/2)Z$. This shows that $\operatorname{dist}(g_t(Z), g_t\circ\hat\gamma[t,\tau])\ge\operatorname{dist}(g_t(Z), g_t((1/2)Z))$ is bounded from below.] Since $Y(t_1,t_2)\le\delta_R\to 0$ as $R\to\infty$, the extremal length of the collection of simple arcs in $\mathbb{H}\setminus g_{t_1}\circ\hat\gamma[t_1,t_2]$ which separate $g_{t_1}\circ\hat\gamma[t_1,t_2]$ from $g_{t_1}(Z)$ goes to 0. By the conformal invariance of the extremal length, it follows that the extremal length of the collection of simple arcs in $\mathbb{H}\setminus\hat\gamma[t_1,t_2]$ which separate $\hat\gamma[t_1,t_2]$ from Z in $\mathbb{H}\setminus\hat\gamma[0,t_1]$ tends to 0 as well. In particular, the shortest such arc for the Euclidean metric, say η, satisfies $\lim_{R\to\infty}\operatorname{length}(\eta)=0$.

We are going to study separately the three cases where η is close to the origin, close to the real line but not to the origin and not close to the real line. In each case, we will see that the existence of such an η is very unlikely. Let $\mathcal{A}$ be the event $|\hat\gamma(t_1)-\hat\gamma(t_2)|\ge\epsilon$. For $s>0$, let $\mathcal{X}_0(s)$ be the event $\operatorname{dist}(0,\eta)<s$ and let $\mathcal{X}_1(s)$ be the event $\operatorname{dist}(\mathbb{R},\eta)<s$. We will prove

$$\forall s_1>0,\ \exists R_0>0,\ \forall R>R_0,\qquad \mathbf{P}[\mathcal{A}\setminus\mathcal{X}_1(s_1)]<\epsilon, \tag{4.9}$$

$$\forall s_0>0,\ \exists s_1>0,\ \exists R_0>0,\ \forall R>R_0,\qquad \mathbf{P}[\mathcal{A}\cap\mathcal{X}_1(s_1)\setminus\mathcal{X}_0(s_0)]<\epsilon, \tag{4.10}$$

$$\exists s_0>0,\ \exists R_0>0,\ \forall R>R_0,\qquad \mathbf{P}[\mathcal{A}\cap\mathcal{X}_0(s_0)]<\epsilon. \tag{4.11}$$

Using these statements, the proof of the lemma is completed by choosing s_0 according to (4.11), then choosing s_1 according to (4.10) and, finally, choosing R_0 according to (4.9), (4.10) and (4.11).

We start with (4.9). Fix some $s_1>0$ and assume that $\mathcal{A}\setminus\mathcal{X}_1(s_1)$ holds. We also assume that $\epsilon<s_1$. There is no loss of generality in that assumption, since $\mathcal{A}$ is monotone decreasing in ϵ. Since $\lim_{R\to\infty}\operatorname{length}(\eta)=0$, for large R the two endpoints of η must be in $\hat\gamma[0,t_1]$. Because $\hat\gamma$ tends to ∞ with t, it is clear that $\hat\gamma[t_2,\infty)\cap\eta\ne\varnothing$. In fact, the crossing number of $\hat\gamma[t_2,\infty)$ and η must be ±1, since $\hat\gamma$ and η are simple curves. Consider the concentric annulus A whose inner circle is the smallest circle surrounding η and whose outer circle has radius $\epsilon/4$. Let $\mathcal{B}$ denote the open disk bounded by the outer circle of A and note that $\mathcal{B}\subset\mathbb{H}$, by our assumption $\epsilon<s_1$. On the event $\mathcal{A}$, there is a $t^*\in[t_1,t_2]$ such

that the distance from $\hat{\gamma}(t^*)$ to η is at least $\epsilon/2$. In particular, $\hat{\gamma}(t^*) \notin \mathcal{B}$. Now, η separates $\hat{\gamma}(t^*)$ from ∞ in $\mathbb{H} \setminus \hat{\gamma}[0, t_1]$. Therefore, if $\mathcal{A} \setminus X_1(s)$ holds, then $\hat{\gamma}[0, t_1] \cup \eta$ separates $\hat{\gamma}(t^*)$ from ∞. Since $\hat{\gamma}$ is a simple path, this implies that the arc of $\hat{\gamma}[0, t_1]$ between the two points $\overline{\eta} \cap \hat{\gamma}[0, t_1]$ does not stay in $\mathcal{B}$. Hence, $\hat{\gamma}[0, \infty) \cap \mathcal{B}$ has three distinct connected components, say $\hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3$, each of which intersects the inner circle of A such that $\hat{\gamma}_1, \hat{\gamma}_3 \subset \hat{\gamma}[0, t^*]$ and $\hat{\gamma}_2 \subset \hat{\gamma}[t^*, \infty)$ and $\hat{\gamma}_2$ separates $\hat{\gamma}_1$ from $\hat{\gamma}_3$ within $\mathcal{B}$. See Figure 8.

Note that adjacent to one side of $\phi_R^{-1}(\hat{\gamma}_2)$ lies T, the UST, and $T^\dagger$, the dual UST, is adjacent to the other side. Both are connected, and they do not intersect $\phi_R^{-1}(\hat{\gamma})$. It follows that there are paths $\chi_1 \subset T$ and $\chi_2 \subset T^\dagger$ with endpoints in $\phi_R^{-1}(\partial\mathcal{B})$, each of which intersects the inner boundary of $\phi_R^{-1}(A)$. But the diameter of the inner boundary of $R^{-1}\phi_R^{-1}(A)$ goes to 0 as $R \to \infty$, and the distance between the two boundary components of $R^{-1}\phi_R^{-1}(A)$ does not. Hence, by [38], Theorem 10.7, the probability that such a configuration appears somewhere goes to 0 with R. (Although the result from [38] refers to the UST in the whole plane, the proof is local, and since we are bounded away from the boundary, the result is applicable here.) This proves (4.9).

Now fix $s_0 > 0$ and let $s_1 > 0$ be much smaller. Assume that $\mathcal{A} \cap \mathcal{X}_1(s_1) \setminus \mathcal{X}_0(s_0)$ holds, $\epsilon < s_1$ and R is large. Also assume that η is closer to $[0, \infty)$ than to $(-\infty, 0]$. Note that η is then bounded away from $(-\infty, 0]$. Let A be defined as above and let $\mathcal{B}$ be the intersection of $\mathbb{H}$ with the disk bounded by the outer boundary component of A. We now need to consider two distinct possibilities. Either both endpoints of η are on $\hat{\gamma}[0, t_1]$, and then the configuration is topologically as in the argument for (4.9), or one endpoint of η is on $[0, \infty)$. But it is easy to see that in either case there is a simple path in $T^\dagger$ which intersects $\phi_R^{-1}(\eta)$ whose endpoints are in $\phi_R^{-1}(\partial\mathcal{B})$, by an argument very similar to the one given above. Now [38], Theorem 11.1(ii), shows that these events have small probabilities if s_1 is small. The case where η is closer to $(-\infty, 0]$ is treated similarly, with the roles of the tree and the dual tree switched. Thus, (4.10) is established.

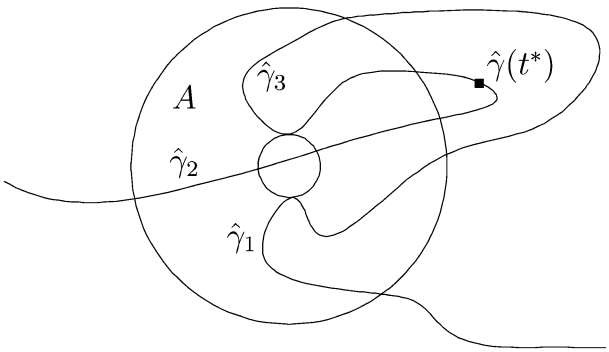


FIG. 8. *The paths $\hat{\gamma}_1, \hat{\gamma}_2$ and $\hat{\gamma}_3$ and the annulus A.*

To prove (4.11), let $s > 0$ and let $v \in \mathbb{Z}^2$ be a vertex closest to $\phi_R^{-1}(is)$. Let $v^\dagger \in (\mathbb{Z} + 1/2)^2$ be a dual vertex adjacent to v. Let χ be the simple path from v to α^R in T and let $\chi^\dagger$ be the simple path from $v^\dagger$ to β^R in $T^\dagger$. We may sample χ by running a simple random walk from v on $H(D^R)$ stopped on hitting α^R and loop-erasing it. It therefore follows by Proposition 4.2 that if s is sufficiently small, then the diameter of $\phi_R(\chi)$ is smaller than $\epsilon/10$ with probability at least $1 - \epsilon/10$. Moreover, there is some $s_0 > 0$ such that for all sufficiently large R with probability at least $1 - \epsilon/10$ the distance from $\phi_R(\chi)$ to 0 is at least s_0, and the same two estimates will hold for $\chi^\dagger$. Let D' be the domain bounded by $[v, v^\dagger] \cup \chi \cup \chi^\dagger \cup \partial D^R$ which has the initial point a^R of γ on its boundary. Note that γ crosses the boundary of D' exactly once, through the segment $[v, v^\dagger]$. In particular, if $\operatorname{diam}(\phi_R(D')) < \epsilon$ and $\mathcal{A}$ holds, then η is not contained in $\phi_R(D')$. This proves (4.11) and completes the proof of the lemma. □

PROOF OF PROPOSITION 4.5. Theorem 4.4 implies that we may find some $r > 0$ such that $\mathbf{P}[\sup\{|W(t)| : t \in [0, \bar{t}\,]\} \geq r] < \epsilon/4$ for all sufficiently large R. Let $\epsilon' := \min\{\epsilon, r^{-1}\}$ and let δ' denote the δ obtained by using Lemma 4.6 with ϵ' in place of ϵ. Since Brownian motion is a.s. continuous, Theorem 4.4 implies that there is some $\delta > 0$ such that if R is large enough we have

$$\mathbf{P}\big[\sup\{|W(t_1) - W(t_2)| : t_1, t_2 \in [0, \bar{t}\,], |t_1 - t_2| < \delta\} \geq \delta'\big] < \epsilon/4.$$

Lemma 2.1 applied to the path $t \mapsto g_{t_1} \circ \gamma(t - t_1) - W(t_1)$ now implies

$$\mathbf{P}\big[\sup\{Y(t_1, t_2) : 0 \leq t_1 \leq t_2 \leq \bar{t}, |t_1 - t_2| < \delta\} \geq C(\delta^{1/2} + \delta')\big] < \epsilon/4.$$

Now the proof is completed by using Lemma 4.6. □

4.4. *Consequences.* In this section we gather some consequences, starting with the following two theorems.

THEOREM 4.7 (Chordal SLE_8 traces a path). *Let $\tilde{g}_t$ denote the chordal* SLE_8 *process driven by $B(8t)$, where $B(t)$ is standard Brownian motion. Then, a.s. for every $t > 0$, the map $\tilde{g}_t^{-1}$ extends continuously to $\overline{\mathbb{H}}$ and $\tilde{\gamma}(t) := \tilde{g}_t^{-1}(B(8t))$ is a.s. continuous. Moreover, a.s. $\tilde{g}_t^{-1}(\mathbb{H})$ is the unbounded component of $\mathbb{H} \setminus \tilde{\gamma}[0, t]$ for every $t \geq 0$.*

THEOREM 4.8 (Peano path convergence). *Let $D \subset \mathbb{C}$ be a domain containing 0 such that ∂D is a C^1-smooth simple closed path. Let $\partial D = \alpha_D \cup \beta_D$ be a partition of the boundary of D into two nontrivial complementary arcs. For $R > 0$, let (D^R, α^R, β^R) be an approximation of $(RD, R\alpha, R\beta)$, as described in Section* 4.3. *Let $\gamma = \gamma^R$ denote the UST Peano curve in D^R with the corresponding boundary conditions. Let $\phi_R : D^R \to \mathbb{H}$ denote the conformal map which takes the initial point of γ to 0, the terminal point to ∞ and satisfies $|\phi_R(0)| = 1$. Let $\hat{\gamma} := \phi_R \circ \gamma$, parameterized by capacity from ∞. Then the law of $\hat{\gamma}$ tends weakly to the law of $\tilde{\gamma}$ from Theorem* 4.7.

Here, we think of $\hat{\gamma}$ and $\tilde{\gamma}$ as elements of the space of continuous maps from $[0, \infty)$ to $\overline{\mathbb{H}}$, with the topology of locally uniform convergence.

A consequence of the theorem is that $R^{-1}\gamma$ is close to $R^{-1}\phi_R^{-1} \circ \tilde{\gamma}$. That is, we may approximate the UST Peano path γ by the image of chordal SLE_8 in D^R.

The analog of Theorem 4.7 was proven in [36] for all $\kappa \neq 8$, but the particular case $\kappa = 8$ could not be handled there. It is fortunate that the convergence of the UST Peano path to SLE_8 settles this problem. By Remark 7.5 in [36], it follows that with the notation of Theorem 4.7 for $\kappa \geq 8$ we have $\tilde{g}_t^{-1}(\mathbb{H}) = \mathbb{H} \setminus \tilde{\gamma}[0, t]$ for every $t \geq 0$ a.s.

PROOF OF THEOREMS 4.7 AND 4.8. Let $W(t) = W_R(t)$ denote the chordal Loewner driving process for $\hat{\gamma}$. Fix a sequence $R_n \to \infty$. First, note that the family of laws of $\hat{\gamma}$ is tight, because of Proposition 4.5 and the Arzela–Ascoli theorem (see, for instance, [17], Theorem 2.4.10). Also, Theorem 4.4 implies that the law of W converges weakly to the law of $B(8t)$. Hence, there is a subsequence of R_n such that the law of the pair $(\hat{\gamma}, W)$ converges weakly to some probability measure μ. Let (γ^*, W^*) be random with law μ. Then we may identify $W^*(t)$ with $B(8t)$. By the chordal analog of Lemma 3.14, which is valid with the same proof, it follows that, for all $t > 0$, $\tilde{g}_t^{-1}(\mathbb{H})$ is the unbounded component of $\mathbb{H} \setminus \gamma^*[0, t]$. Since γ^* is continuous, elementary properties of conformal maps imply that $\tilde{g}_t^{-1}$ extends continuously to $\overline{\mathbb{H}}$ (e.g., Theorem 2.1 in [35]). It is easy to verify that, a.s. for every $t > 0$, $\tilde{g}_t^{-1}(B(8t)) = \gamma^*(t)$, using the fact that $\gamma^*[t, t']$ is contained in a small neighborhood of $\gamma^*(t)$ when $t' - t > 0$ is small. This proves Theorem 4.7. Because the law of the limit path γ^* does not depend on the subsequence, the original sequence converges, and so Theorem 4.8 is proved as well. □

We now list some easy consequences of Theorems 4.7 and 4.8.

COROLLARY 4.9 (Radial SLE_8 traces a path). *Let $\tilde{g}_t$ denote a radial* SLE_8 *process driven by $W(t) = \exp(iB(8t))$, where B is standard Brownian motion. Then, almost surely, for every $t > 0$, the map $\tilde{g}_t^{-1}$ extends continuously to $\overline{\mathbb{U}}$. Moreover, $\tilde{g}_t^{-1}(W(t))$ is almost surely continuous.*

PROOF. This follows readily from Theorem 4.7 and the absolute continuity relation between radial and chordal SLE_8 derived in [28], Proposition 4.2. □

PROOF OF THEOREM 1.3. Define $G_\delta = \delta D^{1/\delta}$, where $D^{1/\delta}$ is defined as in Theorem 4.8. Consider the situation of Theorem 4.8. As previously remarked, it follows from [35], Corollary 2.4, that $\lim_{R\to\infty} R^{-1}\phi_R^{-1}(z) = \phi^{-1}(z)$ uniformly in $\mathbb{H}$. Consequently, Theorem 4.8 shows that, for all $\bar{t} > 0$, the UST Peano curve scaling limit up to capacity $\bar{t}$ from b is equal to $\phi^{-1} \circ \tilde{\gamma}$ up to time $\bar{t}$. It therefore suffices to prove that for all $\epsilon > 0$ there is an $\epsilon' > 0$ such that for all sufficiently

large R with probability at least $1-\epsilon$ the part of γ after the first time it hits the ϵ'-neighborhood of b stays within the ϵ-neighborhood of b. This is easily proved by the same argument used to prove (4.11) applied to the reversal of the UST Peano path, which is also a UST Peano path. □

COROLLARY 4.10 (Path reversal). *The law of the chordal* SLE_8 *curve is invariant under simultaneously reversing time and inverting in the unit circle, up to a monotone increasing time change. More precisely, if* $\tilde{\gamma}$ *is the chordal* SLE_8 *curve from* 0 *to* ∞ *defined in Theorem* 4.7, *then a time change of* $(-1/\tilde{\gamma}(1/t), t \geq 0)$ *has the same law as* $\tilde{\gamma}$.

PROOF. This follows immediately from the fact that the reversed UST Peano curve is also a UST Peano curve. □

5. Random walk estimates. The goal of this section is to prove the remaining random walk estimates and thereby complete the proofs of the theorems. Basically, we show that, under certain boundary conditions, discrete harmonic functions converge to continuous harmonic functions satisfying corresponding boundary conditions, as the mesh of the grid goes to 0. This general principle is not new, of course (see, e.g., [6]), but it seems that the precise statements which are needed here do not appear in the literature. In particular, our results make no smoothness assumptions on the boundary. It should perhaps be noted that some of the following proofs (and most likely the results, too) are special to two dimensions.

5.1. *Preliminary lemmas.* We now state some lemmas on discrete harmonic functions, which will be helpful in the proofs of Proposition 2.2, Lemma 3.5 and Proposition 4.2.

For $\delta > 0$, define the discrete derivatives

$$\partial_x^\delta f(v) := \delta^{-1}\big(f(v+\delta) - f(v)\big),$$
$$\partial_y^\delta f(v) := \delta^{-1}\big(f(v+i\delta) - f(v)\big).$$

Let $\mathfrak{D}_\delta := \{\delta D : D \in \mathfrak{D}\}$, that is, domains adapted to the grid $\delta\mathbb{Z}^2$. Similarly, for $D_\delta = \delta D \in \mathfrak{D}_\delta$, define $V^\delta(D_\delta) := D_\delta \cap \delta\mathbb{Z}^2 = \delta V(D)$ and $V_\partial^\delta(D_\delta) := \delta V_\partial(D)$.

LEMMA 5.1 (Discrete derivative estimate). *There is a constant* $C > 0$ *such that, for every* $D \in \mathfrak{D}$ *and every bounded function* $h : V(D) \cup V_\partial(D) \to \mathbb{R}$ *that is harmonic in* $V(D)$,

$$\partial_x^1 h(0) \leq C\,\mathrm{rad}_0(D)^{-1}\|h\|_\infty, \qquad \partial_y^1 h(0) \leq C\,\mathrm{rad}_0(D)^{-1}\|h\|_\infty. \tag{5.1}$$

This lemma is proved using the Green's functions in [25], Theorem 1.7.1; see also [16], Lemma 7.1, for a proof of the analogous statement in the triangular lattice using the maximum principle. In Section 6, we rewrite and adapt the proof from [25] to more general walks on planar lattices. One can also rather easily prove the lemma using coupling.

LEMMA 5.2. *For all $\epsilon > 0$ and $k \in \mathbb{N}$, there exists a $c = c_k(\epsilon) > 0$ such that the following always holds. Let $\delta \in (0, c^{-1})$ and let $D \in \mathfrak{D}_\delta$ satisfy* $\mathrm{rad}_0(D) \geq 1/2$. *Let $\partial^\delta_{a_1}, \ldots, \partial^\delta_{a_k} \in \{\partial^\delta_x, \partial^\delta_y\}$. Let $h : V^\delta(D) \cup V^\delta_\partial(D) \to [0, \infty)$ be nonnegative and harmonic in $V^\delta(D)$. If $v \in V^\delta(D)$ satisfies $|\psi_D(v)| \leq 1 - \epsilon$, then*

$$|\partial^\delta_{a_1} \partial^\delta_{a_2} \cdots \partial^\delta_{a_k} h(v)| \leq c h(0). \tag{5.2}$$

Note that the case $k = 0$, which is included, is a kind of Harnack inequality.

It is easy to give quantitative estimates for $c_k(\epsilon)$, but they will not be needed here. Only $k \leq 3$ will be used in the rest of the paper.

In the proof of the lemma, the following simple conformal geometry consequences of the Koebe distortion theorem (see [35], Theorem 1.3) will be needed. Let D, ϵ and v be as in the statement of the lemma. First, note that $1/4 \leq \mathrm{rad}_0(D)\psi'_D(0) \leq 1$ follows from the Koebe $1/4$ theorem and the Schwarz lemma, respectively. Let $\ell = \ell(\epsilon)$ be large and set $z_j := j\,\psi_D(v)/\ell$ and $w_j := \psi_D^{-1}(z_j)$, $j = 0, 1, \ldots, \ell$. The Koebe distortion theorem gives upper and lower bounds for $\mathrm{rad}_0(D)\,|\psi'_D|$ on the preimage of the line segment $[0, z_\ell]$. This implies that there is a constant $c_1 = c_1(\epsilon) > 0$ such that $\mathrm{rad}_{w_j}(D) \geq c_1 \mathrm{rad}_0(D)$ and that if $\ell = \ell(\epsilon)$ is large, then, $|w_j - w_{j-1}| \leq c_1 \mathrm{rad}_0(D)/20$, $j = 1, \ldots, \ell$. In particular, if v_j is the vertex in $V^\delta(D)$ closest to w_j, then, provided that δ is sufficiently small, $|v_j - v_{j-1}| \leq \mathrm{rad}_{v_{j-1}}(D)/10$.

PROOF OF LEMMA 5.2. We start with $k = 0$. Suppose first that $|v| \leq \mathrm{rad}_0(D)/10$. Let $W \subset V^\delta(D)$ be the set of vertices w satisfying $h(w) \geq h(v)$. Then W contains a path from v to ∂D. But the probability p that the path traced by a simple random walk from 0 before exiting D separates v from ∂D is bounded away from 0. On that event, the simple random walk hits W before exiting D. Consequently, $h(0) \geq p h(v)$, as needed. For arbitrary $v \in V^\delta(D)$ satisfying $|\psi_D(v)| \leq 1 - \epsilon$, as we have noted, the Koebe distortion theorem implies that there is an $\ell = \ell(\epsilon)$ depending only on ϵ and a sequence $0 = v_0, v_1, \ldots, v_\ell = v$ in $V^\delta(D)$ with $\ell \leq \ell(\epsilon)$ such that $|v_j - v_{j-1}| \leq \mathrm{rad}_{v_j}(D)/10$ for each $j = 1, \ldots, \ell$. Consequently, iterating the above result gives $h(0) \geq p^\ell h(v)$ and proves the case $k = 0$.

Using the above, we know that $h(w) \leq c' h(0)$ on the set of vertices $w \in V^\delta(D)$ such that $|w - v| \leq \mathrm{rad}_v(D)/10$, where $c' = c'(\epsilon)$ is some constant depending only on ϵ. Consequently, the case $k = 1$ now follows from Lemma 5.1 applied with v translated to 0.

For $k > 1$, the proof is by induction. By the above, we may assume $v = 0$. Let M be the maximum of $|\partial^\delta_{a_k} h(w)|/h(0)$ on the set V of vertices $w \in V^\delta(D)$ satisfying $|w| \leq \mathrm{rad}_0(D)/10$. The above shows that M is bounded by a universal constant. Since $\partial^\delta_{a_k} h$ is discrete harmonic on V, the proof is completed by applying the inductive hypotheses to the function $\partial^\delta_{a_k} h(w) + M h(0)$. $\square$

LEMMA 5.3 (Continuous harmonic approximation). *For every $\epsilon > 0$, there is some $r_0 = r_0(\epsilon) > 0$ such that the following holds. If $D \in \mathfrak{D}$ satisfies $\operatorname{rad}_0(D) \geq r_0$ and $h : V(D) \cup V_\partial(D) \to [0, \infty)$ is discrete harmonic in $V(D)$, then there exists a harmonic function $h^* : D \to [0, \infty)$ such that*

$$|h^*(v) - h(v)| \leq \epsilon h(0) \tag{5.3}$$

holds for every vertex $v \in V(D)$ satisfying $|\psi_D(v)| < 1 - \epsilon$.

PROOF. Suppose that the lemma is not true. Then there exists $\epsilon > 0$ and a sequence of pairs (D_n, h_n), where $D_n \in \mathfrak{D}$ satisfies $\operatorname{rad}_0(D_n) \geq n$ and $h_n > 0$ is discrete harmonic in $V(D_n)$, satisfies $h_n(0) = 1$, but (5.3) fails for every harmonic function $\hat{h}$.

Set $\delta = \delta_n := 1/\operatorname{rad}_0(D_n)$. Our objective is to apply compactness to show that the maps $h_n \circ \psi_D^{-1}$ converge locally uniformly in $\mathbb{U}$ as $n \to \infty$ along some subsequence to some harmonic $\hat{h}$, so that (5.3) does hold for some n. We put $h^n(v) := h_n(v/\delta_n)$.

First, standard compactness properties of conformal maps say that one can take a subsequence such that the maps $\delta_n \psi_{D_n}^{-1}$ converge locally uniformly in $\mathbb{U}$ to some conformal map, say ϕ. (This follows, e.g., from the Arzela–Ascoli theorem, together with [35], Corollary 1.4, with $z = 0$ and part two of [35], Theorem 1.3.) If $K \subset \mathbb{U}$ is compact, then Lemma 5.2 shows that there is a constant $C > 0$ such that, for all sufficiently large n in the subsequence, the discrete derivatives $|\partial_x^\delta h^n|$ and $|\partial_y^\delta h^n|$ are bounded by C in $\phi(K) \cap V(\delta D_n)$. By a variant of the Arzela–Ascoli theorem, it then follows that there is some continuous $h^* : \phi(\mathbb{U}) \to [0, \infty)$ and a further subsequence such that, for every compact $K \subset \phi(\mathbb{U})$,

$$\sup\{|h^n(v) - h^*(v)| : v \in K \cap \delta\mathbb{Z}^2\} \to 0$$

along the subsequence. The same argument may also be applied to prove the convergence of the discrete derivatives of h^n to arbitrary order, possibly in a further subsequence. Obviously, the discrete derivatives of h^n will converge to the corresponding continuous derivatives of h^*; that is,

$$\sup\bigl\{\bigl|\partial_{a_1}^\delta \cdots \partial_{a_k}^\delta h^n(v) - \partial_{a_1} \cdots \partial_{a_k} h^*(v)\bigr| : v \in K \cap \delta\mathbb{Z}^2\bigr\} \to 0, \tag{5.4}$$

where $\partial_{a_j}^\delta \in \{\partial_x^\delta, \partial_y^\delta\}$ and $\partial_{a_j} \in \{\partial_x, \partial_y\}$ is the corresponding continuous derivative, $j = 1, 2, \ldots, k$. The fact that h^n is discrete harmonic translates to $(\partial_x^\delta)^2 h^n(v - \delta) + (\partial_y^\delta)^2 h^n(v - i\delta) = 0$. Therefore, (5.4) shows that h^* is harmonic. This completes the proof. □

LEMMA 5.4 (Boundary hitting). *For every $\epsilon_1, \epsilon_2 > 0$, there is a $\delta = \delta(\epsilon_1, \epsilon_2) > 0$ such that if $D \in \mathfrak{D}$ and $w \in V(D)$ is a vertex satisfying $|\psi_D(w)| \geq 1 - \delta$, then the probability that the simple random walk started at w will hit*

$$\{v \in V(D) : |\psi_D(v) - \psi_D(w)| > \epsilon_1\}$$

before hitting ∂D is at most ϵ_2.

PROOF. We first prove the lemma in the case where ϵ_2 is very close to 1. Let $\delta > 0$ be much smaller than ϵ_1. Fix some vertex $w \in V(D)$ and suppose that $|\psi_D(w)| \geq 1 - \delta$. Let

$$\alpha := \{z \in D : |\psi_D(z) - \psi_D(w)| = \epsilon_1\}.$$

Let z_1 be a point in ∂D closest to w and set $r := \text{dist}(w, \partial D) = |z_1 - w|$. Let A_1 be the line segment $[w, z_1]$. Let Q be the connected component of $C(z_1, r) \cap D$ which contains w, where $C(z, r)$ denotes the circle of radius r and center z. Then Q is an arc of a circle. Let A_2 and A_3 denote the two connected components of $Q \setminus \{w\}$. See Figure 9. For $j = 1, 2, 3$, let K_j be the connected component of $D \setminus (A_1 \cup A_2 \cup A_3)$ which does not have A_j as a subset of its boundary.

Because δ is small compared to ϵ_1, the Koebe distortion theorem (e.g., Corollaries 1.4 and 1.5 in [35]) shows that $\alpha \cap C(w, r/8) = \varnothing$. For $j = 1, 2, 3$, let $\mathcal{S}_j$ be the collection of all paths which stay in K_j from the first time they hit $C(w, r/8)$ until they first exit from D. Let $B(t)$ denote Brownian motion started from w. It is easy to see that there is a universal constant $c_1 > 0$ such that $\mathbf{P}[B \in \mathcal{S}_j] > c_1$ for $j = 1, 2, 3$. For example, to prove this for $j = 3$, observe that the collection of Brownian paths which first hit $C(w, r/8)$ in K_3 and later hit A_3 before $A_1 \cup A_2$ has probability bounded away from 0.

Suppose for the moment that α intersects A_1 and A_2. Consider a subarc $\alpha' \subset \alpha$ whose endpoints are in A_1 and A_2, which is minimal with respect to inclusion. Then $\alpha' \subset \overline{K_3}$ or $\alpha' \subset \overline{K_1 \cup K_2}$. If $\alpha' \subset \overline{K_3}$, then α' separates $C(w, r/8)$ from ∂D in K_3. Consequently, on the event $B \in \mathcal{S}_j$, B hits α before hitting ∂D. However, by choosing δ to be sufficiently small and invoking the conformal invariance of the harmonic measure, we may ensure that the latter event has probability smaller than c_1. An entirely similar argument rules out the possibility that $\alpha' \subset \overline{K_1 \cup K_2}$. Similarly, it is not possible that α intersects both A_1 and A_3 or that α intersects

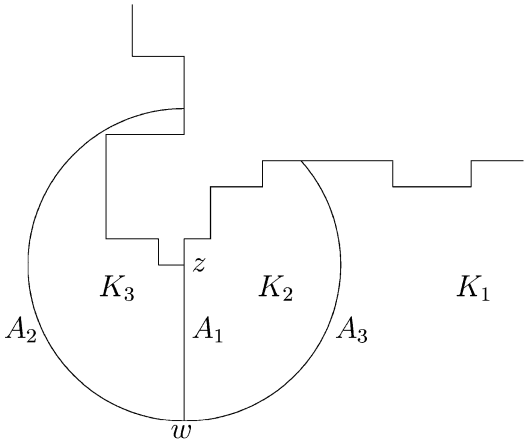


FIG. 9. *The arcs A_j and the components K_j.*

both A_2 and A_3. Hence, there is some $j \in \{1, 2, 3\}$ such that $\alpha \cap K_j = \varnothing$. Let j' be such a j.

By the convergence of a simple random walk to Brownian motion, it is clear that there is some universal constant $r_0 > 0$ such that if $r > r_0$, then the probability that the simple random walk started from w is in $\mathcal{S}_{j'}$ is at least $c_1/2$. This establishes the lemma in the case where $\epsilon_2 \in (1 - c_1/2, 1]$ and $r > r_0$. Suppose $r \le r_0$. Then there are two grid paths of bounded length starting from w to ∂D that are disjoint except at w. If α intersects both these paths, then this gives a lower bound for the continuous harmonic measure of α from w. Consequently, by making δ small enough, we can make sure that this does not happen. Thus, again, with probability bounded away from 0, the random walk from w hits ∂D before α, since it may follow any one of these two paths. This proves the lemma in the case where $\epsilon_2 \in (c_2, 1]$, where c_2 is some universal constant.

The Koebe distortion theorem implies that there is a constant $c > 0$ such that if $v_1, v_2 \in V(D)$ are neighbors, then $1 - |\psi_D(v_2)| \le c(1 - |\psi_D(v_1)|)$. (See, e.g., Corollaries 1.4 and 1.5 in [35].) Consequently, we may iterate the above restricted case of the lemma and use the Markov property, thereby proving the lemma for arbitrary $\epsilon_2 > 0$. □

5.2. *The hitting probability estimate.*

PROOF OF PROPOSITION 2.2. Let $\epsilon_1 > 0$ be much smaller than ϵ. We consider the discrete harmonic function $h(w) := H(w, u)/H(0, u)$. For $\delta > 0$, let

$$V(\delta, \epsilon_1) := \{z \in V(D) : |\psi_D(z)| \ge 1 - \delta, |\psi_D(z) - \psi_D(u)| > \epsilon_1\}.$$

Our first goal is to show that, for every $\epsilon_1 \in (0, 1/4)$, there is some $\delta = \delta(\epsilon_1) > 0$ and some $r_0 = r_0(\epsilon_1) > 0$ such that

$$\max\{h(z) : z \in V(\delta, \epsilon_1)\} < \epsilon_1, \tag{5.5}$$

provided that $\mathrm{rad}_0(D) > r_0$. This will be achieved by first showing that h is not too large on the set

$$W := \{z \in V(D) : \epsilon_1/2 \ge |\psi_D(z) - \psi_D(u)| \ge \epsilon_1/3\}$$

and then letting δ go to 0 and appealing to Lemma 5.4.

Assume that $\mathrm{rad}_0(D)$ is sufficiently large so that any nearest neighbor path from 0 to u in D has a vertex in W. Let M denote the maximum of h on W. We claim that M is bounded by a constant $c = c(\epsilon_1)$ depending only on ϵ_1. Indeed, let K be the set of all $v \in V(D)$ satisfying $h(v) \ge M/2$ and let K' be the union of all edges where both endpoints are in $K \cup \{u\}$. Then the maximum principle shows that K' is connected and contains a simple nearest neighbor path J joining W to u whose vertices are in $\{z \in V(D) : |\psi_D(z) - \psi_D(u)| \le \epsilon_1/2\} \cup \{u\}$. Note that, in particular, $\mathrm{diam}(\psi_D(J)) \ge \epsilon_1/3$. Consequently, the continuous harmonic measure from 0 of J in D is bounded from below by some constant $c_1(\epsilon_1) > 0$.

We claim that the discrete harmonic measure $H_D(0, J)$ of J at the origin is also bounded away from 0 if r_0 is large enough. Indeed, let $D' = D \setminus J$ and let A be the arc on $\partial\mathbb{U}$ corresponding to J under the map $\psi_{D'}$. The length of A is bounded from below, since it is equal to 2π times the harmonic measure of J. Let A' denote the middle subarc of A having half the length of A. By Lemma 5.4 applied to the domain D', it follows that there is a $c_2 = c_2(\epsilon) \in (0, 1/10)$, such that $H_{D'}(v, J) \geq 1/2$ on vertices v such that $\psi_{D'}(v)$ is within distance c_2 of A'. Using Lemma 5.3 with ϵ replaced by $c_2/4$, we find that if r_0 is large, there is a nonnegative harmonic function $h_J^* : D' \to [0, \infty)$ such that $|h_J^*(v) - H_{D'}(v, J)| \leq c_2/4$ for all $v \in V(D')$ satisfying $|\psi_{D'}(v)| < 1 - c_2/4$. Take $a \in A'$ and $z := (1 - c_2/2)a$. Then it follows from the Koebe distortion theorem (as in the argument toward the end of the proof of Proposition 3.4) that we may find a vertex $v \in V(D')$ such that $|\psi_{D'}(v) - z| < c_2/4$, assuming that r_0 is large enough. Thus, $h_J^*(v) > H_{D'}(v, J) - c_2/4 \geq 1/2 - c_2/4 > 1/4$. By the Harnack principle applied to h_J^*, there is a universal constant $c_3 > 0$ such that $h_J^*(z) \geq c_3$. Since this applies to every $z \in (1 - c_2/2)A'$, the mean value property for h_J^* gives $h_J^*(0) \geq c_3 \operatorname{length}(A')/(2\pi)$. Since $|h_J^*(0) - H_{D'}(0, J)| < c_2/4$, our claim that $H_D'(0, J)$ is bounded away from 0 is established. Since h is positive, harmonic $h(0) = 1$ and $h \geq M/2$ on J, this also gives the bound $M \leq c(\epsilon)$.

Since h is harmonic, Lemma 5.4 with $\tilde{\epsilon}_1 = \epsilon_1/2$ and $\epsilon_2 := \tilde{\epsilon}_1/c(\epsilon)$ instead of ϵ_1, ϵ_2 implies that if $\delta = \delta(\epsilon_1)$ is sufficiently small, and $\operatorname{rad}_0(D)$ is large enough to guarantee that W separates $V(\delta, \epsilon_1)$ from u (in the graph-connectivity sense), then $h(z) \leq M\epsilon_2 < \epsilon_1$ for all $z \in V(\delta, \epsilon_1)$: (5.5) holds.

Now apply Lemma 5.3 again to conclude that there is a harmonic function $\hat{h} : \mathbb{U} \to [0, \infty)$ such that

$$|\hat{h} \circ \psi_D(z) - h(z)| < \epsilon_1$$

for all $z \in V(D)$ such that $|\psi_D(z)| < 1 - \delta/4$. Set $\tilde{h}(z) := \hat{h}((1 - \delta/2)z)$. We know that $\tilde{h} \geq 0$ in $\partial\mathbb{U}$, $\tilde{h}(z') \leq 2\epsilon_1$ on the set $S := \{z' \in \partial\mathbb{U} : |z' - \psi_D(u)| \geq 2\epsilon_1\}$. Consequently, the Poisson representation of $\tilde{h}$ gives

$$\tilde{h}(z) = O(\epsilon_1) + \int_{\partial\mathbb{U}\setminus S} \tilde{h}(z') \frac{1 - |z|^2}{|z - z'|^2} \, |dz'|.$$

Since $\tilde{h}(0) = 1 + O(\epsilon_1)$ and ϵ_1 is arbitrary, the proposition follows. □

5.3. *Some Green's function estimates.* As opposed to Proposition 2.2, Lemma 3.5 requires only crude bounds. It is actually possible to prove that $G_0(0, 0) - G_m(0, 0)$ is close to t_m, but we do not need this result here.

PROOF OF LEMMA 3.5. We start with (3.5). Let S be the set of vertices in $V(D)$ satisfying (3.3) and assume $S \neq \varnothing$. For a random walk starting from a vertex

in S, there is probability bounded away from 0 that within $\mathrm{rad}_0(D)^2$ steps it will exit D. This gives

$$\sum_{w\in S} G_D(0,w) \le O(1)\,\mathrm{rad}_0(D)^2. \tag{5.6}$$

On the other hand, with probability bounded away from 0, the number of steps into vertices in S for the random walk started at 0 that is stopped on exiting D is greater than $\mathrm{rad}_0(D)^2$. Therefore,

$$O(1)\sum_{w\in S} G_D(0,w) \ge \mathrm{rad}_0(D)^2. \tag{5.7}$$

By reversing the walk, we know that $G_D(0,w) = G_D(w,0)$. Since $G_D(w,0)$ is harmonic on $V(D)\setminus\{0\}$, the Harnack principle [i.e., $k=0$ in (5.2)] can be used to show that $G_D(w,0)/G_D(w',0) = O(1)$ when $w, w' \in S$. Combining this with $G_D(0,w) = G_D(w,0)$ and the estimates (5.6), (5.7) gives (3.5).

By Lemma 2.1, we have

$$\mathrm{diam}(\psi_D \circ \gamma[0,m]) = O(\delta). \tag{5.8}$$

In the following, we fix $\gamma[0,m]$ (i.e., it will be considered deterministic). Let z be the vertex where a simple random walk from 0 first exits D_m. By considering what happens to the random walk after first hitting z, we get the identity $G_0(0,v) - G_m(0,v) = \mathbf{E}[G_0(z,v)]$ [where $G_0(z,v) = 0$ for $z \notin V(D)$, by definition]. Consequently,

$$G_0(0,v) - G_m(0,v) \le \mathbf{P}\big[z \in \gamma[0,m]\big]\max\{\mathbf{E}[G_0(\gamma_j,v)] : j = 1,\dots,m\}.$$

By (5.8), the continuous harmonic measure from 0 of $\psi_D \circ \gamma[0,m]$ in $\mathbb{U}$ is $O(\delta)$. Therefore, the continuous harmonic measure from 0 of $\gamma[0,m]$ in D is also $O(\delta)$. As in the argument given in Section 5.2, this implies that if $\mathrm{rad}_0(D)$ is large enough, $\mathbf{P}[z \in \gamma[0,m]] = O(\delta)$.

Let K denote the disk $\{w \in D : |w - v| < \mathrm{rad}_0(D)/10\}$ and fix some $j \in \{1, 2, \dots, m\}$. Since $\psi_D \circ \gamma[0,m]$ is contained in $\overline{\mathbb{U}} \setminus (1 - O(\delta))\mathbb{U}$. It follows that the continuous harmonic measure of K from γ_j in D is $O(\delta)$. If $\psi_D(\gamma_j)$ is sufficiently close to $\partial\mathbb{U}$ (how close may depend on δ), then we can make sure that the corresponding discrete harmonic measure $H_D(\gamma_j, K)$ is less than δ by Lemma 5.4. If $\psi_D(\gamma_j)$ is not close to $\partial\mathbb{U}$, then when $\mathrm{rad}_0(D)$ is large the bound $H_D(\gamma_j, K) \le O(\delta)$ follows by the convergence of the discrete harmonic measure to the continuous harmonic measure, as we have seen before. If $w \in V(D) \cap K$ neighbors with a vertex outside of K, then $G_0(w,v) = O(1)$ follows from (5.7) by translating w to 0. Hence, $G_0(\gamma_j, v) = O(1)\, H_D(\gamma_j, K) = O(\delta)$. Putting these estimates together completes the proof. $\square$

5.4. *Mixed boundary conditions.* Recall that Proposition 4.2, which we will now prove, is not used in the proof of Theorem 1.1.

PROOF OF PROPOSITION 4.2. Suppose first that the distance between A_0 and A_1 is at least ϵ. Let A_0^* and A_1^* denote the two connected components of $\eta \setminus (A_0' \cup A_1')$ such that the sequence $(A_0', A_0^*, A_1', A_1^*)$ conforms to the counterclockwise order along η. This induces a corresponding partition $E_2 = E_0^* \cup E_1^*$ of E_2, according to whether or not the first point on the edge is in A_0^* or in A_1^*.

We need to use the discrete harmonic conjugate function $\hat{k}$ of $\hat{h}$. To be perfectly precise, it is necessary to set some combinatorial infrastructure: we first define a (multi-) graph $\hat{H}$, and $\hat{k}$ will be defined on the planar dual $\hat{H}^\dagger$ of $\hat{H}$. The vertices of $\hat{H}$ are $V_H \cup \{v_0, v_1\}$ (where v_0 and v_1 are new symbols not appearing in V_H). As edges of $\hat{H}$, we take all the edges of H, and, additionally, for every $j = 0, 1$ and every directed edge $[v, u]$ in E_j, there is a corresponding edge $[v, v_j]$ in $\hat{H}$. Finally, there is also the edge $[v_0, v_1]$ in $\hat{H}$. Consider a planar embedding of $\hat{H}$ which extends the planar embedding of H such that v_0 and v_1 are in the unbounded component of $\mathbb{C} \setminus H$. Let $\hat{H}^\dagger$ denote the planar dual of $\hat{H}$. Then there is a unique edge $[v_0^\dagger, v_1^\dagger]$ in $\hat{H}^\dagger$ which crosses $[v_0, v_1]$. We choose the labels so that $v_j^\dagger$ naturally corresponds to A_j^*, $j = 0, 1$. Set $\hat{h}(v_j) := j$, $j = 0, 1$. If we consider $\hat{h}$ as a function on $\hat{H}$, then it is discrete harmonic except at v_0 and v_1. This easily implies (see, e.g., [7] or, more explicitly, [4]) that there is a discrete harmonic conjugate $\hat{k}$ defined on the vertices of $\hat{H}^\dagger$; that is, for every directed edge $e = [u, v]$ in $\hat{H}$ if $\{u, v\} \neq \{v_0, v_1\}$, then the discrete Cauchy–Riemann equation $\hat{h}(v) - \hat{h}(u) = \hat{k}(v^\dagger) - \hat{k}(u^\dagger)$ holds, where $[u^\dagger, v^\dagger]$ is the edge of $\hat{H}^\dagger$ intersecting e from right to left. In fact, $\hat{k}$ is harmonic in $\hat{H}^\dagger$ except at $v_0^\dagger$ and $v_1^\dagger$. The function $\hat{k}$ is unique, up to an additive constant. We choose the additive constant so that $\hat{k}(v_0^\dagger) = 0$. Since $\hat{h} \geq 0$, by considering the neighbors of v_0 and the orientation, it follows that $\hat{k}(v_1^\dagger) \geq 0$.

Consider a sequence D_n of such domains satisfying $\mathrm{rad}_0(D_n) \geq n$, with arcs $\eta = \eta_n$ and such harmonic functions $\hat{h}_n, \hat{k}_n$. Let L_n denote the maximum value of $\hat{k}_n$, which is the value of $\hat{k}_n$ on $v_1^\dagger$.

Since $\epsilon > 0$ is fixed, we can consider a subsequence of $n \to \infty$ such that the arcs A_0 and A_1 converge to arcs $\tilde{A}_0$ and $\tilde{A}_1$ of length at least ϵ, and the distance between them is at least ϵ. Let $\tilde{A}_0^*$ and $\tilde{A}_1^*$ denote the two components of $\partial\mathbb{U} \setminus (\tilde{A}_0 \cup \tilde{A}_1)$, so that $\tilde{A}_0, \tilde{A}_0^*, \tilde{A}_1, \tilde{A}_1^*$ is the positive order along $\partial\mathbb{U}$ of these arcs.

We now separate the argument into two cases according to whether or not $L_n > 1$. Suppose that $L_n > 1$ for infinitely many n and take a further subsequence of n such that $L_n > 1$ along that subsequence. Then $\hat{k}_n / L_n$ and $\hat{h}_n / L_n$ are both bounded by 1. It follows from Lemma 5.3 that after taking a further subsequence, if necessary, there are harmonic functions h and k on $\mathbb{U}$

such that $L_n^{-1}\hat{h}_n \circ \psi_{D_n}^{-1} \to h$ and $L_n^{-1}\hat{k}_n \circ \psi_{D_n}^{-1} \to k$ uniformly on compact subsets of $\mathbb{U}$ (appropriately interpreted, since $\hat{h}_n$ and $\hat{k}_n$ are only defined on vertices and dual vertices, not on every point of D_n). Moreover, (5.4) shows that h and k are harmonic conjugates, because the discrete Cauchy–Riemann equations tend to the continuous Cauchy–Riemann equations.

By Lemma 5.4, it follows that k is, respectively, equal to 0 and 1 in the relative interior of $\tilde{A}_0^*$, $\tilde{A}_1^*$, and similarly h has boundary values 0 and $1/\tilde{L}$ in $\tilde{A}_0$ and $\tilde{A}_1$, where $\tilde{L} := \lim_{n\to\infty} L_n$ (where the limit is along the subsequence and must exist and be finite). By Schwarz reflection, say, this implies that h and k satisfy Neumann boundary conditions in $\tilde{A}_0^* \cup \tilde{A}_1^*$ and $\tilde{A}_0 \cup \tilde{A}_1$, respectively. It now easily follows (e.g., from the maximum principle) that $h + ik$ is the (unique) conformal map taking $\mathbb{U}$ to the rectangle $[0, 1/\tilde{L}] \times [0, 1]$ which takes the four arcs $\tilde{A}_0$, $\tilde{A}_0^*$, $\tilde{A}_1$, $\tilde{A}_1^*$ to the corresponding sides of the rectangle.

The argument in the case where $L_n \le 1$ for infinitely many n proceeds in the same manner, except that one should not divide $\hat{h}_n \circ \psi_{D_n}^{-1}$ and $\hat{k}_n \circ \psi_{D_n}^{-1}$ by L_n.

It remains to remove the assumption that the distance between A_0 and A_1 is at least ϵ. Observe that the probabilistic description of $\hat{h}$ shows that it is monotone increasing in A_1' and monotone decreasing in A_0'. Take $\epsilon' > 0$ much smaller than ϵ. Then $\hat{h}(0)$ for the given configuration is bounded from above by the value of $\hat{h}(0)$ for the configuration where arcs of length ϵ' are removed at the two ends of A_1, and A_1' is adjusted accordingly. Similarly, $\hat{h}(0)$ is bounded from below by the value of $\hat{h}(0)$ for the configuration where such arcs are removed at the two ends of A_0. The difference between the value of h for the original versus any of the modified configurations goes to 0 as $\epsilon' \to 0$, since h depends continuously on (A_1, A_2), as long as the length of $A_1 \cup A_2$ is not 0. Consequently, we get the proposition by applying the restricted version proved above with ϵ' in place of ϵ and by "sandwiching." □

6. Other lattices. For convenience and simplicity, the proofs up to now have been written for the loop-erased random walk and UST Peano curve on the square grid. The purpose of this section is to briefly indicate how to adapt the proofs to more general walks on more general grids. In order to keep this section short, we will not try to consider the most general cases.

Let L be a (strictly two-dimensional) lattice in $\mathbb{R}^2$; that is, L is a discrete additive subgroup of $\mathbb{R}^2$ that is not contained in a line. Discrete means that there is some neighborhood of 0 whose intersection with L is $\{0\}$. Suppose that G is a planar graph whose vertices are the elements of L, and G is invariant under translation by elements of L. That is, if $u, v \in L$ are neighbors in G and $\ell \in L$, then $\ell + u$ neighbors $\ell + v$. It is not hard to verify that there is a linear map taking L to the triangular lattice such that neighbors in G are mapped to vertices at distance 1. In particular, as a graph, G is isomorphic to the triangular grid or to the square grid.

Let N be the set of neighbors of 0 in G and let $N' := \{0\} \cup N$. Let X be an N'-valued random variable and let $X_1, X_2, \ldots$ be an i.i.d. sequence where each X_n has the same law as X. Consider the random walk

$$S_n := \sum_{j=1}^{n} X_j$$

on G. We are interested in the situation where the scaling limit of S_n is standard Brownian motion. For this purpose, we require that $\mathbf{E}[X] = 0$ and that the covariance matrix of X is the identity matrix. (Note that if the covariance matrix of X is nondegenerate but not equal to the identity, we can always apply a linear transformation to the system to convert to the above situation. Therefore, what we say below also applies in that case, provided that we appropriately modify the linear complex structure on $\mathbb{R}^2$.)

Note that, under these assumptions, the Markov chain corresponding to the walk S_n does not need to be reversible. An interesting particular example the reader may wish to keep in mind is where $\mathbf{P}[X = \exp(2\pi ij/3)] = 1/3$ for $j = 0, 1, 2$.

THEOREM 6.1. *Theorem* 1.1 *applies to the loop erasure of the random walk* S_n.

PROOF. An inspection of the proof of Theorem 1.1, including all the necessary lemmas, shows that only the generalization of the proof of Lemma 5.1 to the present framework requires special justification, which is given below. □

LEMMA 6.2. *Let* τ_r *denote the first time* n *with* $|S_n| \geq r$. *There exists a constant* C, *depending on* X *but not on* r, *such that, for all* $r \geq C$, $w \in N$ *and* $y \in L$,

$$\left|\mathbf{P}^0[S_{\tau_r} = y] - \mathbf{P}^w[S_{\tau_r} = y]\right| \leq Cr^{-1}\mathbf{P}^0[S_{\tau_r} = y].$$

Here, $\mathbf{P}^w$ denotes the law of the Markov chain started from w; that is, the law of $(S_n + w : n \in \mathbb{N})$ under $\mathbf{P} = \mathbf{P}^0$. This lemma is clearly sufficient to provide the necessary analog of Lemma 5.1 for S_n.

PROOF OF LEMMA 6.2. There are various ways to prove the lemma (via coupling, for instance). We give here a proof based on the Green's functions, as in [25]. Without loss of generality, we assume that $\mathbf{P}[X = 0] > 0$ and that L is the minimal lattice containing $\{w \in N : p_w > 0\}$. Then the random walk is irreducible on L. The discrete Laplacian Δ_X associated with X is defined by

$$\Delta_X f(z) := \mathbf{E}[f(z + X)] - f(z).$$

Let a be the potential kernel for the random walk,

$$a(z) := \sum_{j=0}^{\infty} (\mathbf{P}^0[S_j = 0] - \mathbf{P}^0[S_j = -z]).$$

It is known that the series converges and, in fact,

$$a(z) = c_1 \log|z| + c_2 + O(|z|^{-1}) \tag{6.1}$$

as $|z| \to \infty$, $z \in L$ (where c_1, c_2 depend on the law of X). This is proved in [12] for the lattice $\mathbb{Z}^2$ with arbitrary nondegenerate covariance matrix (with an appropriate dependence on the matrix), so the above follows for other L by applying a linear transformation. Since $\mathbf{P}^0[S_j = -z] = \mathbf{P}^z[S_j = 0]$, it follows that

$$\Delta_X a(z) = \begin{cases} 1, & z = 0, \\ 0, & z \neq 0. \end{cases}$$

Let G_r denote the Green's function for the walk in $L \cap r\mathbb{U}$, that is, $G_r(z, z') := \sum_{j \in \mathbb{N}} \mathbf{P}^z[j < \tau_r,\ S_j = z']$. Note that, for all $z, w \in L \cap r\mathbb{U}$,

$$a(z - w) + G_r(z, w) = \mathbf{E}^z\big[a\big(S_{\tau_r} - w\big)\big], \tag{6.2}$$

since for fixed w both sides are Δ_X-harmonic for $z \in L \cap r\mathbb{U}$ and equality holds for $z \in L \setminus r\mathbb{U}$. Set $M := \max\{|w| : w \in N\}$ and $Z = \{z \in L : r/2 \le |z| < M + r/2\}$. By (6.1) and (6.2), $G_r(z, w) = c_1 \log 2 + O(r^{-1})$ for $z \in Z$ and $w \in N'$. The same argument applied to the reverse walk $-S_j$, which has potential kernel $\bar{a}(z) = a(-z)$ and Green's function $\bar{G}_r(z, w) = G_r(w, z)$, gives

$$\forall\, w \in N',\ \forall\, z \in Z, \qquad G_r(w, z) = c_1 \log 2 + O(r^{-1}). \tag{6.3}$$

Assuming $r > 4M$, by considering the last vertex in Z visited by the walk before time τ_r, we obtain, for all $w \in N'$ and all $y \in L$,

$$\mathbf{P}^w\big[S_{\tau_r} = y\big] = \sum_{z \in Z} G_r(w, z)\mathbf{P}^z\big[S_{\tau_r} = y, \min\{j \ge 1 : S_j \in Z\} > \tau_r\big].$$

Together with (6.3), this completes the proof of the lemma. □

Observe that Theorem 1.1 also holds for the simple random walk on the honeycomb grid, because two steps on the honeycomb lattice are the same as a single step on a triangular grid containing every other vertex on the honeycomb grid, and so Lemma 6.2 may be applied.

We now turn our attention to spanning trees and the generalizations of Corollary 1.2 and Theorem 1.3. Suppose that X and $-X$ have the same distribution, so that the walk S is reversible. For an edge $e = [x, y]$, define $p_e = \mathbf{P}[X = y - x] = \mathbf{P}[X = x - y]$. In this case, it is easy to generalize the definition of UST to a measure on trees related to the law of X. This can be done either by using Wilson's algorithm or, equivalently, by giving to each tree T a probability that is proportional to the product of the transition probabilities along the edges of T. In other words, $\mathbf{P}[T] = Z^{-1} \prod_{e \in T} p_e$, where Z is a normalizing constant. (The equivalence is proved in [44]; see also [26].) We call this the UST corresponding to the walk S (even if this probability measure is not uniform). Note that Lemma 4.1 also holds in the present setting because the probability $\mathbf{P}[T]$ is given in terms of a product.

THEOREM 6.3. *Assuming that $-X$ has the same distribution as X (i.e., S_j is reversible), Corollary* 1.2 *and Theorem* 1.3 *hold for the UST corresponding to the walk S.*

PROOF. The proof of Corollary 1.2 holds in this generality. In the proof of Theorem 1.3, the only significant changes concern the discrete harmonic conjugate function, used in the proof of Proposition 4.2. Recall that there is an appropriate definition for the discrete harmonic conjugate for reversible walks on planar graphs, where the discrete Cauchy–Riemann equation is modified (see [7] or [19], Section 6.1). If G is graph isomorphic to the square grid, the same is true for the dual graph. If G is graph isomorphic to the triangular grid, then the dual is graph isomorphic to the honeycomb grid. As pointed out above, Lemma 6.2 may therefore be applied to the harmonic conjugate. The details are left to the reader. □

In the nonreversible setting, instead of a spanning tree, one should consider a spanning arborescence, which is an oriented tree with a root and the edges are oriented toward the root. Fix a finite Markov chain with state space V and a root $o \in V$. Consider the measure on spanning arborescences of V with root o, where the probability for T is proportional to the product of the transition probabilities along the directed edges of T. This is the analog of the UST in the nonreversible setting. Wilson's algorithm holds in this generality (see [44]); however, the choice of the root o clearly matters.

If we consider a finite piece of the lattice L, and we wire part or all of the boundary, it is natural to pick the wired vertex as the root. With this convention, Corollary 1.2 holds for the wired tree. It would be interesting to see if the free tree with root chosen at $0 \in D$ is invariant under conformal maps preserving 0, say (in the nonreversible setting). Of course, one needs to choose a grid approximation of D where there is an oriented path from each vertex to the root 0.

In the proof of Theorem 1.3, we have used reversibility in two places. The proofs of Theorems 10.7 and 11.1 of [38], which we quoted, currently require reversibility. However, these results were only used to improve the topology of convergence to SLE. More seriously, Section 5.4 uses the conjugate harmonic function, whose definition in the nonreversible setting is not clear. Notwithstanding the obstacles, it seems likely that these results can be proven in the nonreversible setting, too.

REFERENCES

[1] AHLFORS, L. V. (1973). *Conformal Invariants: Topics in Geometric Function Theory*. McGraw-Hill, New York.

[2] AIZENMAN, M. and BURCHARD, A. (1999). Hölder regularity and dimension bounds for random curves. *Duke Math. J.* **99** 419–453.

[3] AIZENMAN, M., BURCHARD, A., NEWMAN, C. M. and WILSON, D. B. (1997). Scaling limits for minimal and random spanning trees in two dimensions. *Random Structures Algorithms* **15** 319–367.
[4] BENJAMINI, I. and SCHRAMM, O. (1996). Random walks and harmonic functions on infinite planar graphs using square tilings. *Ann. Probab.* **24** 1219–1238.
[5] CARDY, J. L. (1992). Critical percolation in finite geometries. *J. Phys. A* **25** L201–L206.
[6] COLLATZ, L. (1960). *The Numerical Treatment of Differential Equations*, 3rd ed. Springer, Berlin.
[7] DEHN, M. (1903). Über die Zerlegung von Rechtecken in Rechtecke. *Math. Ann.* **57** 314–332.
[8] DUBINS, L. E. (1968). On a theorem of Skorohod. *Ann. Math. Statist.* **39** 2094–2097.
[9] DUDLEY, R. M. (1989). *Real Analysis and Probability*. Wadsworth and Brooks/Cole, Pacific Grove, CA.
[10] DUPLANTIER, B. (1987). Critical exponents of Manhattan Hamiltonian walks in two dimensions, from Potts and $o(n)$ models. *J. Statist. Phys.* **49** 411–431.
[11] DUPLANTIER, B. (1992). Loop-erased self-avoiding walks in two dimensions: Exact critical exponents and winding numbers. *Phys. A* **191** 516–522.
[12] FOMIN, S. (2001). Loop-erased walks and total positivity. *Trans. Amer. Math. Soc.* **353** 3563–3583.
[13] FUKAI, Y. and UCHIYAMA, K. (1996). Potential kernel for two-dimensional random walk. *Ann. Probab.* **24** 1979–1992.
[14] GUTTMANN, A. and BURSILL, R. (1990). Critical exponent for the loop-erased self-avoiding walk by Monte-Carlo methods. *J. Statist. Phys.* **59** 1–9.
[15] HÄGGSTRÖM, O. (1995). Random-cluster measures and uniform spanning trees. *Stochastic Process. Appl.* **59** 267–275.
[16] HE, Z.-X. and SCHRAMM, O. (1998). The C^∞-convergence of hexagonal disk packings to the Riemann map. *Acta Math.* **180** 219–245.
[17] KARATZAS, I. and SHREVE, S. E. (1988). *Brownian Motion and Stochastic Calculus*. Springer, New York.
[18] KASTELEYN, P. W. (1963). A soluble self-avoiding walk problem. *Physica* **29** 1329–1337.
[19] KENYON, R. (1998). Tilings and discrete Dirichlet problems. *Israel J. Math.* **105** 61–84.
[20] KENYON, R. (2000). The asymptotic determinant of the discrete Laplacian. *Acta Math.* **185** 239–286.
[21] KENYON, R. (2000). Long-range properties of spanning trees. *J. Math. Phys.* **41** 1338–1363.
[22] KENYON, R. (2000). Conformal invariance of domino tilings. *Ann. Probab.* **28** 759–795.
[23] KOZMA, G. (2002). Scaling limit of loop-erased random walks: A naive approach. Unpublished manuscript.
[24] LAWLER, G. F. (1980). A self-avoiding random walk. *Duke Math. J.* **47** 655–693.
[25] LAWLER, G. F. (1991). *Intersections of Random Walks*. Birkhäuser, Boston.
[26] LAWLER, G. F. (1999). Loop-erased random walk. In *Perplexing Problems in Probability* 197–217. Birkhäuser, Boston.
[27] LAWLER, G. F., SCHRAMM, O. and WERNER, W. (2001). Values of Brownian intersection exponents I. Half-plane exponents. *Acta Math.* **187** 237–273.
[28] LAWLER, G. F., SCHRAMM, O. and WERNER, W. (2001). Values of Brownian intersection exponents II. Plane exponents. *Acta Math.* **187** 275–308.
[29] LAWLER, G. F., SCHRAMM, O. and WERNER, W. (2002). Analyticity of intersection exponents for planar Brownian motion. *Acta Math.* **189** 179–201.
[30] LAWLER, G. F., SCHRAMM, O. and WERNER, W. (2002). One-arm exponent for critical 2D percolation. *Electron. J. Probab.* **7** No. 2.
[31] LAWLER, G. F., SCHRAMM, O. and WERNER, W. (2003). Conformal restriction properties: The chordal case. *J. Amer. Math. Soc.* **16** 917–955.

[32] LYONS, R. (1998). A bird's-eye view of uniform spanning trees and forests. In *Microsurveys in Discrete Probability* 135–162. Amer. Math. Soc., Providence, RI.
[33] MAJUMDAR, S. N. (1992). Exact fractal dimension of the loop-erased self-avoiding walk in two dimensions. *Phys. Rev. Lett.* **68** 2329–2331.
[34] PEMANTLE, R. (1991). Choosing a spanning tree for the integer lattice uniformly. *Ann. Probab.* **19** 1559–1574.
[35] POMMERENKE, CH. (1992). *Boundary Behaviour of Conformal Maps*. Springer, Berlin.
[36] REVUZ, D. and YOR, M. (1991). *Continuous Martingales and Brownian Motion*. Springer, Berlin.
[37] ROHDE, S. and SCHRAMM, O. (2001). Basic properties of SLE. Unpublished manuscript.
[38] SCHRAMM, O. (2000). Scaling limits of loop-erased random walks and uniform spanning trees. *Israel J. Math.* **118** 221–288.
[39] SCHRAMM, O. (2001). A percolation formula. *Electron. Comm. Probab.* **6** 115–120.
[40] SMIRNOV, S. (2001). Critical percolation in the plane: Conformal invariance, Cardy's formula, scaling limits. *C. R. Acad. Sci. Paris Sér. I Math.* **333** 239–244.
[41] SMIRNOV, S. (2001). Critical percolation in the plane. I. Conformal invariance and Cardy's formula. II. Continuum scaling limit. Preprint.
[42] SMIRNOV, S. and WERNER, W. (2001). Critical exponents for two-dimensional percolation. *Math. Res. Lett.* **8** 729–744.
[43] STRASSEN, V. (1967). Almost sure behavior of sums of independent random variables and martingales. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* 315–343. Univ. California Press.
[44] WILSON, D. B. (1996). Generating random spanning trees more quickly than the cover time. In *Proceedings of the 28th Annual ACM Symposium on the Theory of Computing* 296–303. ACM, New York.

G. LAWLER
DEPARTMENT OF MATHEMATICS
310 MALOTT HALL
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853-4201
USA
E-MAIL: lawler@math.cornell.edu

O. SCHRAMM
MICROSOFT CORPORATION
ONE MICROSOFT WAY
REDMOND, WASHINGTON 98052
USA
E-MAIL: schramm@microsoft.com

W. WERNER
DÉPARTEMENT DE MATHÉMATIQUES
BÂT. 425
UNIVERSITÉ PARIS-SUD
91405 ORSAY CEDEX
FRANCE
E-MAIL: wendelin.werner@math.u-psud.fr

Annals of Mathematics, **161** (2005), 883–924

Basic properties of SLE

By Steffen Rohde* and Oded Schramm

Dedicated to Christian Pommerenke on the occasion of his 70th *birthday*

Abstract

SLE_κ is a random growth process based on Loewner's equation with driving parameter a one-dimensional Brownian motion running with speed κ. This process is intimately connected with scaling limits of percolation clusters and with the outer boundary of Brownian motion, and is conjectured to correspond to scaling limits of several other discrete processes in two dimensions.

The present paper attempts a first systematic study of SLE. It is proved that for all $\kappa \neq 8$ the SLE trace is a path; for $\kappa \in [0,4]$ it is a simple path; for $\kappa \in (4,8)$ it is a self-intersecting path; and for $\kappa > 8$ it is space-filling.

It is also shown that the Hausdorff dimension of the SLE_κ trace is almost surely (a.s.) at most $1+\kappa/8$ and that the expected number of disks of size ε needed to cover it inside a bounded set is at least $\varepsilon^{-(1+\kappa/8)+o(1)}$ for $\kappa \in [0,8)$ along some sequence $\varepsilon \searrow 0$. Similarly, for $\kappa \geq 4$, the Hausdorff dimension of the outer boundary of the SLE_κ hull is a.s. at most $1+2/\kappa$, and the expected number of disks of radius ε needed to cover it is at least $\varepsilon^{-(1+2/\kappa)+o(1)}$ for a sequence $\varepsilon \searrow 0$.

1. Introduction

Stochastic Loewner Evolution (SLE) is a random process of growth of a set K_t. The evolution of the set over time is described through the normalized conformal map $g_t = g_t(z)$ from the complement of K_t. The map g_t is the solution of Loewner's differential equation with driving parameter a one-dimensional Brownian motion. SLE, or SLE_κ, has one parameter $\kappa \geq 0$, which is the speed of the Brownian motion. A more complete definition appears in Section 2 below.

The SLE process was introduced in [Sch00]. There, it was shown that under the assumption of the existence and conformal invariance of the scaling limit of loop-erased random walk, the scaling limit is SLE_2. (See Figure 9.1.) It was also stated there without proof that SLE_6 is the scaling limit of the

*Partially supported by NSF Grants DMS-0201435 and DMS-0244408.

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_31, © Springer Science+Business Media, LLC 2011

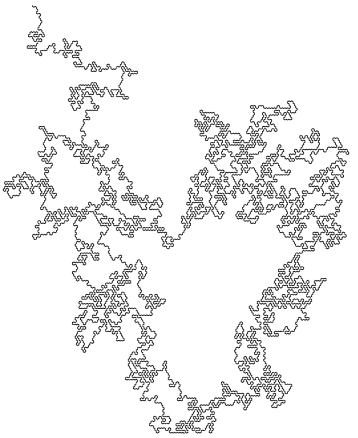


Figure 1.1: The boundary of a percolation cluster in the upper half plane, with appropriate boundary conditions. It converges to the chordal SLE_6 trace.

boundaries of critical percolation clusters, assuming their conformal invariance. Smirnov [Smi01] has recently proved the conformal invariance conjecture for critical percolation on the triangular grid and the claim that SLE_6 describes the limit. (See Figure 1.1.) With the proper setup, the outer boundary of SLE_6 is the same as the outer boundary of planar Brownian motion [LSW03] (see also [Wer01]). SLE_8 has been conjectured [Sch00] to be the scaling limit of the uniform spanning tree Peano curve (see Figure 9.2), and there are various further conjectures for other parameters. Most of these conjectures are described in Section 9 below. Also related is the work of Carleson and Makarov [CM01], which studies growth processes motivated by DLA via Loewner's equation.

SLE is amenable to computations. In [Sch00] a few properties of SLE have been derived; in particular, the winding number variance. In the series of papers [LSW01a], [LSW01b], [LSW02], a number of other properties of SLE have been studied. The goal there was not to investigate SLE for its own sake, but rather to use SLE_6 as a means for the determination of the Brownian motion intersection exponents.

As the title suggests, the goal of the present paper is to study the fundamental properties of SLE. There are two main variants of SLE, *chordal* and *radial*. For simplicity, we concentrate on chordal SLE; however, all the main results of the paper carry over to radial SLE as well. In chordal SLE, the set K_t, $t \geq 0$, called the SLE *hull*, is a subset of the closed upper half plane $\overline{\mathbb{H}}$ and $g_t : \mathbb{H} \setminus K_t \to \mathbb{H}$ is the conformal uniformizing map, suitably normalized at infinity.

We show that with the possible exception of $\kappa = 8$, a.s. there is a (unique) continuous path $\gamma : [0, \infty) \to \overline{\mathbb{H}}$ such that for each $t > 0$ the set K_t is the union of $\gamma[0, t]$ and the bounded connected components of $\overline{\mathbb{H}} \setminus \gamma[0, t]$. The path γ is called the SLE trace. It is shown that $\lim_{t\to\infty} |\gamma(t)| = \infty$ a.s. We also

describe two phase transitions for the SLE process. In the range $\kappa \in [0,4]$, a.s. $K_t = \gamma[0,t]$ for every $t \geq 0$ and γ is a simple path. For $\kappa \in (4,8)$ the path γ is not a simple path and for every $z \in \mathbb{H}$ a.s. $z \notin \gamma[0,\infty)$ but $z \in \bigcup_{t>0} K_t$. Finally, for $\kappa > 8$ we have $\overline{\mathbb{H}} = \gamma[0,\infty)$ a.s. The reader may wish to examine Figures 9.1, 1.1 and 9.2, to get an idea of what the SLE_κ trace looks like for $\kappa = 2$, 6 and 8, respectively.

We also discuss the expected number of disks needed to cover the SLE_κ trace and the outer boundary of K_t. It is proved that the Hausdorff dimension of the trace is a.s. at most $1 + \kappa/8$, and that the Hausdorff dimension of the outer boundary ∂K_t is a.s. at most $1 + 2/\kappa$ if $\kappa \geq 4$. For $\kappa \in [0,8)$, we also show that the expected number of disks of size ε needed to cover the trace inside a bounded set is at least $\varepsilon^{-(1+\kappa/8)+o(1)}$ along some sequence $\varepsilon \searrow 0$. Similarly, for $\kappa \geq 4$, the expected number of disks of radius ε needed to cover the outer boundary is at least $\varepsilon^{-(1+2/\kappa)+o(1)}$ for a sequence of $\varepsilon \searrow 0$. Richard Kenyon has earlier made the conjecture that the Hausdorff dimension of the outer boundary is a.s. $1 + 2/\kappa$. These results offer strong support for this conjecture.

It is interesting to compare our results to recent results for the deterministic Loewner evolution, i.e., the solutions to the Loewner equation with a deterministic driving function $\xi(t)$. In [MR] it is shown that if ξ is Hölder continuous with exponent $1/2$ and small norm, then K_t is a simple path. On the other hand, there is a function ξ, Hölder continuous with exponent $1/2$ and having large norm, such that K_t is not even locally connected, and therefore there is no continuous path γ generating K_t. In this example, K_t spirals infinitely often around a disk D, accumulating on ∂D, and then spirals out again. It is easy to see that the disk D can be replaced by any compact connected subset of $\mathbb{H}$. Notice that according to the law of the iterated logarithm, a.s. Brownian motion is not Hölder continuous with exponent $1/2$. Therefore, it seems unlikely that the results of the present paper can be obtained from deterministic results.

Our results are based on the computation and estimates of the distribution of $|g_t'(z)|$ where $z \in \mathbb{H}$. Note that in [LSW01b] the derivatives $g_t'(x)$ are studied for $x \in \mathbb{R}$.

The organization of the paper is as follows. Section 2 introduces the basic definitions and some fundamental properties. The goal of Section 3 is to obtain estimates for quantities related to $\mathbf{E}\big[|g_t'(z)|^a\big]$, for various constants a (another result of this nature is Lemma 6.3), and to derive some resulting continuity properties of g_t^{-1}. Section 4 proves a general criterion for the existence of a continuous trace, which does not involve randomness. The proof that the SLE_κ trace is continuous for $\kappa \neq 8$ is then completed in Section 5. There, it is also proved that g_t^{-1} is a.s. Hölder continuous when $\kappa \neq 4$. Section 6 discusses the two phase transitions $\kappa = 4$ and $\kappa = 8$ for SLE_κ. Besides some quantitative

properties, it is shown there that the trace is a.s. a simple path if and only if $\kappa \in [0,4]$, and that the trace is space-filling for $\kappa > 8$. The trace is proved to be transient when $\kappa \neq 8$ in Section 7. Estimates for the dimensions of the trace and the boundary of the hull are established in Section 8. Finally, a collection of open problems is presented in Section 9.

Update. Since the completion and distribution of the first version of this paper, there has been some further progress. In [LSW] it was proven that the scaling limit of loop-erased random walk is SLE_2 and the scaling limit of the UST Peano path is SLE_8. As a corollary of the convergence of the UST Peano path to SLE_8, it was also established there that SLE_8 is generated by a continuous transient path, thus answering some of the issues left open in the current paper. However, it is quite natural to ask for a more direct analytic proof of these properties of SLE_8.

Recently, Vincent Beffara [Bef] has announced a proof that the Hausdorff dimension of the SLE_κ trace is $1 + \kappa/8$ when $4 \neq \kappa \leq 8$.

The paper [SS] proves the convergence of the harmonic explorer to SLE_4.

2. Definitions and background

2.1. *Chordal* SLE. Let B_t be Brownian motion on $\mathbb{R}$, started from $B_0 = 0$. For $\kappa \geq 0$ let $\xi(t) := \sqrt{\kappa} B_t$ and for each $z \in \overline{\mathbb{H}} \setminus \{0\}$ let $g_t(z)$ be the solution of the ordinary differential equation

$$\partial_t g_t(z) = \frac{2}{g_t(z) - \xi(t)}, \qquad g_0(z) = z. \tag{2.1}$$

The solution exists as long as $g_t(z) - \xi(t)$ is bounded away from zero. We denote by $\tau(z)$ the first time τ such that 0 is a limit point of $g_t(z) - \xi(t)$ as $t \nearrow \tau$. Set

$$H_t := \left\{ z \in \mathbb{H} : \tau(z) > t \right\}, \qquad K_t := \left\{ z \in \overline{\mathbb{H}} : \tau(z) \leq t \right\}.$$

It is immediate to verify that K_t is compact and H_t is open for all t. The parametrized collection of maps $(g_t : t \geq 0)$ is called *chordal* SLE_κ. The sets K_t are the *hulls* of the SLE. It is easy to verify that for every $t \geq 0$ the map $g_t : H_t \to \mathbb{H}$ is a conformal homeomorphism and that H_t is the unbounded component of $\mathbb{H} \setminus K_t$. The inverse of g_t is obtained by flowing backwards from any point $w \in \mathbb{H}$ according to the equation (2.1). (That is, the fact that g_t is invertible is a particular case of a general result on solutions of ODE's.) One only needs to note that in this backward flow, the imaginary part increases, hence the point cannot hit the singularity. It also cannot escape to infinity in finite time. The fact that $g_t(z)$ is analytic in z is clear, since the right-hand side of (2.1) is analytic in $g_t(z)$.

The map g_t satisfies the so-called *hydrodynamic* normalization at infinity:

$$\lim_{z\to\infty} g_t(z) - z = 0\,. \tag{2.2}$$

Note that this uniquely determines g_t among conformal maps from H_t onto $\mathbb{H}$. In fact, (2.1) implies that g_t has the power series expansion

$$g_t(z) = z + \frac{2t}{z} + \cdots\,, \qquad z \to \infty\,. \tag{2.3}$$

The fact that $g_t \neq g_{t'}$ when $t' > t$ implies that $K_{t'} \neq K_t$. The relation $K_{t'} \supset K_t$ is clear.

Two important properties of chordal SLE are scale-invariance and a sort of stationarity. These are summarized in the following proposition. (A similar statement appeared in [LSW01a].)

PROPOSITION 2.1. (i) SLE_κ *is scale-invariant in the following sense. Let* K_t *be the hull of* SLE_κ, *and let* $\alpha > 0$. *Then the process* $t \mapsto \alpha^{-1/2} K_{\alpha t}$ *has the same law as* $t \mapsto K_t$. *The process* $(t,z) \mapsto \alpha^{-1/2}\, g_{\alpha t}(\sqrt{\alpha} z)$ *has the same law as the process* $(t,z) \mapsto g_t(z)$.

(ii) *Let* $t_0 > 0$. *Then the map* $(t,z) \mapsto \tilde{g}_t(z) := g_{t+t_0} \circ g_{t_0}^{-1}\big(z + \xi(t_0)\big) - \xi(t_0)$ *has the same law as* $(t,z) \mapsto g_t(z)$; *moreover,* $(\tilde{g}_t)_{t\geq 0}$ *is independent from* $(g_t)_{0\leq t\leq t_0}$.

The scaling property easily follows from the scaling property of Brownian motion, and the second property follows from the Markov property and translation invariance of Brownian motion. One just needs to write down the expression for $\partial_t \tilde{g}_t$. We leave the details as an exercise to the reader.

The following notations will be used throughout the paper.

$$f_t := g_t^{-1}, \qquad \hat{f}_t(z) := f_t\big(z + \xi(t)\big).$$

The *trace* γ of SLE is defined by

$$\gamma(t) := \lim_{z\to 0} \hat{f}_t(z)\,,$$

where z tends to 0 within $\mathbb{H}$. If the limit does not exist, let $\gamma(t)$ denote the set of all limit points. We say that the SLE trace is a continuous path if the limit exists for every t and $\gamma(t)$ is a continuous function of t.

2.2. *Radial* SLE. Another version of SLE_κ is called *radial* SLE_κ. It is similar to chordal SLE but appropriate for the situation where there is a distinguished point in the interior of the domain. Radial SLE_κ is defined as follows. Let $B(t)$ be Brownian motion on the unit circle $\partial\mathbb{U}$, started from a uniform-random point $B(0)$, and set $\xi(t) := B(\kappa t)$. The conformal maps g_t are defined as the solution of

$$\partial_t g_t(z) = -g_t(z)\,\frac{g_t(z) + \xi(t)}{g_t(z) - \xi(t)}\,, \qquad g_0(z) = z\,,$$

for $z \in \mathbb{U}$. The sets K_t and H_t are defined as for chordal SLE. Note that the scaling property 2.1.(i) fails for radial SLE. Mainly due to this reason, chordal SLE is easier to work with. However, appropriately stated, all the main results of this paper are valid also for radial SLE. This follows from [LSW01b, Prop. 4.2], which says in a precise way that chordal and radial SLE are equivalent.

2.3. *Local martingales and martingales.* The purpose of this subsection is to present a slightly technical lemma giving a sufficient condition for a local martingale to be a martingale. Although we have not been able to find an appropriate reference, the lemma must be known (and is rather obvious to the experts).

See, for example, [RY99, §IV.1] for a discussion of the distinction between a local martingale and a martingale.

While the stochastic calculus needed for the rest of the paper is not much more than familiarity with Itô's formula, this subsection does assume a bit more.

LEMMA 2.2. *Let B_t be stardard one dimensional Brownian motion, and let a_t be a progressive real valued locally bounded process. Suppose that X_t satisfies*

$$X_t = \int_0^t a_s \, dB_s \,,$$

and that for every $t > 0$ there is a finite constant $c(t)$ such that

$$a_s^2 \le c(t) X_s^2 + c(t) \tag{2.4}$$

for all $s \in [0,t]$ a.s. Then X is a martingale.

Proof. We know that X is a local martingale. Let $M > 0$ be large, and let $T := \inf\{t : |X_t| \ge M\}$. Then $Y_t := X_{t\wedge T}$ is a martingale (where $t \wedge T = \min\{t, T\}$). Let $f(t) := \mathbf{E}\big[Y_t^2\big]$. Itô's formula then gives

$$f(t') = \mathbf{E}\Big[\int_0^{t'} a_s^2 \, 1_{s<T} \, ds\Big] \,.$$

Our assumption $a_s^2 \le c(t) X_s^2 + c(t)$ therefore implies that for $t' \in [0,t]$

$$f(t') \le c(t)\, t' + c(t) \int_0^{t'} f(s) \, ds \,. \tag{2.5}$$

This implies $f(s) < \exp(2\, c(t)\, s)$ for all $s \in [0,t]$, since (2.5) shows that t' cannot be the least $s \in [0,t]$ where $f(s) \ge \exp(2\, c(t)\, s)$. Thus,

$$\mathbf{E}\big[\langle X, X\rangle_{t\wedge T}\big] = \mathbf{E}\big[\langle Y, Y\rangle_t\big] = \mathbf{E}\big[Y_t^2\big] = f(t) < \exp(2\, c(t)\, t) \,.$$

Taking $M \to \infty$, we get by monotone convergence $\mathbf{E}\big[\langle X, X\rangle_t\big] \le \exp(2\, c(t)\, t)$ $< \infty$. Thus X is a martingale (for example, by [RY99, IV.1.25]). □

3. Derivative expectation

In this section, g_t is the SLE_κ flow; that is, the solution of (2.1) where $\xi(t) := B(\kappa t)$, and B is standard Brownian motion on $\mathbb{R}$ starting from 0. Our only assumption on κ is $\kappa > 0$. The goal of the section is to derive bounds on quantities related to $\mathbf{E}\big[|g_t'(z)|^a\big]$. Another result of this nature is Lemma 6.3, which is deferred to a later section.

3.1. *Basic derivative expectation.* We will need estimates for the moments of $|\hat{f}_t'|$. In this subsection, we will describe a change of time and obtain derivative estimates for the changed time.

For convenience, we take B to be two-sided Brownian motion. The equation (2.1) can also be solved for negative t, and g_t is a conformal map from $\mathbb{H}$ into a subset of $\mathbb{H}$ when $t < 0$. Notice that the scale invariance (Proposition 2.1.(i)) also holds for $t < 0$.

LEMMA 3.1. *For all fixed $t \in \mathbb{R}$ the map $z \mapsto g_{-t}(z)$ has the same distribution as the map $z \mapsto \hat{f}_t(z) - \xi(t)$.*

Proof. Fix $t_1 \in \mathbb{R}$, and let

$$\hat{\xi}_{t_1}(t) = \xi(t_1 + t) - \xi(t_1)\,. \tag{3.1}$$

Then $\hat{\xi}_{t_1} : \mathbb{R} \to \mathbb{R}$ has the same law as $\xi : \mathbb{R} \to \mathbb{R}$. Let

$$\hat{g}_t(z) := g_{t_1+t} \circ g_{t_1}^{-1}\big(z + \xi(t_1)\big) - \xi(t_1)\,,$$

and note that $\hat{g}_0(z) = z$ and $\hat{g}_{-t_1}(z) = \hat{f}_{t_1}(z) - \xi(t_1)$. Since

$$\partial_t \hat{g}_t = \frac{2}{\hat{g}_t + \xi(t_1) - \xi(t + t_1)} = \frac{2}{\hat{g}_t - \hat{\xi}_{t_1}(t)}\,,$$

the lemma follows from (2.1). □

Note that (2.1) implies that $\mathrm{Im}\big(g_t(z)\big)$ is monotone decreasing in t for every $z \in \mathbb{H}$. For $z \in \mathbb{H}$ and $u \in \mathbb{R}$ set

$$T_u = T_u(z) := \sup\Big\{t \in \mathbb{R} : \mathrm{Im}\big(g_t(z)\big) \geq e^u\Big\}\,. \tag{3.2}$$

We claim that for all $z \in \mathbb{H}$ a.s. $T_u \neq \pm\infty$. By (2.1),

$$|\partial_t g_t(z)| = 2|g_t(z) - \xi(t)|^{-1}.$$

Setting $\bar{\xi}(t) := \sup\big\{|\xi(s)| : s \in [0,t]\big\}$, the above implies that $|g_t(z)| \leq |z| + \bar{\xi}(t) + 2\sqrt{t}$ for $t < \tau(z)$, since $\partial_t|g_t(z)| \leq 2/(|g_t(z)| - \bar{\xi}(t))$ whenever $|g_t(z)| > \bar{\xi}(t)$. Setting $y_t := \mathrm{Im} g_t(z)$, we get from (2.1),

$$-\partial_t \log y_t \geq 2\,(|z| + 2\,\bar{\xi}(t) + 2\sqrt{t})^{-2}\,.$$

The law of iterated logarithms implies that the right-hand side is not integrable over $[0,\infty)$ nor over $(-\infty,0]$. Thus, $|T_u|<\infty$ a.s.

We will need the formula

$$\begin{aligned}\partial_t \log|g_t'(z)| &= \operatorname{Re}\Big(\frac{\partial_z\partial_t g_t(z)}{g_t'(z)}\Big) = \operatorname{Re}\Big(g_t'(z)^{-1}\,\partial_z \frac{2}{g_t(z)-\xi(t)}\Big)\\ &= -2\operatorname{Re}\Big(\big(g_t(z)-\xi(t)\big)^{-2}\Big)\,.\end{aligned} \tag{3.3}$$

Set $u=u(z,t):=\log \operatorname{Im} g_t(z)$. Observe that (2.1) gives

$$\partial_t u = -2\,|g_t(z)-\xi(t)|^{-2}\,, \tag{3.4}$$

and (3.3) gives

$$\partial_u \log\big|g_t'(z)\big| = \frac{\operatorname{Re}\big((g_t(z)-\xi(t))^2\big)}{|g_t(z)-\xi(t)|^2}\,. \tag{3.5}$$

Fix some $\hat z=\hat x+i\hat y\in\mathbb{H}$. For every $u\in\mathbb{R}$, let

$$z(u):=g_{T_u(\hat z)}(\hat z)-\xi(T_u),\qquad \psi(u):=\frac{\hat y}{y(u)}\,\big|g_{T_u}'(\hat z)\big|,$$

and

$$x(u):=\operatorname{Re}(z(u)),\qquad y(u):=\operatorname{Im}(z(u))=\exp(u).$$

THEOREM 3.2. *Let $\hat z=\hat x+i\hat y\in\mathbb{H}$ as above. Assume that $\hat y\neq 1$, and set $\nu:=-\operatorname{sign}(\log\hat y)$. Let $b\in\mathbb{R}$. Define a and λ by*

$$a:=2\,b+\nu\,\kappa\,b\,(1-b)/2,\qquad \lambda:=4\,b+\nu\,\kappa\,b\,(1-2b)/2\,. \tag{3.6}$$

Set

$$F(\hat z)=F_b(\hat z):=\hat y^a\,\mathbf{E}\Big[\big(1+x(0)^2\big)^b\,\big|g_{T_0(\hat z)}'(\hat z)\big|^a\Big].$$

Then

$$F(\hat z)=\Big(1+(\hat x/\hat y)^2\Big)^b\,\hat y^\lambda\,. \tag{3.7}$$

Before we give the short proof of the theorem, a few remarks may be of help to motivate the formulation and the proof. Our goal was to find an expression for $\mathbf{E}\Big[\big|g_{T_0(\hat z)}'(\hat z)\big|^a\Big]$. It turns out to be more convenient to consider

$$\bar F(\hat z):=\hat y^a\,\mathbf{E}\Big[\big|g_{T_0(\hat z)}'(\hat z)\big|^a\Big]=\mathbf{E}\Big[\psi(0)^a\Big].$$

The obvious strategy is to find a differential equation which $\bar F$ must satisfy and search for a solution. The first part is not too difficult, and proceeds as follows. Let $\mathcal{F}_u$ denote the σ-field generated by $\langle\xi(t):(t-T_u)\nu\ge 0\rangle$. Note that the

strong Markov property for ξ and the chain rule imply that for u between 0 and $\hat{u} := \log \hat{y}$,

$$\mathbf{E}\Big[\psi(0)^a \;\Big|\; \mathcal{F}_u\Big] = \psi(u)^a \, \bar{F}\big(z(u)\big)\,. \tag{3.8}$$

Hence, the right-hand side is a martingale. Observe that

$$dx = \frac{2\,x\,dt}{x^2+y^2} - d\xi, \qquad dy = \frac{-2\,y\,dt}{x^2+y^2}, \qquad d\log\psi = \frac{4\,y^2\,dt}{(x^2+y^2)^2}\,. \tag{3.9}$$

(The latter easily follows from (3.3) and (3.4).) We assume for now that $\bar{F}$ is smooth. Itô's formula may then be calculated for the right-hand side of (3.8). Since it is a martingale, the drift term of the Itô derivative must vanish; that is,

$$\psi^a \cdot \Lambda \bar{F} = 0\,,$$

where

$$\Lambda G := -\frac{4\,\nu\,a\,y^2}{(x^2+y^2)^2}\,G - \frac{2\,\nu\,x}{x^2+y^2}\,\partial_x G + \frac{2\,\nu\,y}{x^2+y^2}\,\partial_y G + \frac{\kappa}{2}\,\partial_x^2 G\,. \tag{3.10}$$

(The $-\nu$ factor comes from the fact that t is monotone decreasing with respect to the filtration $\mathcal{F}_u$ if and only if $\nu = 1$.) Guessing a solution for the equation $\Lambda G = 0$ is not too difficult (after changing to coordinates where scale invariance is more apparent). It is easy to verify that

$$\hat{F}(x+iy) = \hat{F}_{b,\lambda}(x+iy) := \Big(1 + (x/y)^2\Big)^b\, y^\lambda\,,$$

satisfies $\Lambda\hat{F} = 0$. Unfortunately, $\hat{F}$ does not satisfy the boundary values $\hat{F} = 1$ for $y = 1$, which hold for $\bar{F}$. Consequently, the theorem gives a formula for F, rather than for $\bar{F}$. (Remark 3.4 concerns the problem of determining $\bar{F}$.) Assuming that F is C^2, the above derivation does apply to F, and shows that $\Lambda F = 0$. However, we have not found a clean reference to the fact that $F \in C^2$. Fortunately, the proof below does not need to assume this.

Proof of Theorem 3.2. Note that by (3.4)

$$du = -2\,|z|^{-2}\,dt.$$

Let

$$\hat{B}(u) := -\sqrt{2/\kappa}\int_{t=0}^{T_u} |z|^{-1}\,d\xi.$$

Then $\hat{B}$ is a local martingale and

$$d\langle \hat{B}\rangle = -2\,|z|^{-2}\,dt = du.$$

Consequently, $\hat{B}(u)$ is Brownian motion (with respect to u). Set $M_u := \psi(u)^a\,\hat{F}(z(u))$. Itô's formula gives

$$dM_u = -2M\,\frac{b\,x}{x^2+y^2}\,d\xi = \sqrt{2\,\kappa}\,M\,\frac{b\,x}{\sqrt{x^2+y^2}}\,d\hat{B}\,.$$

Thus M is a local martingale. In fact, Lemma 2.2 then tells us that M is a martingale. Consequently, we have

$$\psi(\hat{u})^a \, \hat{F}(\hat{z}) = \mathbf{E}\big[\psi(0)^a \, \hat{F}(z(0))\big] \tag{3.11}$$

and the theorem clearly follows. □

In Section 8 we will estimate the expected number of disks needed to cover the boundary of K_t. To do this, we need the following lower bound on the expectation of the derivative.

LEMMA 3.3. *Let $\kappa > 0$ and $b < (\kappa+4)/(4\kappa)$, and define a and λ by (3.6) with $\nu = 1$. Then there is a constant $c = c(\kappa, b) > 0$ such that*

$$\mathbf{E}\Big[|\hat{f}_1'(\hat{z})|^a \, 1_{\mathrm{Im}(\hat{f}_1(\hat{z}))\geq c}\Big] \geq c\Big(1 + (\hat{x}/\hat{y})^2\Big)^b \, \hat{y}^{\lambda-a}$$

holds for every $\hat{z} = \hat{x} + i\hat{y} \in \mathbb{H}$ satisfying $|\hat{z}| \leq c$.

Proof. As before, let $u := \log y$. Set $v = v(u) := \sinh^{-1}(x/y)$; that is, $x = y\sinh(v)$. Then Itô's formula gives

$$dv = -\frac{d\xi}{|z|} + \frac{1}{2}\,(8 + \kappa\,\nu)\,|z|^{-3}\,x\,dt\,.$$

The formula for dv in terms of $\hat{B}$ and du is

$$dv = \sqrt{\kappa/2}\,d\hat{B} - \frac{x}{4\,|z|}\,(8 + \kappa\,\nu)\,du\,. \tag{3.12}$$

Recall that $\psi(u)^a\,F\big(z(u)\big)$ is a martingale, since $F = \hat{F}$. Define a new probability measure by

$$\tilde{\mathbf{P}}[\mathcal{A}] := \frac{\mathbf{E}\Big[1_{\mathcal{A}}\,\big(1 + x(0)^2\big)^b \psi(0)^a\Big]}{\mathbf{E}\Big[\big(1 + x(0)^2\big)^b \psi(0)^a\Big]}\,,$$

for every event $\mathcal{A}$. This is the so-called Doob-transform (or h-transform) corresponding to the martingale $\psi(u)^a\,F\big(z(u)\big)$. Recall that if $\alpha(w)$ is a positive martingale for a diffusion process $dw = q_1(w,t)\,dB(t) + q_2(w,t)\,dt$, $t \leq t_1$, where $B(t)$ is Brownian motion, then for $t < t_1$, $dw = q_1(w,t)\,d\tilde{B}(t) + q_2(w,t)\,dt + q_1(w,t)\,\partial_w \log\alpha(w)\,q_1(w,t)\,dt$, where $\tilde{B}$ is Brownian motion with respect to the probability measure weighted by $\alpha(w(t_1))$; that is, the Doob transform of α. This follows, for example, from Girsanov's Formula [Dur96, §2.12]. We apply this with $w = (v,\psi)$ and u as the time parameter (in this case, q_1 and q_2 are vectors and $\partial_w \log\alpha(w)$ is a linear functional), and get by (3.12)

$$dv = \sqrt{\kappa/2}\,d\tilde{B} - \frac{x}{4\,|z|}\,(8 + \kappa)\,du + \frac{1}{2}\,\kappa\,\partial_v \log F\,du\,,$$

where $\tilde{B}(u)$ is Brownian motion under $\tilde{\mathbf{P}}$. This simplifies to

$$dv = \sqrt{\kappa/2}\, d\tilde{B} - \tilde{b}\, \tanh(v)\, du\,, \tag{3.13}$$

where

$$\tilde{b} := 2 - \big(b - \frac{1}{4}\big)\kappa > 1\,.$$

Thus, $v(u)$ is a very simple diffusion process. When $|v|$ is large, $\tanh(v)$ is close to $\operatorname{sign}(v)$, and v has almost constant drift $-\tilde{b}\operatorname{sign}(v)$, pushing it towards 0.

At this point, do not assume $|\hat{z}| \le c$, but only $|\hat{z}| \le 1$. Let $\Psi : [\hat{u}, 0] \to \mathbb{R}$ be the continuous function that is equal to 1 at 0, equal to $|\hat{u}|+1$ at $\hat{u}$, has slope $-\big(\tilde{b} \wedge 2 + 1\big)/2$ in the interval $[\hat{u}, \hat{u}/2]$ and has constant slope in the interval $[\hat{u}/2, 0]$, and let $\mathcal{A}$ be the event

$$\mathcal{A} := \Big\{\forall u \in [\hat{u}, 0], \quad |v(u)| \le \Psi(u)\Big\}\,.$$

Note that our assumption $|\hat{z}| \le 1$ implies that $|\hat{v}| \le |\hat{u}| + \log 2$. Since $\tilde{b} > 1$ it easily follows from (3.13) that there is a constant $c_1 = c_1(b, \kappa) > 0$ such that $\tilde{\mathbf{P}}[\mathcal{A}] > c_1$. In particular, c_1 does not depend on $\hat{z}$. This means

$$\mathbf{E}\Big[1_{\mathcal{A}}\, \big(1 + x(0)^2\big)^b\, \psi(0)^a\Big] \ge c_1\, \mathbf{E}\Big[\big(1 + x(0)^2\big)^b\, \psi(0)^a\Big] = c_1\, \Big(1 + (\hat{x}/\hat{y})^2\Big)^b\, \hat{y}^\lambda\,.$$

However, note that $v(0)$ and hence $x(0)$ are bounded on $\mathcal{A}$. Therefore, there is some constant $c_2 > 0$ such that

$$\mathbf{E}\Big[1_{\mathcal{A}}\, \big|g'_{T_0}(\hat{z})\big|^a\Big] \ge c_2\, \Big(1 + (\hat{x}/\hat{y})^2\Big)^b\, \hat{y}^{\lambda - a}\,. \tag{3.14}$$

We now estimate T_0 on the event $\mathcal{A}$. Recall that $T_0 \le 0$. From (3.4) we have $\partial_u T_u = -|z|^2/2$ and therefore on $\mathcal{A}$

$$\begin{aligned} T_0 &= -\int_{\hat{u}}^0 \frac{y^2}{2}\, du - \int_{\hat{u}}^0 \frac{x^2}{2}\, du = \frac{\hat{y}^2}{4} - \frac{1}{4} - \frac{1}{2}\int_{\hat{u}}^0 \sinh(v)^2\, e^{2\,u}\, du \\ &\ge -\frac{1}{4} - \frac{1}{2}\int_{\hat{u}}^0 e^{2\Psi(u) + 2u}\, du \ge -c_3\,, \end{aligned}$$

where $c_3 = c_3(b, \kappa) < \infty$ is some constant. That is, we have $T_0 \in [-c_3, 0]$ on $\mathcal{A}$. On the event $T_0 \ge -c_3$, we clearly have $\operatorname{Im}\big(g_{-c_3}(\hat{z})\big) \ge 1$ and also $\big|g'_{T_0}(\hat{z})\big| \le \big|g'_{-c_3}(\hat{z})\big|$, by (3.3). Consequently, Lemma 3.1 and (3.14) give

$$\mathbf{E}\Big[\big|\hat{f}'_{c_3}(\hat{z})\big|^a\, 1_{\operatorname{Im}(\hat{f}_{c_3}(\hat{z})) \ge 1}\Big] \ge c_4\, \Big(1 + (\hat{x}/\hat{y})^2\Big)^b\, \hat{y}^{\lambda - a}\,. \tag{3.15}$$

This is almost the result that we need. However, we want to replace c_3 by 1. For this, we apply scale invariance. In this procedure, the assumption $|\hat{z}| \le 1$ needs to be replaced by $|\hat{z}| \le 1/\sqrt{c_3}$. The proof of the lemma is now complete. □

It is not too hard to see that for every constant $C > 0$ the statement of the lemma can be strengthened to allow $c < |\hat{z}| < C$. The constant c must then also depend on C.

Remark 3.4. Suppose that we take

$$b = \frac{1}{4} + \frac{2\,\nu}{\kappa}\,,$$

and define a and λ using (3.6). Define $\tilde{\mathbf{P}}$ as in the proof of the lemma. Then as the proof shows, v becomes Brownian motion times $\sqrt{\kappa/2}$ under $\tilde{\mathbf{P}}$, since $\tilde{b}$ vanishes. This allows a detailed understanding of the distribution of $|g'_{T_0}(\hat{z})|$. For example, one can determine in this way the decay of $\mathbf{P}\big[\psi(0) > 1/2\big]$ as $\hat{y} \searrow 0$. It also follows that for such a, b, λ, and every $A \subset \mathbb{R}$ one can write down an explicit expression for $\mathbf{E}\Big[|g'_{T_0}(\hat{z})|^a\, 1_{x(0)\in A}\Big]$, since for every $v_0 \in \mathbb{R}$ we have

$$\Lambda\Big(\big(1 + (x/y)^2\big)^b\, y^\lambda\, \frac{1}{\sqrt{\nu\,u}}\, \exp\frac{(v_0 - v)^2}{\nu\,\kappa\,u}\Big) = 0\,.$$

This equation is the PDE facet of the fact that v is Brownian motion times $\sqrt{\kappa/2}$ under $\tilde{\mathbf{P}}$. However, these results will not be needed in the present paper.

3.2. *Derivative upper bounds at a fixed time t.* From Theorem 3.2 it is not hard to obtain estimates for $|\hat{f}'_t|$:

Corollary 3.5. *Let $b \in [0, 1 + 4\,\kappa^{-1}]$, and define λ and a by* (3.6) *with $\nu = 1$. There is a constant $C(\kappa, b)$, depending only on κ and b, such that the following estimate holds for all $t \in [0, 1]$, $y, \delta \in (0, 1]$ and $x \in \mathbb{R}$.*

$$\mathbf{P}\big[|\hat{f}'_t(x + i\,y)| \geq \delta\, y^{-1}\big] \leq C(\kappa, b)\,(1 + x^2/y^2)^b\,(y/\delta)^\lambda\,\vartheta(\delta, a - \lambda)\,, \tag{3.16}$$

where

$$\vartheta(\delta, s) = \begin{cases} \delta^{-s} & s > 0, \\ 1 + |\log\delta| & s = 0, \\ 1 & s < 0. \end{cases}$$

Proof. Note that $a \geq 0$. We assume that $\delta > y$, for otherwise the right-hand side is at least $C(\kappa, b)$, and we take $C(\kappa, b) \geq 1$. Take $z = x + i\,y$. By Lemma 3.1, $\hat{f}'_t(z)$ has the same distribution as $g'_{-t}(z)$. Let $u_1 := \log \operatorname{Im}\big(g_{-t}(x + iy)\big)$. Recall the notation T_u from (3.2). Note that

$$\big|g'_{-t}(z)/g'_{T_u(z)}(z)\big| \leq \exp\big(|u - u_1|\big), \tag{3.17}$$

since $\big|\partial_u \log|g'_t(z)|\big| \leq 1$, by (3.5). Moreover, it is clear that there is a constant c such that $u_1 \leq c$, because $t, y \leq 1$. Consequently,

$$\mathbf{P}\big[|g'_{-t}(z)| \geq O(1)\delta\, y^{-1}\big] \leq O(1) \sum_{j=\lceil \log y\rceil}^{0} \mathbf{P}\big[|g'_{T_j(z)}(z)| \geq \delta\, y^{-1}\big].$$

The Schwarz lemma implies that $y|g'(z)| \le \mathrm{Im}\big(g(z)\big)$ if $g : \mathbb{H} \to \mathbb{H}$. Therefore, the above gives

$$\mathbf{P}\big[|g'_{-t}(z)| \ge O(1)\,\delta\,y^{-1}\big] \le O(1) \sum_{j=\lceil \log \delta \rceil}^{0} \mathbf{P}\big[|g'_{T_j(z)}(z)| \ge \delta\,y^{-1}\big]. \tag{3.18}$$

By scale invariance, we have for all $j \in [\log y, 0]$

$$\mathbf{E}\big[y^a\,e^{-ja}\,|g'_{T_j(z)}(z)|^a\big] \le F_b\big(e^{-j}z\big)\,,$$

where $F = F_b$ is as in Theorem 3.2. Hence

$$\begin{aligned}\mathbf{P}\big[|g'_{T_j(z)}(z)| \ge \delta\,y^{-1}\big] &= \mathbf{P}\big[|g'_{T_j(z)}(z)|^a\,y^a\,\delta^{-a} \ge 1\big] \\ &\le \mathbf{E}\big[|g'_{T_j(z)}(z)|^a\,y^a\,\delta^{-a}\big] \le \delta^{-a}\,e^{j\,a}\,F_b\big(e^{-j}z\big)\,.\end{aligned}$$

Consequently, by (3.18) and Theorem 3.2,

$$\mathbf{P}\big[|g'_{-t}(z)| \ge O(1)\,\delta\,y^{-1}\big] \le O(1)\,(1 + x^2/y^2)^b\,\delta^{-a}\,y^\lambda \sum_{j=\lceil \log \delta \rceil}^{0} e^{j(a-\lambda)}.$$

The corollary follows. □

3.3. *Continuity of* $\hat{f}_t(0)$. In the deterministic example of nonlocally connected hulls described in the introduction, there is a time t_0 for which $\lim_{z\to 0} \hat{f}_{t_0}(z)$ does not exist (the limit set is the outer boundary of the prescribed compact set). Even when the SLE trace is a continuous path, it is not always true that $(z,t) \mapsto f_t(z)$ extends continuously to $\overline{\mathbb{H}} \times [0,\infty)$ (this is only true for simple paths). The next theorem shows that $\hat{f}_t(0) = f_t(\xi(t))$ exists as a radial limit and is continuous. Together with the result of Section 4 below, this is enough to show that the SLE trace is a path.

THEOREM 3.6. *Define*

$$H(y,t) := \hat{f}_t(i\,y), \qquad y > 0,\ t \in [0,\infty).$$

If $\kappa \ne 8$, *then a.s.* $H(y,t)$ *extends continuously to* $[0,\infty) \times [0,\infty)$.

Proof. Fix $\kappa \ne 8$. By scale invariance, it is enough to show continuity of H on $[0,\infty) \times [0,1)$. Given $j,k \in \mathbb{N}$, with $k < 2^{2j}$, let $R(j,k)$ be the rectangle

$$R(j,k) := [2^{-j-1}, 2^{-j}] \times [k\,2^{-2j}, (k+1)\,2^{-2j}]\,,$$

and set

$$d(j,k) := \mathrm{diam} H\big(R(j,k)\big)\,.$$

Take $b = (8+\kappa)/(4\kappa)$ and let a and λ be given by (3.6) with $\nu = 1$. Note that $\lambda > 2$. Set $\sigma_0 := (\lambda - 2)/\max\{a,\lambda\}$, and let $\sigma \in (0,\sigma_0)$. Our objective is to

prove

$$\sum_{j=0}^{\infty} \sum_{k=0}^{2^{2j}-1} \mathbf{P}\big[d(j,k) \geq 2^{-j\sigma}\big] < \infty\,. \tag{3.19}$$

Fix some such pair (j,k). The idea of the proof is to decompose the time interval $[k\,2^{-2j}, (k+1)\,2^{-2j}]$ into (random) sub-intervals using partition points $\hat{t}_N, \hat{t}_{N-1}, \ldots, \hat{t}_0$ such that the change of ξ on each interval is controlled. Then it is not hard to control $\hat{f}_t$ on each sub-interval. Now, the details. Set $t_0 = (k+1)\,2^{-2j}$. Inductively, let

$$t_{n+1} := \sup\big\{t < t_n : |\xi(t) - \xi(t_n)| = 2^{-j}\big\}\,,$$

and let N be the least $n \in \mathbb{N}$ such that $t_N \leq t_0 - 2^{-2j}$. Also set $t_\infty := t_0 - 2^{-2j} = k\,2^{-2j}$ and

$$\hat{t}_n := \max\{t_n, t_\infty\}\,.$$

Observe that the scaling property of Brownian motion shows that there is a constant $\rho < 1$, which does not depend on j or k, such that $\mathbf{P}[N > 1] = \rho$. Moreover, the Markov property implies that $\mathbf{P}\big[N \geq m+1 \;\big|\; N \geq m\big] \leq \rho$ for every $m \in \mathbb{N}$. In particular, $\mathbf{P}[N > m] \leq \rho^m$.

For every $s \in [0,\infty)$, the map $\hat{f}_s$ is measurable with respect to the σ-field generated by $\xi(t)$ for $t \in [0,s]$, while $\hat{t}_n$ is determined by $\xi(t)$ for $t \geq t_n$ (i.e., $\hat{t}_n$ is a stopping time for the reversed time filtration). Therefore, by the strong Markov property, for every $n \in \mathbb{N}$, $s \in [t_\infty, t_0]$ and $\delta > 0$,

$$\mathbf{P}\Big[|\hat{f}'_{\hat{t}_n}(i\,2^{-j})| > \delta \;\Big|\; \hat{t}_n = s\Big] = \mathbf{P}\Big[|\hat{f}'_s(i\,2^{-j})| > \delta\Big]\,,$$

and consequently,

$$\mathbf{P}\Big[|\hat{f}'_{\hat{t}_n}(i\,2^{-j})| > \delta \;\Big|\; \hat{t}_n > t_\infty\Big] \leq \sup_{s\in[0,1]} \mathbf{P}\Big[|\hat{f}'_s(i\,2^{-j})| > \delta\Big]\,.$$

This allows us to estimate,

$$\begin{aligned}
&\mathbf{P}\Big[\exists n \in \mathbb{N} \;:\; |\hat{f}'_{\hat{t}_n}(i\,2^{-j})| > \delta\Big] \\
&\quad\leq \mathbf{P}\big[|\hat{f}'_{t_\infty}(i\,2^{-j})| > \delta\big] + \sum_{n=0}^{\infty} \mathbf{P}[\hat{t}_n > t_\infty]\,\mathbf{P}\Big[|\hat{f}'_{\hat{t}_n}(i\,2^{-j})| > \delta \;\Big|\; \hat{t}_n > t_\infty\Big] \\
&\quad\leq \mathbf{E}[N+1] \sup_{s\in[0,1]} \mathbf{P}\big[|\hat{f}'_s(i\,2^{-j})| > \delta\big] \leq O(1) \sup_{s\in[0,1]} \mathbf{P}\big[|\hat{f}'_s(i\,2^{-j})| > \delta\big]\,.
\end{aligned}$$

By Corollary 3.5, and because $\sigma < (\lambda-2)/\max\{a,\lambda\}$, this gives

$$\mathbf{P}\Big[\exists n \in \mathbb{N} \;:\; \big|\hat{f}'_{\hat{t}_n}(i\,2^{-j})\big| > 2^j\,2^{-j\sigma}/j^2\Big] \leq O(1)\,2^{-2j}\,2^{-\varepsilon j}\,, \tag{3.20}$$

for some $\varepsilon = \varepsilon(\kappa) > 0$. Let S be the rectangle

$$S := \big\{x + iy : |x| \leq 2^{-j+3},\, y \in [2^{-j-1}, 2^{-j+3}]\big\}\,.$$

We claim that

$$H\big(R(j,k)\big) \subset \bigcup_{n=0}^{N} \hat{f}_{\hat{t}_n}(S)\,, \tag{3.21}$$

and

$$\forall n \in \mathbb{N} \qquad \hat{f}_{\hat{t}_n}(S) \cap \hat{f}_{\hat{t}_{n+1}}(S) \neq \emptyset\,. \tag{3.22}$$

Indeed, to prove (3.21), let $t \in [\hat{t}_{n+1}, \hat{t}_n]$ and $y \in [2^{-j-1}, 2^{-j}]$. Then we may write

$$\hat{f}_t(i\,y) = \hat{f}_{\hat{t}_{n+1}}\Big(g_{\hat{t}_{n+1}}\big(\hat{f}_t(i\,y)\big) - \xi(\hat{t}_{n+1})\Big). \tag{3.23}$$

We want to prove (3.21) by showing that $g_{\hat{t}_{n+1}}\big(\hat{f}_t(i\,y)\big) - \xi(\hat{t}_{n+1}) \in S$. Set $\varphi(s) := g_s\big(\hat{f}_t(i\,y)\big)$ for $s \le t$. Then $\varphi(t) = i\,y + \xi(t)$ and by (2.1)

$$\partial_s \varphi(s) = 2\,\big(\varphi(s) - \xi(s)\big)^{-1}\,.$$

Note that $\partial_s \mathrm{Im}\big(\varphi(s)\big) < 0$, and hence $\mathrm{Im}\big(\varphi(s)\big) \ge \mathrm{Im}\big(\varphi(t)\big) \ge 2^{-j-1}$. This then implies that $|\partial_s \varphi(s)| \le 2^{j+2}$, and therefore $|t - \hat{t}_{n+1}| \le 2^{-2j}$ gives $|\varphi(\hat{t}_{n+1}) - \varphi(t)| \le 2^{2-j}$. Since $|\xi(t) - \xi(\hat{t}_{n+1})| \le 2^{1-j}$, by (3.23) this gives $\hat{f}_t(i\,y) \in \hat{f}_{\hat{t}_{n+1}}(S)$ and verifies (3.21). We also have (3.22), because taking $t = \hat{t}_n$ in the above gives $\hat{f}_{\hat{t}_n}(i\,y) \in \hat{f}_{\hat{t}_{n+1}}(S)$. Since $|\hat{f}_t'(z)|/|\hat{f}_t'(i\,2^{-j})|$ is bounded by some constant if $z \in S$ (this follows from the Koebe distortion theorem, see [Pom92, §1.3]), we find that

$$\mathrm{diam}\big(\hat{f}_t(S)\big) \le O(1)\,2^{-j}\,|\hat{f}_t'(i\,2^{-j})|\,.$$

Therefore, the relations (3.21) and (3.22) give

$$\begin{aligned} d(j,k) &\le \sum_{n=0}^{N} \mathrm{diam}\big(\hat{f}_{\hat{t}_n}(S)\big) \le O(1)\,2^{-j} \sum_{n=0}^{N} \big|\hat{f}_{\hat{t}_n}'(i\,2^{-j})\big| \\ &\le O(1)\,2^{-j}\,N\,\max\Big\{\big|\hat{f}_{\hat{t}_n}'(i\,2^{-j})\big| : n = 0, 1, \ldots, N\Big\}\,. \end{aligned} \tag{3.24}$$

By (3.20), we get

$$\begin{aligned} \mathbf{P}\big[d(j,k) > 2^{-j\sigma}\big] &\le \mathbf{P}[N > j^2] + O(1)\,2^{-2j}\,2^{-\varepsilon j} \\ &\le \rho^{j^2} + O(1)\,2^{-2j}\,2^{-\varepsilon j} \le O(1)\,2^{-2j}\,2^{-\varepsilon j}\,, \end{aligned}$$

which proves (3.19).

From (3.19) we conclude that a.s. there are at most finitely many pairs $j, k \in \mathbb{N}$ with $k \le 2^{2j} - 1$ such that $d(j,k) > 2^{-j\sigma}$. Hence $d(j,k) \le C(\omega)\,2^{-j\sigma}$ for all j, k, where $C = C(\omega)$ is random (and the notation $C(\omega)$ is meant to suggest that). Let (y', t') and (y'', t'') be points in $(0,1)^2$. Let j_1 be the smallest integer larger than $\min\big\{-\log_2 y', -\log_2 y'', -\frac{1}{2}\log_2 |t' - t''|\big\}$. Note that a rectangle $R(j_1, k')$ that intersects the line $t = t'$ is adjacent to a rectangle

$R(j_1, k'')$ that intersects the line $t = t''$. Consequently, $\left|H(y', t') - H(y'', t'')\right| \le \sum_{j \ge j_1} (d(j, k'_j) + d(j, k''_j)) \le O(1)\, C(\omega)\, 2^{-\sigma j_1}$, where $R(j, k'_j)$ is a rectangle meeting the line $t = t'$ and $R(j, k''_j)$ is a rectangle meeting the line $t = t''$. This shows that for every $t_0 \in [0, 1)$ the limit of $H(y, t)$ as $(y, t) \to (0, t_0)$ exists, and thereby extends the definition of H to $[0, \infty) \times [0, 1)$. Since H is clearly continuous in $(0, \infty) \times [0, t)$, the proof is now complete. □

For $\kappa = 8$, we are unable to prove Theorem 3.6. However, a weaker result does follow easily, namely, a.s. for almost every $t \ge 0$ the limit $\lim_{y \searrow 0} \hat{f}_t(i\, y)$ exists.

Update. It follows from [LSW] and the results of the current paper that the Theorem holds also when $\kappa = 8$.

4. Reduction

The following theorem provides a criterion for hulls to be generated by a continuous path. In this section we do not assume that ξ is a (time scaled) Brownian motion.

THEOREM 4.1. *Let $\xi : [0, \infty) \to \mathbb{R}$ be continuous, and let g_t be the corresponding solution of* (2.1). *Assume that $\beta(t) := \lim_{y \searrow 0} g_t^{-1}(\xi(t) + iy)$ exists for all $t \in [0, \infty)$ and is continuous. Then g_t^{-1} extends continuously to $\overline{\mathbb{H}}$ and H_t is the unbounded connected component of $\mathbb{H} \setminus \beta([0, t])$, for every $t \in [0, \infty)$.*

In the proof, the following basic properties of conformal maps will be needed. Suppose that $g : D \to \mathbb{U}$ is a conformal homeomorphism. If $\alpha : [0, 1) \to D$ is a path such that the limit $l_1 = \lim_{t \nearrow 1} \alpha(t)$ exists, then $l_2 = \lim_{t \nearrow 1} g \circ \alpha$ exists, too. (However, if $\alpha : [0, 1) \to \mathbb{U}$ is a path such that the above limit exists, it does not follow that $\lim_{t \nearrow 1} g^{-1} \circ \alpha(t)$ exists. In other words, it is essential that the image of g is a nice domain such as $\mathbb{U}$.) Moreover, $\lim_{t \nearrow 1} g^{-1}(t l_2)$ exists and equals l_1. Consequently, if $\tilde{\alpha} : [0, 1) \to D$ is another path with $\lim_{t \nearrow 1} \tilde{\alpha}$ existing and with $\lim_{t \nearrow 1} g \circ \alpha(t) = \lim_{t \nearrow 1} g \circ \tilde{\alpha}(t)$, then $\lim_{t \nearrow 1} \alpha(t) = \lim_{t \nearrow 1} \tilde{\alpha}(t)$.

These statements are well known and easily established, for example with the notion of extremal length. See [Pom92, Prop. 2.14] for the first statement and [Ahl73, Th. 3.5] implies the second claim.

Proof. Let $S(t) \subset \overline{\mathbb{H}}$ be the set of limit points of $g_t^{-1}(z)$ as $z \to \xi(t)$ in $\mathbb{H}$. Fix $t_0 \ge 0$, and let $z_0 \in S(t_0)$. We want to show that $z_0 \in \overline{\beta([0, t_0))}$, and hence $z_0 \in \beta([0, t_0])$. Fix some $\varepsilon > 0$. Let

$$t' := \sup\left\{t \in [0, t_0] : K_t \cap \overline{D(z_0, \varepsilon)} = \emptyset\right\},$$

where $D(z_0, \varepsilon)$ is the open disk of radius ε about z_0. We first show that

$$\beta(t') \in \overline{D(z_0, \varepsilon)}. \tag{4.1}$$

Indeed, $D(z_0, \varepsilon) \cap H_{t_0} \neq \emptyset$ since $z_0 \in S(t_0)$. Let $p \in D(z_0, \varepsilon) \cap H_{t_0}$, and let $p' \in K_{t'} \cap \overline{D(z_0, \varepsilon)}$. Let p'' be the first point on the line segment from p to p' which is in $K_{t'}$. We want to show that $\beta(t') = p''$. Let L be the line segment $[p, p'')$, and note that $L \subset H_{t'}$. Hence $g_{t'}(L)$ is a curve in $\mathbb{H}$ terminating at a point $x \in \mathbb{R}$. If $x \neq \xi(t')$, then $g_t(L)$ terminates at points $x(t) \neq \xi(t)$ for all $t < t'$ sufficiently close to t'. Because $g_\tau(p'')$ has to hit the singularity $\xi(\tau)$ at some time $\tau \leq t'$, this implies $p'' \in K_t$ for $t < t'$ close to t'. This contradicts the definition of t' and shows $x = \xi(t')$. Now $\beta(t') = p''$ follows because the conformal map $g_{t'}^{-1}$ of $\mathbb{H}$ cannot have two different limits along two arcs with the same terminal point.

Having established (4.1), since $\varepsilon > 0$ was arbitrary, we conclude that $z_0 \in \overline{\beta([0, t_0))}$ and hence $z_0 \in \beta([0, t_0])$. This gives $S(t) \subset \beta([0, t])$ for all $t \geq 0$. We now show that H_t is the unbounded component of $\mathbb{H} \setminus \overline{\bigcup_{\tau \leq t} S(\tau)}$. First, H_t is connected and disjoint from $\overline{\bigcup_{\tau \leq t} S(\tau)}$. On the other hand, as the argument in the previous paragraph shows, $\partial H_t \cap \mathbb{H}$ is contained in $\overline{\bigcup_{\tau \leq t} S(\tau)}$. Therefore, H_t is a connected component of $\mathbb{H} \setminus \overline{\bigcup_{\tau \leq t} S(\tau)}$; that is, H_t is the unbounded connected component of $\mathbb{H} \setminus \beta([0, t])$. Since β is a continuous path, it follows from [Pom92, Th. 2.1] that g_t^{-1} extends continuously to $\overline{\mathbb{H}}$ (which also proves that $S(t) = \{\beta(t)\}$). □

5. Continuity

We have now established all the results needed to show that the SLE_κ trace is a continuous path a.s.

THEOREM 5.1 (Continuity). *Let $\kappa \neq 8$. The following statements hold almost surely. For every $t \geq 0$ the limit*

$$\gamma(t) := \lim_{z \to 0,\, z \in \mathbb{H}} \hat{f}_t(z)$$

exists, $\gamma : [0, \infty) \to \overline{\mathbb{H}}$ is a continuous path, and H_t is the unbounded component of $\mathbb{H} \setminus \gamma\big([0, t]\big)$.

We believe the theorem to be valid also for $\kappa = 8$. (This is stated as Conjecture 9.1.) Despite repeated efforts, the proof eluded us.

Update. This extension to $\kappa = 8$ is proved in [LSW].

Proof of Theorem 5.1. By Theorem 3.6, a.s. $\lim_{y \searrow 0} \hat{f}_t(iy)$ exists for all t and is continuous. Therefore we can apply Theorem 4.1, and the theorem follows. □

It follows from Theorem 5.1 that f_t extends continuously to $\overline{\mathbb{H}}$ a.s. The next result gives more information about the regularity of f_t on $\overline{\mathbb{H}}$. It neither follows from Theorem 5.1, nor does it imply 5.1.

THEOREM 5.2 (Hölder continuity). *For every $\kappa \neq 4$ there is some $h(\kappa) > 0$ such that for every bounded set $A \subset \mathbb{H}$ and every $t > 0$, a.s. $\hat{f}_t$ is Hölder continuous with exponent $h(\kappa)$ on A,*

$$|\hat{f}_t(z) - \hat{f}_t(z')| \leq C|z - z'|^{h(\kappa)} \qquad \text{for all} \quad z, z' \in A,$$

where $C = C(\omega, t, A)$ is random and may depend on t and A. Moreover, $\lim_{\kappa \searrow 0} h(\kappa) = \frac{1}{2}$ and $\lim_{\kappa \nearrow \infty} h(\kappa) = 1$.

Since $f_t(z) - z \to 0$ as $z \to \infty$, it easily follows that for every t a.s.

$$|\hat{f}_t(z) - \hat{f}_t(z')| \leq C(\omega, t) \max(|z - z'|, |z - z'|^{h(\kappa)}). \tag{5.1}$$

We do not believe that the theorem holds for $\kappa = 4$, for then the trace is a simple path which "almost" touches itself.

Update. For $\kappa \leq 4$, the fact that γ is a simple path in $\mathbb{H}$, Theorem 6.1 below, implies $h(\kappa) \leq 1/2$. Thus the estimate $\lim_{\kappa \searrow 0} h(\kappa) = \frac{1}{2}$ is best possible. On the other hand, this nonsmoothness of $\hat{f}_t$ is localized at $\hat{f}_t^{-1}(0)$: Joan Lind (manuscript in preparation) has shown that the Hölder exponent $h(\kappa)$ of $(\hat{f}_t(\sqrt{z}))^2$ satisfies $\lim_{\kappa \searrow 0} h(\kappa) = 1$.

Proof. Fix $\kappa \neq 4$ and $t > 0$. By scaling, we may assume $0 < t \leq 1$ and $A = [-1, 1] \times (0, 1]$. Denote

$$z_{j,n} = (j + i)2^{-n}, \quad 0 \leq n < \infty, \quad -2^n < j < 2^n.$$

We will first show that there is an $h = h(\kappa) > 0$ such that a.s.

$$|\hat{f}_t'(z_{j,n})| \leq C(\omega, t) 2^{n(1-h)}, \qquad \forall j, n. \tag{5.2}$$

Using Corollary 3.5 with $\delta = 2^{-nh}$ we have

$$\mathbf{P}\big[|\hat{f}_t'(z_{j,n})| \geq 2^{n(1-h)}\big] \leq C(\kappa, b)\,(1 + 2^{2\,n})^b\, 2^{-n(1-h)\lambda}\, \vartheta(2^{-nh}, a - \lambda).$$

Hence

$$\sum_{n=0}^{\infty} \sum_{j=-2^n}^{2^n} \mathbf{P}\big[|\hat{f}_t'(z_{j,n})| \geq 2^{n(1-h)}\big] < \infty$$

provided that

$$1 + 2b - (1 - h)\lambda < 0 \quad \text{and} \quad a - \lambda \leq 0,$$

or that

$$1 + 2b - \lambda + ah < 0 \quad \text{and} \quad a - \lambda \geq 0.$$

If $0 < \kappa \leq 12$, $b = 1/4 + 1/\kappa$ and $h < (\kappa - 4)^2/((\kappa + 4)(\kappa + 12))$, the first condition is satisfied. For $\kappa > 12$, $b = 4/\kappa$ and $h < 1/2 - 4/\kappa$ the second condition is satisfied, and (5.2) follows.

To see that one can actually achieve $h(\kappa) \to 1/2$ as $\kappa \searrow 0$, set $b := -1/2 + \sqrt{1/2 + 2/\kappa}$ for $0 < \kappa < 4$, and let h be smaller than but close to $(\lambda - 1 - 2b)/\lambda$. To get $\lim_{\kappa \nearrow \infty} h(\kappa) = 1$, take $b := (\sqrt{2\kappa}\sqrt{\kappa^2 + 10\kappa + 16} - 2\kappa)/(\kappa^2 + 4\kappa)$ for $\kappa \geq 2(3 + \sqrt{17})$ and let h be smaller than but close to $(\lambda - 1 - 2b)/a$.

From the Koebe distortion Theorem and (5.2) we obtain

$$|\hat{f}_t'(z)| \leq O(1)\, C(\omega, t)\, y^{h-1}$$

for all $z \in A$. It is well-known and easy to see, by integrating $|f'|$ over the hyperbolic geodesic from z to z' (similarly to the end of the proof of Theorem 3.6), that this implies Hölder continuity with exponent h on A. The theorem follows. □

The following corollary is an immediate consequence of Theorem 5.2. In Section 8 below, we will present more precise estimates.

COROLLARY 5.3. *For $\kappa \neq 4$ and every t, the Hausdorff dimension of ∂K_t a.s. satisfies*

$$\dim \partial K_t < 2.$$

In particular, a.s. $\operatorname{area} \partial K_t = 0$.

Proof. By [JM95] (see also [KR97] for an easier proof), the boundary of the image of a disk under a Hölder continuous conformal map has Hausdorff dimension bounded away from 2. Consider the conformal map $T(z) = (z - i)/(z + i)$ from $\mathbb{H}$ onto $\mathbb{U}$. By (5.1), $T \circ f_t \circ T^{-1}$ is a.s. Hölder continuous in $\mathbb{U}$. Since T preserves Hausdorff dimension, the corollary follows. □

6. Phases

In this section, we will investigate the topological behavior of SLE, and will identify three very different phases for the parameter κ, namely, $[0, 4]$, $(4, 8)$, and $[8, \infty)$.

The following result was conjectured in [Sch00]. There, it was proved that for $\kappa > 4$, a.s. K_t is not a simple path. The proof was based on the calculation of the harmonic measure $F(x)$ below, which we will repeat here for the convenience of the reader.

THEOREM 6.1. *In the range $\kappa \in [0, 4]$, the* SLE_κ *trace γ is a.s. a simple path and $\gamma[0, \infty) \subset \mathbb{H} \cup \{0\}$.*

LEMMA 6.2. *Let $\kappa \in [0,4]$, and let γ be the* SLE_κ *trace. Then a.s.*

$$\gamma[0,\infty) \subset \mathbb{H} \cup \{0\}\,.$$

Set $Y_x(t) := g_t(x) - \xi(t)$, $x \in \mathbb{R}$, $t \geq 0$. Then $Y_x(t)/\sqrt{\kappa}$ is a Bessel process of dimension $1 + 4/\kappa$. The basic theory of Bessel processes then tells us that a.s. $Y_x(t) \neq 0$ for $t \geq 0$, $x \neq 0$ and $\kappa \in [0,4]$, which amounts to $x \neq K_t$. However, we will prove this for the convenience of the reader unfamiliar with Bessel processes.

Proof. Let $b > a > 0$. For $x \in [a,b]$, let

$$T = T_x := \inf\{t \geq 0 : Y_x(t) \notin (a,b)\}.$$

Set

$$\hat{F}(x) := \begin{cases} x^{(\kappa-4)/\kappa} & \kappa \neq 4, \\ \log(x) & \kappa = 4\,, \end{cases}$$

and

$$F(x) := \frac{\hat{F}(x) - \hat{F}(a)}{\hat{F}(b) - \hat{F}(a)}\,.$$

Itô's formula shows that $F(Y_x(t \wedge T))$ is a local martingale, and since F is bounded in $[a,b]$, this is a martingale. Consequently, the optional sampling theorem gives $F(x) = \mathbf{E}\big[F(Y_x(T)\big] = \mathbf{P}\big[Y_x(T) = b\big]$.

Note that $F(x) \to 1$ when $a \searrow 0$. (That's where the assumption $\kappa \leq 4$ is crucial.) Hence, given $x > 0$, a.s. for all $b > x$, there is some $s > 0$ such that $g_t(x)$ is well defined for $t \in [0,s]$, $\inf\big\{Y_x(t) : t \in [0,s]\big\} > 0$ and $Y_x(s) = b$. Note that the Itô derivative of $Y_x(t)$ (with respect to t) is

$$dY_x = (2/Y_x)\,dt + d\xi\,.$$

It follows easily that a.s. $Y_x(t)$ does not escape to ∞ in finite time. Observe that if $x' > x$, then $Y_{x'}(t) \geq Y_x(t)$. Therefore, a.s. for every $x > 0$ we have $Y_x(t)$ well defined and in $(0,\infty)$ for all $t \geq 0$. This implies that a.s. for every $x > 0$ and every $s > 0$ there is some neighborhood N of x in $\mathbb{C}$ such that the differential equation (2.1) has a solution in the range $z \in N$, $t \in [0,s]$. This proves that a.s. $\gamma[0,\infty)$ does not intersect $(0,\infty)$. The proof that it a.s. does not intersect $(-\infty,0)$ is the same. □

Proof of Theorem 6.1. Let $t_2 > t_1 > 0$. The theorem will be established by proving that $\gamma[0,t_1] \cap \gamma[t_2,\infty) = \emptyset$. Let $s \in (t_1,t_2)$ be rational, and set

$$\hat{g}_t(z) := g_{t+s} \circ g_s^{-1}\big(z + \xi(s)\big) - \xi(s).$$

By Proposition 2.1 $(\hat{g}_t : t \geq 0)$ has the same distribution as $(g_t : t \geq 0)$. Let $\hat{\gamma}_s(t)$ be the trace for the collection $(\hat{g}_t : t \geq 0)$; that is,

$$\hat{\gamma}_s(t) = g_s \circ g_{t+s}^{-1}\big(\xi(t+s)\big) - \xi(s) = g_s \circ \gamma(t+s) - \xi(s)\,.$$

We know from Lemma 6.2 that a.s. for all rational $s \in [t_1, t_2]$,

$$\hat{\gamma}_s[0,\infty) \subset \mathbb{H} \cup \{0\}\,.$$

Consequently, by applying g_s^{-1}, we conclude that a.s. for every rational $s > 0$,

$$\gamma[s,\infty) \cap (\mathbb{R} \cup K_s) = \{\gamma(s)\}\,.$$

Since $g_{t_2} \circ g_{t_1}^{-1}$ is not the identity, it follows from Theorem 5.1 that there is some $s \in (t_1, t_2)$ such that $\gamma(s) \notin \mathbb{R} \cup K_{t_1}$, and this holds for an open set of $s \in (t_1, t_2)$. In particular, there is some rational $s \in (t_1, t_2)$ with $\gamma(s) \notin \mathbb{R} \cup K_{t_1}$. Since $\gamma[0, t_1] \subset K_{t_1} \cup \mathbb{R}$, it follows that $\gamma[0, t_1] \cap \gamma[t_2, \infty) = \emptyset$, proving the theorem. □

Recall the definition of the hypergeometric function in $|z| < 1$,

$$ {}_2F_1(\eta_0, \eta_1, \eta_2, z) := \sum_{n=0}^{\infty} \frac{(\eta_0)_n\,(\eta_1)_n}{(\eta_2)_n\,n!}\,z^n\,, \tag{6.1}$$

where $(\eta)_n := \prod_{j=0}^{n-1}(\eta + j)$. Let

$$\hat{G}(x + i\,y) = \hat{G}_{a,\kappa}(x + i\,y) := {}_2F_1\big(\eta_0, \eta_1, 1/2, x^2/(x^2+y^2)\big)\,,$$

where

$$\eta_j = \eta_j(a,\kappa) := \frac{1}{2} - \frac{2}{\kappa} - (-1)^j\,\frac{\sqrt{32\,a\,\kappa + (2\,\kappa - 8)^2}}{4\,\kappa}\,.$$

LEMMA 6.3. *Let $\kappa > 0$ and $z = x + iy \in \mathbb{H}$. Then the limit*

$$Z(z) := \lim_{t \nearrow \tau(z)} y\,\big|g_t'(z)\big|/\mathrm{Im}\big(g_t(z)\big)$$

exists a.s. We have $Z(z) = \infty$ a.s. if $\kappa \geq 8$ and $Z(z) < \infty$ a.s. if $\kappa < 8$. Moreover, for all $a \in \mathbb{R}$ we have

$$\mathbf{E}\big[Z(z)^a\big] = \begin{cases} \hat{G}_{a,\kappa}(z)/\hat{G}_{a,\kappa}(1) & a < 1 - \kappa/8,\ \kappa < 8, \\ \infty & a \geq 1 - \kappa/8,\ a > 0, \\ 0 & a < 0,\ \kappa \geq 8. \end{cases}$$

Before proving the lemma, let us discuss its geometric meaning. For $z_0 \in \mathbb{H}$ and $t < \tau(z_0)$ we have $z_0 \in H_t$. Set $r_t := \mathrm{dist}(z_0, \partial H_t)$, and let $\phi : \mathbb{H} \to \mathbb{U}$ be a conformal homeomorphism satisfying $\phi\big(g_t(z_0)\big) = 0$. Note that $\Big|\phi'\big(g_t(z_0)\big)\Big| = \big(2\,\mathrm{Im}\,g_t(z_0)\big)^{-1}$. Hence, the Schwarz lemma applied to the map $z \mapsto \phi \circ g_t(z_0 + r_t\,z)$ proves $r_t\,\big|g_t'(z_0)\big| \leq 2\,\mathrm{Im}\,g_t(z_0)$. On the other hand, the Koebe 1/4 Theorem applied to the function $g_t^{-1} \circ \phi^{-1}$ implies that $\mathrm{Im}\,g_t(z_0) \leq 2\,r_t\,|g_t'(z_0)|$. Therefore, since $\lim_{t \nearrow \tau(z_0)} r_t = \mathrm{dist}\big(z_0, \gamma[0,\infty) \cup \mathbb{R}\big)$,

$$Z(z_0)^{-1}\,\mathrm{Im}\,z_0/2 \leq \mathrm{dist}\big(z_0, \gamma[0,\infty) \cup \mathbb{R}\big) \leq 2\,Z(z_0)^{-1}\,\mathrm{Im}\,z_0\,. \tag{6.2}$$

In particular, the lemma tells us that $\gamma[0,\infty)$ is dense in $\mathbb{H}$ if and only if $\kappa \geq 8$. (For $\kappa = 8$ we have not proven that γ is a path. In that case, $\gamma[0,\infty)$ is defined as the union over t of all limit points of sequences $\hat{f}_t(z_j)$ where $z_j \to 0$ in $\mathbb{H}$. By an argument in the proof of Theorem 4.1, $\overline{\bigcup_{t\geq 0} \partial K_t} = \overline{\gamma[0,\infty)}$.)

The way one finds the function $\hat{G}$ is explained within the proof below, and is similar to the situation in Theorem 3.2.

Proof. Let $G(z) = G_{a,\kappa}(z) := \mathbf{E}\big[Z(z)^a\big]$. Fix some $\hat{z} = \hat{x} + i\,\hat{y} \in \mathbb{H}$, and abbreviate $Z := Z(\hat{z})$. Let $z_t = x_t + i\,y_t := g_t(\hat{z}) - \xi(t)$. From (3.9) we find that

$$\partial_t \log\big|g_t'(z)/y_t\big| = 4\,y_t^2\,|z_t|^{-4}\,,$$

which implies that

$$Z = \exp\Big(\int_0^{\tau(\hat{z})} 4\,y_t^2\,|z_t|^{-4}\,dt\Big) = \exp\Big(\int_0^{\tau(\hat{z})} \frac{-2\,y_t^2}{x_t^2+y_t^2}\,d\log y_t\Big)\,. \tag{6.3}$$

In particular, the limit in the definition of Z exists a.s. and $G(\hat{z}) = \mathbf{E}[Z^a]$ (possibly $+\infty$).

Let $\mathcal{F}_t$ denote the σ-field generated by $\big(\xi(s) : s \leq t\big)$. A direct application of Itô's formula shows that

$$M_t := \big(\hat{y}\,|g_t'(\hat{z})|/y_t\big)^a\,\hat{G}(z_t)$$

is a local martingale[1].

As before, let $u_t = \log y_t$. Note that $w(u) = w_t := x_t/y_t$ is a time-homogeneous diffusion process as a function of u. It is easy to see that $\tau(\hat{z})$ is a.s. the time at which u hits $-\infty$. (For example, see the argument following (3.2).)

If $t_1 < \tau(\hat{z})$ is a stopping time, then by (6.3) we may write

$$\log Z(\hat{z}) = \int_0^{t_1} + \int_{t_1}^{\tau(\hat{z})} 4\,y_t^2\,|z_t|^{-4}\,dt\,. \tag{6.4}$$

Note that if $s > 0$ then the distribution of $Z(\hat{z})$ is the same as that of $Z(s\,\hat{z})$, by scale invariance. It is clear that for every $z' \in \mathbb{H}$ there is positive probability that there will be some first time $t_1 < \tau(\hat{z})$ such that $z_{t_1}/|z_{t_1}| = z'/|z'|$. Conditioned on $t_1 < \tau(\hat{z})$ and on $\mathcal{F}_{t_1}$, the second integral in (6.4) has the same distribution as $\log Z(z')$, by the strong Markov property and scale invariance. Consequently, we find that $\mathbf{P}\big[Z(\hat{z}) > s\big] \geq c\,\mathbf{P}\big[Z(z') > s\big]$ holds for every s, where $c = c(\hat{z}, z', \kappa) > 0$. In fact, later we will need the following stronger

[1]The Markov property can be used to show that $\big(\hat{y}\,|g_t'(\hat{z})|/y_t\big)^a\,G(z_t)$ is a local martingale. One then assumes that G is smooth and applies Itô's formula, to obtain a PDE satisfied by G. Scale invariance can then be used to transform the PDE to the ODE (6.9) appearing below. Then one obtains $\hat{G}$ as a solution. That's how the function $\hat{G}$ was arrived at.

statement (which is not needed for the present lemma, but we find it convenient to give the argument here). Let $\overline{w} \in (0,\infty)$. Then assuming $|w_0| \le \overline{w}$, there is a $c = c(\overline{w},\kappa) > 0$ such that

$$\forall s > 0\,, \qquad \mathbf{P}\big[Z(i) > s\big] \ge \mathbf{P}\big[Z(\hat{z}) > s\big] \ge c\,\mathbf{P}\big[Z(i) > 2\,s\big]\,. \tag{6.5}$$

The right-hand inequality is proved by taking t_1 to be the first time smaller than $\tau(z)$ such that $\operatorname{Re} z_{t_1} = 0$ and the first integral in (6.4) is greater than $\log 2$, on the positive probability event that there is such a time. The left-hand inequality is clear, for if $|\operatorname{Re} g_t(i)/\operatorname{Im} g_t(i)|$ never hits $|w_0|$, then we have $Z(i) = \infty$, by (6.3).

Note that $\hat{G}(i) = 1$. Assume, for now, that $G(i) < \infty$. We claim that in this case $\hat{G}(x+i) > 0$ for each $x \in \mathbb{R}$. Let $s_0 > 0$, let $T := \inf\{t \ge 0 : |w_t| = s_0\}$. Since $w(u)$ is time-homogeneous, clearly, $T < \tau(\hat{z})$. If $\hat{G}(x+i) \le 0$ for some $x \in \mathbb{R}$, we may choose $s_0 := \inf\{x > 0 : \hat{G}(x+i) = 0\}$. Then $M_T = 0$ if $|w_0| \le s_0$. Note that $1 \le \hat{y}\,|g_T'(\hat{z})|/y_T < Z$. Since M_t is a local martingale on $t < \tau(\hat{z})$, there is an increasing sequence of stopping times $t_n < \tau(\hat{z})$ with $\lim_n t_n = \tau(\hat{z})$ a.s. such that $M_{t\wedge t_n}$ is a martingale. Set $\bar{t}_n := T \wedge t_n$. The optional sampling theorem then gives for $\hat{z} = i$

$$\begin{aligned} 1 = \hat{G}(\hat{z}) = M_0 = \mathbf{E}[M_{\bar{t}_n}] = \mathbf{E}\big[\big(\hat{y}\,|g_{\bar{t}_n}'(\hat{z})|/y_{\bar{t}_n}\big)^a\,\hat{G}(z_{\bar{t}_n})\big] \\ \le \mathbf{E}\big[\max\{1, Z^a\}\,\hat{G}(z_{\bar{t}_n})\big]\,. \end{aligned}$$

Since $\hat{G}(z_{t\wedge T})$ is bounded, $\lim_{n\to\infty} \hat{G}(z_{\bar{t}_n}) = 0$ a.s., and $\mathbf{E}[Z^a] < \infty$, we have $\lim_{n\to\infty} \mathbf{E}\big[\max\{1, Z^a\}\,\hat{G}(z_{\bar{t}_n})\big] = 0$, contradicting the above inequality. Consequently, $\hat{G}(x+i) > 0$ for every $x \in \mathbb{R}$.

We maintain the assumption $G(i) < \infty$. By (6.5), this implies $G(z) < \infty$ for all $z \in \mathbb{H}$. Let $s_1, s_2, \ldots$ be a sequence in $(0,\infty)$ with $s_n \to \infty$ such that $L := \lim_n \hat{G}(s_n + i)$ exists (allowing $L = \infty$). Now set $T_m := \inf\{t \ge 0 : |w_t| = s_m\}$. The argument of the previous paragraph shows that $M_0 = \mathbf{E}\big[M_{T_m\wedge t_n}\big] \to \mathbf{E}\big[M_{T_m}\big]$, as $n \to \infty$. Therefore,

$$\hat{G}(\hat{z}) = M_0 = \mathbf{E}\big[M_{T_m}\big] \to L\,G(\hat{z})\,, \qquad m \to \infty\,,$$

unless $L = \infty$, which happens if and only if $G(\hat{z}) = 0$. Moreover, L does not depend on the particular sequence s_n, and we may write $\hat{G}(1)$ instead of L. Thus, when $G(i) < \infty$,

$$G(\hat{z}) = \hat{G}(\hat{z})/\hat{G}(1)\,. \tag{6.6}$$

When $a = 1 - \kappa/8$, $\hat{G}$ simplifies to $(y/|z|)^{(8-\kappa)/\kappa}$, with $\hat{G}(1) = \infty$ if $\kappa > 8$ and $\hat{G}(1) = 0$ if $\kappa < 8$. It follows that $G(z) = \infty$ if $a = 1 - \kappa/8$ and $\kappa < 8$. Clearly, $G(z) = \infty$ also if $a > 1 - \kappa/8$, and $\kappa < 8$. Similarly, we have $G(z) = 0$ if $\kappa > 8$ and $a = 1 - \kappa/8$. This implies that $Z = \infty$ a.s. when $\kappa > 8$, which implies $G = 0$ when $\kappa > 8$ and $a < 0$. There are various ways to argue that $Z = \infty$ a.s. when $\kappa = 8$. In particular, it is easy to see that $\hat{G}_{\frac{1}{4},8}(1) = \infty$

(if a_n denotes the n-th coefficient of ${}_2F_1(\frac{1}{4},\frac{1}{4},\frac{1}{2},z)$, then $(n+1)a_{n+1} = na_nb_n$ with $b_n = (n+1/4)^2/(n(n+1/2))$ and it follows from $\prod_n b_n > 0$ that na_n is bounded away from zero) so that the above argument applies. It remains to prove that $G(i) < \infty$ when $a < 1 - \kappa/8$ and $\kappa < 8$.

Assume now that $\kappa < 8$ and $a < 1 - \kappa/8$ and a is sufficiently close to $1 - \kappa/8$ so that $\eta_1 > 0$. We prove below that in this range

$$\inf_{z\in\mathbb{H}} \hat{G}(z) > 0\,. \tag{6.7}$$

Since $\hat{G}(\hat{z}) = M_0 = \mathbf{E}\big[M_{t_n}\big]$ when $\hat{z} = i$, we immediately get $G(i) < \infty$ from (6.7). Since $G(i) < \infty$ for some a implies $G(i) < \infty$ for any smaller a, it only remains to establish (6.7) in the above specified range.

To prove (6.7), first note that $\eta_0 \le \eta_1$, $0 < \eta_1 < 1/2$ and $\eta_0 + \eta_1 = 1 - 4/\kappa < 1/2$. By [EMOT53, 2.1.3.(14)], when $\eta_2 > \eta_1 + \eta_0$ and $\eta_2 > \eta_1 > 0$, the right-hand side of (6.1) converges at $z = 1$ and is equal to

$$\frac{\Gamma(\eta_2)\,\Gamma(\eta_2 - \eta_1 - \eta_0)}{\Gamma(\eta_2 - \eta_1)\,\Gamma(\eta_2 - \eta_0)}\,. \tag{6.8}$$

This gives $\hat{G}(1) > 0$. Now write

$$g(x) := \hat{G}_{a,\kappa}(x+i)/\hat{G}_{1-\kappa/8,\kappa}(x+i) = (1+x^2)^{(8-\kappa)/(2\kappa)}\,\hat{G}_{a,\kappa}(x+i)\,.$$

Since $\hat{G}$ satisfies

$$\frac{4\,a\,y^2}{|z|^4}\,\hat{G} + \frac{\kappa}{2}\,\partial_x^2\hat{G} + \frac{4x}{|z|^2}\,\partial_x\hat{G} = 0 \tag{6.9}$$

for $y = 1$, it follows that

$$(\kappa - 8 + 8a)\,g + 2\,(\kappa - 4)\,x\,(1+x^2)\,g' + \kappa\,(1+x^2)^2\,g'' = 0\,. \tag{6.10}$$

It is immediate that $g(0) = 1$, $g'(0) = 0$ and $g''(0) > 0$, e.g., by differentiation or by considering (6.9). Consequently, $g > 1$ for small $|x| > 0$. Suppose that there is some $x \neq 0$ such that $g(x) = 1$. Then there must be some $x_0 \neq 0$ such that $g(x_0) > 1$ and g has a local maximum at x_0. Then $g'(x_0) = 0 \ge g''(x_0)$. But this contradicts (6.10), giving $g > 1$ for $x \neq 0$. Together with $\hat{G}(1) > 0$, this implies (6.7), and completes the proof. □

We say that $z \in \overline{\mathbb{H}}$ is *swallowed* by the SLE if $z \notin \gamma[0,\infty)$ but $z \in K_t$ for some $t \in (0,\infty)$. The smallest such t will be called the *swallowing time* of z.

THEOREM 6.4. *If $\kappa \in (4,8)$ and $z \in \overline{\mathbb{H}} \setminus \{0\}$, then a.s. z is swallowed. If $\kappa \notin (4,8)$, then a.s. no $z \in \overline{\mathbb{H}} \setminus \{0\}$ is swallowed.*

LEMMA 6.5. *Let $z \in \overline{\mathbb{H}} \setminus \{0\}$. If $\kappa > 4$, then $\mathbf{P}\big[\tau(z) < \infty\big] = 1$.*

In fact, in this range $\kappa > 4$, the stronger statement that a.s. for all $z \in \overline{\mathbb{H}}$ we have $\tau(z) < \infty$ holds as well. This follows from Theorem 7.1 (and its extension to $\kappa = 8$ based on [LSW]).

Proof. Let $b := 1 - 4/\kappa$, and $\theta := \exp\bigl(i\pi(1-b)/2\bigr)$. Set

$$h(z) := \mathrm{Im}\bigl(\theta\, z^b\bigr), \qquad z \in \overline{\mathbb{H}}\,.$$

A direct calculation shows that the drift term in Itô's formula for $\bigl(g_t(z)-\xi(t)\bigr)^b$ is zero. Consequently, $h\bigl(g_t(z)-\xi(t)\bigr)$ is a local martingale. Note that $\kappa > 4$ implies that $h(z) \to \infty$ as $z \to \infty$ in $\overline{\mathbb{H}}$, that $h(z) > 0$ in $\overline{\mathbb{H}} \setminus \{0\}$ and $h(0) = 0$.

Fix some $z_0 \in \mathbb{H}$ and some $\varepsilon > 0$. For $t \geq 0$ set $z_t := g_t(z_0) - \xi(t)$. We shall prove that there is some $T = T(z_0, \varepsilon)$ such that $\mathbf{P}\bigl[z_0 \in K_T\bigr] > 1 - \varepsilon$. Indeed, let $R > 1$ be sufficiently large so that $\varepsilon \inf\{h(z) : |z| = R\} > 2h(z_0)$. Let

$$\tau_R := \inf\Bigl(\bigl\{t \in [0, \tau(z_0)) : |z_t| \geq R\bigr\} \cup \bigl\{\tau(z_0)\bigr\}\Bigr)\,.$$

We claim that if $T = T(z_0, \varepsilon)$ is sufficiently large, then $\mathbf{P}\bigl[\tau_R > T\bigr] < \varepsilon/2$. Let $t \geq 0$ and let $\lceil t \rceil := \inf\{k \in \mathbb{Z} : k \geq t\}$. Say that t is a *sprint time* if $|\xi(t) - \xi(\lceil t \rceil)| = R$ and there is no $s \in [t, \lceil t \rceil]$ such that $|\xi(s) - \xi(\lceil t \rceil)| = R$. Note that for every integer $k \geq 0$ there is positive probability that there is a sprint time in $[k, k+1]$ and this event is independent from $\bigl(\xi(t) : t \in [0,k]\bigr)$. Therefore, there are infinitely many sprint times a.s. Let $t_0 < t_1 < t_2 < \cdots$ be the sequence of sprint times. Let $\mathcal{A}_j$ be the event that $\xi(\lceil t_j \rceil) > \xi(t_j)$. Observe that for all $j \in \mathbb{N}$,

$$\mathbf{P}\bigl[\mathcal{A}_j \bigm| t_j,\ \mathcal{F}_{t_j}\bigr] = 1/2\,, \tag{6.11}$$

where $\mathcal{F}_t$ denotes the σ-field generated by $\bigl(\xi(s) : s \in [0,t]\bigr)$, because given t_j and $\bigl(\xi(t) : t \in [0, t_j]\bigr)$, the distribution of $\bigl(\xi(t) : t > t_j\bigr)$ is the same as that of $\bigl(2\xi(t_j) - \xi(t) : t > t_j\bigr)$, by reflection symmetry. Let $x_t := \mathrm{Re}(z_t)$, and note that

$$dx_t = 2x_t\, |z_t|^{-2}\, dt - d\xi(t)\,.$$

Consequently, for $s \geq t_j$,

$$x_s = x_{t_j} + 2\int_{t_j}^{s} x_t\, |z_t|^{-2}\, dt - \xi(s) + \xi(t_j)\,. \tag{6.12}$$

On $\mathcal{A}_j$ we have $\xi(s) - \xi(t_j) > 0$ for all $s \in [t_j, \lceil t_j \rceil]$. Hence, it follows from (6.12) that on $\mathcal{A}_j$ and assuming $x_{t_j} \leq 0$, we have $x_s \leq 0$ for all $s \in [t_j, \lceil t_j \rceil]$ and moreover $x_{\lceil t_j \rceil} < -R$. Similarly, if $x_{t_j} \geq 0$ and $\neg\mathcal{A}_j$ holds, then $x_{\lceil t_j \rceil} > R$. Hence, by (6.11) we have $\mathbf{P}\bigl[\tau_R > \lceil t_j \rceil\bigr] \leq 2^{-j}$, which implies that we may choose T such that $\mathbf{P}[\tau_R > T] < \varepsilon/2$. Set $\tau^* := \tau_R \wedge T$, where T is as above.

Since $h(z_t)$ is a local martingale, and $h(z_{t\wedge\tau^*})$ is bounded, $h(z_{t\wedge\tau^*})$ is a martingale. Therefore,

$$h(z_0) = \mathbf{E}[h(z_{\tau^*})] \geq \inf\{h(z) : |z| = R\}\, \mathbf{P}\bigl[|z_{\tau^*}| = R\bigr]\,.$$

By our choice of R, this gives $\mathbf{P}\bigl[|z_{\tau^*}| = R\bigr] < \varepsilon/2$. Since $\mathbf{P}[\tau_R > T] < \varepsilon/2$, it follows that $\mathbf{P}[\tau(z_0) \leq T] \geq 1 - \varepsilon$, and the lemma is thereby established. □

We now prove a variant of Cardy's formula. (See [Car] for a survey of Cardy's formula.) A proof of (a generalized) Cardy's formula for SLE_κ appears in [LSW01a]. However the following variant appears to be different, except in the special case $\kappa = 6$, in which they are equivalent due to the locality property of SLE_6. (See [LSW01a] for an explanation of the locality property.)

LEMMA 6.6. *Fix $\kappa \in (4,8)$ and let*

$$X := \inf\big([1,\infty) \cap \gamma[0,\infty)\big)\,.$$

Then for all $s \geq 1$,

$$\mathbf{P}[X \geq s] = \frac{4^{(\kappa-4)/\kappa}\sqrt{\pi}\,{}_2F_1(1-4/\kappa, 2-8/\kappa, 2-4/\kappa, 1/s)\,s^{(4-\kappa)/\kappa}}{\Gamma(2-4/\kappa)\,\Gamma(4/\kappa-1/2)}\,. \tag{6.13}$$

The proof will be brief, since it is similar to the detailed proof of the more elaborate result in [LSW01a].

Proof. Let $s > 1$, and set

$$Y_t := \frac{g_t(1) - \xi(t)}{g_t(s) - \xi(t)}\,,$$

for $t < \tau(1)$. It is not hard to see that $\lim_{t \nearrow \tau(1)} Y_t = 1$ if $X > s$ and $\lim_{t \nearrow \tau(1)} Y_t = 0$ if $X < s$: Just notice that for $X > s$ and $t \nearrow \tau(1)$, the extremal distance within H_t from $[1,s]$ to $[\mathrm{Re}\,\gamma(t), \gamma(t)]$ tends to ∞, whereas for $X < s$ we have $\tau(1) < \tau(s)$. Direct inspection shows that the right-hand side of (6.13) is zero when $s = \infty$ and is 1 when $s = 1$. In particular, it is bounded for $s \in [1,\infty)$. Moreover, Itô's formula shows that when we plug $1/Y_t$ in place of s in the right-hand side of (6.13), we get a local martingale. Since $Y_0 = 1/s$ and $Y_t \in (0,1)$ for $t < \tau(1)$, it therefore suffices to prove that $\lim_{t \nearrow \tau(1)} Y_t \in \{0,1\}$ a.s. The only way this could fail is if $X = s$ occurs with positive probability. However, scale invariance shows that $\mathbf{P}[X = s'] \geq \mathbf{P}[X = s]$ when $s' \in (1,s)$. Hence $\mathbf{P}[X = s] = 0$ for all $s \neq 1$. □

Proof of Theorem 6.4. The case $\kappa \in [0,4]$ is covered by Theorem 6.1 and Theorem 5.1.

Suppose that $\kappa \in (4,8)$ and $z \in \overline{\mathbb{H}}$. Lemma 6.5 shows that $z \in \bigcup_{t>0} K_t$ a.s. If $z \in \mathbb{H}$, then Lemma 6.3 and (6.2) show a.s. $z \notin \gamma[0,\infty)$. Lemma 6.6 shows that $1 \notin \gamma[0,\infty)$ a.s., and the same follows for every $z \in \mathbb{R} \setminus \{0\}$ by scale and reflection invariance. This covers the range $\kappa \in (4,8)$.

Now suppose that z_0 is swallowed, and let t_0 be the swallowing time. If $\delta > 0$ is smaller than the distance from z_0 to $\gamma[0,t_0]$, then all the points within distance δ from z_0 are swallowed with z_0. Therefore, if the probability that some $z_0 \in \overline{\mathbb{H}} \setminus \{0\}$ is swallowed is positive, then by (6.2) there is some z_0 such that $\mathbf{P}\big[Z(z_0) < \infty\big] > 0$. By Lemma 6.3, this implies that for $\kappa \geq 8$ a.s. there is no $z \in \overline{\mathbb{H}} \setminus \{0\}$ which is swallowed. □

Exercise 6.7. Prove that $\mathbf{P}\big[\tau(z) < \infty\big] = 0$ when $\kappa \in (0, 4]$ and $z \in \mathbb{H}$.

7. Transience

THEOREM 7.1 (Transience). *For all $\kappa \neq 8$ the SLE_κ trace $\gamma(t)$ is transient a.s. That is,*

$$\mathbf{P}\Big[\liminf_{t\to\infty} |\gamma(t)| = \infty\Big] = 1\,.$$

LEMMA 7.2. *Suppose that $\kappa \leq 4$, and let $x \in \mathbb{R} \setminus \{0\}$. Then a.s. $x \notin \overline{\gamma[0,\infty)}$.*

Proof. By symmetry, we may assume that $x > 0$. Moreover, scaling shows that it is enough to prove the lemma in the case $x = 1$. So assume that $x = 1$. Let $s \in (0, 1/4)$ and suppose that there is some first time $t_0 > 0$ such that $\big|\gamma(t_0) - 1\big| = s$. This implies that $\big|g_{t_0}(1/2) - \xi(t_0)\big| = O(s)$. Indeed, $\big|g_{t_0}(1/2) - \xi(t_0)\big|$ is the limit as $y \to \infty$ of y times the harmonic measure in $\mathbb{H} \setminus \gamma[0, t_0]$ from $i\,y$ of the union of $[0, 1/2]$ and the "right-hand side" of $\gamma[0, t_0]$. By the maximum principle, this harmonic measure is bounded above by the harmonic measure in $\mathbb{H}$ of the interval $[1, \gamma(t_0)]$, which is $O(s/y)$.

In the proof of Lemma 6.2, we have determined the probability that $Y_x(t) = g_t(x) - \xi(t)$ hits b before a, where $x \in [a, b]$ and $0 < a < b$. For $\kappa < 4$, that probability goes to 1 as $a \searrow 0$, uniformly in b. Consequently, a.s. there is a random $\varepsilon > 0$ such that for all $b > x$, $Y_x(t)$ hits b before ε. Since $\sup_{t\in[0,T]} Y_x(t) < \infty$ a.s. for all finite T, it follows that $\inf_{t\geq 0} Y_x(t) > 0$ a.s. Using this with $x = 1/2$, we conclude from the previous paragraph that a.s. $1 \notin \overline{\gamma[0,\infty)}$, at least when $\kappa < 4$.

Assume now that $\kappa = 4$. Although the above proof does not apply (since $Y_x(t)$ is recurrent), there are various different ways to handle this case. Let $0 < y < x < \infty$. Let $\bar\gamma(t)$ denote the complex conjugate of $\gamma(t)$, and set $D_t := \mathbb{C} \setminus \big(\gamma[0,t] \cup \bar\gamma[0,t] \cup (-\infty, y]\big)$. Using Schwarz reflection, g_t may be conformally extended to D_t. Note that $g_t(D_t) = \mathbb{C} \setminus \big(-\infty, g_t(y)\big]$. By the Koebe 1/4 Theorem applied to g_t^{-1}, the distance from x to ∂D_t is at least $\frac{1}{4}\big(g_t(x) - g_t(y)\big)/g_t'(x)$. Consequently, it suffices to show that a.s.

$$\sup_{t\geq 0} g_t'(x)/\big(g_t(x) - g_t(y)\big) < \infty\,.$$

Set $Q(t) = Q_{x,y}(t) := \log g_t'(x) - \log\big(g_t(x) - g_t(y)\big)$. Then the above reduces to verifying that $\sup_t Q(t) < \infty$. The time derivative of Q is

$$\partial_t Q(t) = -2\,Y_x(t)^{-2} + 2\,Y_x(t)^{-1}\,Y_y(t)^{-1}\,.$$

In particular, $Q(t)$ is increasing in t. Define

$$G(s) := \log s\,\log(1+s) - \frac{1}{2}\log^2(1+s) + \mathrm{Li}_2(-s)\,,$$

where $\mathrm{Li}_2(z) = \int_z^0 s^{-1}\log(1-s)\,ds$. Note that G satisfies $s\,(1+s)^2\,G''(s) + s\,(1+s)\,G'(s) = 1$. A straightforward application of Itô's formula shows that $Q(t) - G_t$ is a local martingale, where

$$G_t := G\Big(\frac{g_t(x) - g_t(y)}{g_t(y) - \xi(t)}\Big)\,.$$

Consequently, there is an increasing sequence of stopping times t_n tending to ∞ a.s. such that $\mathbf{E}[Q(t_n)] = \mathbf{E}[G_{t_n}] + Q(0) - G_0$. It is immediate to verify that $\sup\{G(x) : x > 0\} < \infty$. Hence, $\limsup_{n\to\infty} \mathbf{E}[Q(t_n)] < \infty$. Because Q is increasing in t, it follows that $\mathbf{E}[\sup_t Q(t)] < \infty$ and hence $\sup_t Q(t) < \infty$ a.s., which completes the proof. □

LEMMA 7.3. *Suppose that $\kappa > 4$, $\kappa \neq 8$, and let $t > 0$. Then a.s. there is some $\varepsilon > 0$ such that $K_t \supset \{z \in \mathbb{H} : |z| < \varepsilon\}$.*

Proof. Consider first the case $\kappa \in (4, 8)$. We know that $\gamma(t)$ is a continuous path, by Theorem 5.1. Note that by scale and reflection invariance, Lemma 6.6 shows that for $a < b < 0$ there is a positive probability that the interval $[a, b]$ is swallowed all at once; i.e., $\mathbf{P}\big[[a, b] \cap \gamma[0, \infty) = \emptyset\big] > 0$, but a.s. there is some t such that $[a, b] \subset K_t$. When this happens, we have $[a, b] \cap \overline{H_t} = \emptyset$.

Let S be the set of limit points of $g_{\tau(1)}(z)$ as $z \to 0$ in $H_{\tau(1)}$. (If $0 \notin \overline{H_{\tau(1)}}$, set $S = \emptyset$.) Then $S \subset \big(-\infty, \xi(\tau(1))\big)$. Note that $\big(g_{\tau(1)}(\gamma(t + \tau(1))) - \xi(\tau(1)) : t \geq 0\big)$ has the same distribution as $\big(\gamma(t) : t \geq 0\big)$. Consequently, by the above paragraph, there is positive probability that $g_{\tau(1)}(\gamma[\tau(1), \infty))$ is disjoint from S, but there is some $t > \tau(1)$ such that $g_{\tau(1)}(K_t \setminus K_{\tau(1)}) \supset S$. This implies that with positive probability there is some $t > 0$ such that

$$\delta := \mathbf{P}\big[0 \notin \overline{H_t}\big] > 0\,.$$

By scale invariance, δ does not depend on t. Since these events are monotone in t, it follows that

$$\mathbf{P}\big[\forall t > 0,\ 0 \notin \overline{H_t}\big] = \delta\,.$$

By Blumenthal's 0-1 law it therefore follows that $\delta = 1$. This completes the proof in the case $\kappa \in (4, 8)$.

Now suppose that $\kappa > 8$. The above argument implies that it is sufficient to show that there is some $t > 0$ such that with positive probability $0 \notin \overline{H_t}$. Striving for a contradiction, we assume that for all $t > 0$ a.s. $0 \in \overline{H_t}$.

Let $z_0 \in \mathbb{H}$ be arbitrary. We know from Theorem 6.4 and Lemma 6.5 that a.s. $z_0 \in \gamma[0, \infty)$. Set $t_0 := \tau(z_0)$. Then $z_0 = \gamma(t_0)$. Let $t > 0$. Conditioned on $\mathcal{F}_{t_0}$, the set $g_{t_0}(K_{t+t_0} \setminus K_{t_0})$ has the same distribution as $K_t + \xi(t_0)$. It therefore follows by our above assumption that a.s. $\xi(t_0) \in \overline{g_{t_0}(H_{t+t_0})}$. Mapping by $g_{t_0}^{-1}$, this implies that a.s. $z_0 \in \overline{H_{t+t_0}}$; that is, $z_0 \in \partial K_{t+t_0}$. We conclude that for

every $z \in \mathbb{H}$ and every $t > 0$ a.s. $z \in \partial K_{t+\tau(z)}$. Therefore, a.s.

$$\text{area}(\partial K_t) = \int_{z\in\mathbb{H}} 1_{z\in\partial K_t}\, dx\, dy \geq \int_{z\in\mathbb{H}} 1_{t>\tau(z)}\, dx\, dy\,,$$

and Fubini implies that ∂K_t has positive measure with positive probability for large t. However, this contradicts Corollary 5.3, completing the proof of the lemma. □

Proof of Theorem 7.1. First suppose that $\kappa > 4$. Then we know from Lemma 7.3 that there is a random but positive $r > 0$ such that $K_1 \supset \{z \in \mathbb{H} : |z| < r\}$. Let $\varepsilon > 0$, and let $r_0 = r_0(\kappa, \varepsilon) > 0$ be a constant such that $\mathbf{P}[r < r_0] < \varepsilon$. Then the scaling property of SLE shows that for all $t > 0$

$$\mathbf{P}\Big[K_t \supset \{z \in \mathbb{H} : |z| < r_0\sqrt{t}\}\Big] \geq 1 - \varepsilon\,.$$

This implies that

$$\mathbf{P}\big[\exists t' > t \ : |\gamma(t')| < r_0\sqrt{t}\big] < \varepsilon\,.$$

Since $\varepsilon > 0$ was arbitrary, this clearly implies transience.

Consider now the case $\kappa \in [0, 4]$. In this range, γ is a simple path. Then a.s. there are two limit points x_0, x_1 for $g_1(z)$ as $z \to 0$ in H_1. Note that $g_1(\gamma[1,\infty))$ has the same distribution as $\gamma[0,\infty)$ translated by $\xi(1)$. Consequently, Lemma 7.2 shows that a.s. x_0 and x_1 are not in the closure of $g_1(\gamma[1,\infty))$. This means that $0 \notin \overline{\gamma[1,\infty)}$ a.s. The proof is then completed as in the case $\kappa > 4$ above. □

COROLLARY 7.4. *Suppose that* $\kappa > 8$, *then* $\gamma[0,\infty) = \overline{\mathbb{H}}$ *a.s.*

Proof. From Lemma 6.3 and (6.2) we know that $\gamma[0,\infty)$ is a.s. dense in $\overline{\mathbb{H}}$. Since γ is a.s. transient, $\gamma[0,\infty)$ is closed in $\mathbb{C}$, and hence is equal to $\overline{\mathbb{H}}$. □

Update. Corollary 7.4 and Theorem 7.1 are true also for $\kappa = 8$. The proofs are based on the extension [LSW] to $\kappa = 8$ of Theorem 5.1, and are otherwise the same.

8. Dimensions

In this section we study the size of the SLE trace and the hull boundary. When $\kappa \leq 4$, these two sets are the same, and it is interesting to compare the methods and results. For sets $A \subset \mathbb{C}$, denote by $N(\varepsilon) = N(\varepsilon, A)$ the minimal number of disks of radius ε needed to cover$A$.

8.1. *The size of the trace.*

THEOREM 8.1. *For every nonempty bounded open set A with $\overline{A} \subset \mathbb{H}$,*

$$\limsup_{\varepsilon \to 0} \frac{\log \mathbf{E}\Big[N\big(\varepsilon, \gamma[0,\infty) \cap A\big)\Big]}{|\log \varepsilon|} = \begin{cases} 1 + \kappa/8 & \kappa < 8, \\ 2 & \kappa \geq 8\,. \end{cases}$$

Proof. We may assume $\kappa < 8$ because $\gamma[0,\infty) = \overline{\overline{\mathbb{H}}}$ if $\kappa > 8$, by Corollary 7.4, and when $\kappa = 8$ we know that for every $z \in \mathbb{H}$ we have $z \in \gamma[0,\infty)$ a.s., by Theorem 6.4 and Lemma 6.5. Fix $\varepsilon > 0$ smaller than the distance from A to $\mathbb{R}$, and denote those points $z \in \varepsilon\mathbb{Z}^2$ with $\operatorname{dist}(z, A) < \varepsilon$ by $z_j, 1 \leq j \leq n$. Then n is comparable to $1/\varepsilon^2$. Set $N := N\big(\varepsilon, \gamma[0,\infty) \cap A\big)$. Obviously $N \leq \#\{z_j : \operatorname{dist}(z_j, \gamma[0,\infty)) < \varepsilon\} \leq 100N$. From (6.2) and (6.5) we get that $\mathbf{E}[N]$ is comparable to

$$\sum_{j=1}^{n} \mathbf{P}\big[Z(z_j) > \varepsilon^{-1}\big].$$

Write $F(t) = \mathbf{P}[Z(i) > t]$. Using (6.5), the proof of the theorem reduces to showing that

$$\limsup_{t \to \infty} \frac{\log F(t)}{\log t} = \kappa/8 - 1.$$

By (6.8),

$$\lim_{a \nearrow 1-\kappa/8} \frac{\hat{G}_{a,\kappa}(1)}{1 - \kappa/8 - a} = \frac{\sqrt{\pi}\,\Gamma(4/\kappa - 1/2)}{2\,\Gamma(4/\kappa)}$$

and since $\hat{G}_{a,\kappa}(i) = 1$ from Lemma 6.3 we obtain

$$C^{-1} \frac{1}{1 - \kappa/8 - a} \leq \mathbf{E}\big[Z(i)^a\big] \leq C \frac{1}{1 - \kappa/8 - a}$$

for all $a < 1 - \kappa/8$ sufficiently close to $1 - \kappa/8$. Hence, for small $\Delta = \Delta(\kappa) > 0$,

$$C\Delta \geq \int_0^{\Delta} \mathbf{E}\big[Z(i)^{1-\kappa/8-\delta}\big]\,\delta\,d\delta = \int_0^{\Delta} \int_0^{\infty} (1 - \kappa/8 - \delta)\, t^{-\kappa/8-\delta}\, F(t)\, dt\, \delta\, d\delta\,.$$

From $\int_0^{\Delta} t^{-\kappa/8-\delta}\,\delta\,d\delta \geq C\,t^{-\kappa/8}/(\log t)^2$ for $t \geq e$ (where Δ is treated as a constant), we obtain

$$\int_e^{\infty} \frac{t^{-\kappa/8}\,F(t)}{(\log t)^2}\,dt = O(1).$$

Using $F(t)/F(2t) = O(1)$ from (6.5) this easily implies for large t

$$F(t) \leq C\,t^{\kappa/8-1}\,(\log t)^2.$$

There is a sequence $t_n \to \infty$ with

$$F(t_n) \geq \frac{t_n^{\kappa/8-1}}{(\log t_n)^2},$$

for otherwise we would have $\mathbf{E}\big[Z(i)^{1-\kappa/8}\big] = O(1)\int_e^\infty \frac{dt}{t(\log t)^2} < \infty$, contradicting Lemma 6.3. □

An immediate consequence is

COROLLARY 8.2. *For $\kappa < 8$, the Hausdorff dimension of $\gamma[0,\infty)$ is a.s. bounded above by $1 + \kappa/8$.*

8.2. *The size of the hull boundary.* When studying the size of the hull boundary ∂K_t, it suffices to restrict to $t = 1$, by scaling. We will first estimate the size of $\partial H_1 \cap \{y > h\}$ for small $h > 0$ in terms of the convergence exponent of the Whitney decomposition of H_1. A Whitney decomposition of H_1 is a covering by essentially disjoint closed squares $Q \subset H_1$ with sides parallel to the coordinate axes such that the side length $d(Q)$ is comparable to the distance of Q from the boundary of H_1. One can always arrange the side lengths to be integer powers of 2 and that

$$d(Q) \le \operatorname{dist}(Q, \partial H_1) \le 8\, d(Q). \tag{8.1}$$

For example, one can take all squares of the form $Q = 2^n\big([k,k+1]\times[j,j+1]\big)$ such that $Q \subset H_1$, $d(Q) \le \operatorname{dist}(Q,\partial H_1)$, and Q is not strictly contained in any other Q' satisfying these conditions. One important feature of the Whitney decomposition is that the hyperbolic diameters of the tiles in the decomposition are bounded away from 0 and ∞. (The hyperbolic metric on H_1 is the metric for which the conformal map $g_1 : H_1 \to \mathbb{H}$ is an isometry, where $\mathbb{H}$ is taken with the hyperbolic metric with length element $y^{-1}\,|dz|$.) This follows from the well-known consequence of the Koebe 1/4 theorem saying that the hyperbolic length element in H_1 is comparable to $\operatorname{dist}(z,\partial H_1)^{-1}\,|dz|$. This description of the hyperbolic metric will be useful below.

Fix a Whitney decomposition of H_1, and denote by W the collection of those squares Q for which $\operatorname{dist}(Q, K_1) \le 1$ and $\sup\{\operatorname{Im} z : z \in Q\} \ge h$. For $a > 0$ let

$$S(a) = S_h(a) := \sum_{Q\in W} d(Q)^a.$$

THEOREM 8.3. *Set*

$$\delta(\kappa) := \begin{cases} 1 + \kappa/8 & \kappa \le 4, \\ 1 + 2/\kappa & \kappa > 4. \end{cases}$$

For $\kappa > 0$, $\kappa \ne 4$, and all sufficiently small $h \in (0,1)$ we have

$$\mathbf{E}[S(a)] \begin{cases} < \infty & a > \delta(\kappa), \\ = \infty & a \le \delta(\kappa). \end{cases}$$

For $j, n \in \mathbb{Z}$, let $x_{j,n} := j\,2^{-n}$, $y_n := 2^{-n}$ and $z_{j,n} = x_{j,n} + i\,y_n$. Let I be the interval $g_1(\partial K_1) - \xi(1)$; that is, $\hat{f}^{-1}(\partial K_1)$.

LEMMA 8.4. *There are constants $C = C(h), c = c(h) > 0$ such that the following holds. Every $Q \in W$ is within distance C in the hyperbolic metric of H_1 from some $\hat{f}_1(z_{j,n})$ with $x_{j,n} \in I$ and $n \geq 0$, and every $\hat{f}_1(z_{j,n})$ such that $x_{j,n} \in I$, $n \geq 0$ and $\operatorname{Im} \hat{f}_1(z_{j,n}) \geq h$ is within hyperbolic distance C from some $Q \in W$. Moreover, in each of these cases,*

$$c^{-1}\, d(Q) \leq y_n \left|\hat{f}_1'(z_{j,n})\right| \leq c\, d(Q)\,.$$

Proof. Observe that the diameter of K_1 is bounded from below, because the uniformizing map $g : \mathbb{H} \setminus K \to \mathbb{H}$ with the hydrodynamic normalization (2.2) tends to the identity as K tends to a point, and the uniformizing map g_1 for K_1 satisfies $g_1(z) = z + 2/z + \cdots$ near ∞, by (2.3).

Let $\hat{z} \in Q \in W$. It follows from the definition of W and the fact that $\operatorname{diam}(K_1)$ is bounded from zero that the harmonic measure of K_1 from $\hat{z}$ in H_1 is bounded from zero. Therefore, the harmonic measure of I from $\hat{f}_1^{-1}(\hat{z})$ in $\mathbb{H}$ is bounded from zero. This implies that there is some $z_{j,n}$ within bounded distance from $\hat{f}_1^{-1}(\hat{z})$ in the hyperbolic metric of $\mathbb{H}$ such that $x_{j,n} \in I$. Then $\hat{f}_1(z_{j,n})$ is within bounded hyperbolic distance from Q. Because $\sup\{\operatorname{Im} z : z \in K_1\} = O(1)$, we also have $\operatorname{Im} \hat{z} = O(1)$, and therefore $\operatorname{Im} \hat{f}_1(z_{j,n}) = O(1)$. Since $\operatorname{Im} \hat{f}_1(z_{j,n}) \geq \operatorname{Im} z_{j,n}$, it follows that $\max\{0, -n\} = O(1)$. Then $z' := x_{j,n} + i\, 2^{n \wedge 0}$ is equal to $z_{j',n'}$ for some $j' \in \mathbb{Z}$, $n' \in \mathbb{N}$, and satisfies the requirements in the first statement of the lemma.

From the fact that $\operatorname{diam}(K_1)$ is bounded from zero it also follows that $|I|$, the length of I, is bounded from zero. Suppose that we have $x_{j,n} \in I$, $j \in \mathbb{Z}$, $n \in \mathbb{N}$. Then the harmonic measure of I from $z_{j,n}$ is bounded from zero. Consequently, the harmonic measure of K_1 from $\hat{f}_1(z_{j,n})$ in H_1 is bounded from zero. Now suppose also that $\operatorname{Im} \hat{f}_1(z_{j,n}) \geq h$. Since $\sup\{\operatorname{Im} z : z \in K_1\} = O(1)$ and $\operatorname{Im} \hat{f}_1(z_{j,n}) = O(1)$, it follows that the Euclidean distance from $\hat{f}_1(z_{j,n})$ to K_1 is bounded from above. If $\operatorname{dist}\bigl(\hat{f}_1(z_{j,n}), K_1\bigr) \leq 1$, then the Whitney tile Q containing $\hat{f}_1(z_{j,n})$ is in W. Otherwise, let p be a point closest to $\hat{f}_1(z_{j,n})$ among the points z satisfying $\operatorname{Im} z \geq h$ and $\operatorname{dist}(z, K_1) \leq 1$. Since $h < 1$ and $\operatorname{dist}\bigl(\hat{f}_1(z_{j,n}), K_1\bigr) = O(1)$, it follows that $\bigl|p - \hat{f}_1(z_{j,n})\bigr| = O(1)$. Since the points in $\bigl[p, \hat{f}_1(z_{j,n})\bigr]$ have distance at least h from ∂H_1, it follows that the hyperbolic distance from p to $\hat{f}_1(z_{j,n})$ is bounded above. For the second statement of the lemma we may then take Q to be the Whitney tile containing p.

The last statement holds because $\hat{f}_1$ is an isometry from the hyperbolic metric of $\mathbb{H}$ to the hyperbolic metric of H_1. □

Proof of Theorem 8.3. Set

$$\tilde{S}_h(a) := \sum_{j=-\infty}^{\infty} \sum_{n=0}^{\infty} 1_{x_{j,n} \in I}\, 1_{\operatorname{Im} \hat{f}_1(z_{j,n}) > h}\, \bigl|\hat{f}_1'(z_{j,n})\, y_n\bigr|^a\,.$$

We now show that this quantity is comparable to $S(a)$. Let C be the constant in the lemma. Since each square $Q \in W$ has a bounded diameter in the hyperbolic metric of H_1, and the hyperbolic distance in $\mathbb{H}$ between any two distinct $z_{j,n}$ is bounded from below, it follows that there are at most a bounded number of $\hat{f}_1(z_{j,n})$ within distance C, in the hyperbolic metric of H_1, from any single $Q \in W$. Similarly, there are at most a bounded number of $Q \in W$ within hyperbolic distance C from any point in H_1. Moreover, if z is within hyperbolic distance C from some $Q \in W$, then $\operatorname{Im} z$ is bounded away from zero, because the hyperbolic distance in H_1 between any two points is an upper bound on the distance between them in the hyperbolic metric of $\mathbb{H}$ (by the Schwarz lemma, for instance). Consequently, Lemma 8.4 implies

$$c_1^{-1}\, \tilde{S}_h(a) \le S_h(a) \le c_1\, \tilde{S}_{h'}(a)\,, \tag{8.2}$$

for some constant c_1, h', which depends only on h and a. Now,

$$\mathbf{E}\big[\tilde{S}_h(a)\big] = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \sum_{j=-\infty}^{\infty} \mathbf{E}\Big[\big|\hat{f}_1'(z_{j,n})\; y_n\big|^a\, 1_{\mathcal{A}(j,n,m)} \Big] \tag{8.3}$$

where $\mathcal{A}(j,n,m)$ denotes the event $\{x_{j,n} \in I, \operatorname{Im}\hat{f}_1(z_{j,n}) \ge h, m \le |I| < m+1\}$. We split this sum into two pieces, depending on whether $m \le n$ or $m > n$. Clearly, $0 \in I$. It is easy to see that

$$|I| \le O\Big(1 + \max_{0\le t\le 1} \big|\xi(t)\big|\Big)\,,$$

for example, by considering the differential equation (2.1) directly. This implies $\mathbf{P}\big[|I| \ge m\big] \le O(1)\, e^{-C(\kappa)\, m^2}$, for some $C(\kappa) > 0$. Since

$$\big|y\, \hat{f}_1'(x + i\, y)\big| \le O(1) \operatorname{Im} \hat{f}_1(x + i\, y) \le O(1)$$

for $0 < y \le 1$, we have the trivial estimate

$$\begin{aligned} \mathbf{E}\Big[\sum_{j=-\infty}^{\infty} |\hat{f}_1'(z_{j,n})\; y_n|^a 1_{\mathcal{A}(j,n,m)} \Big] &\le O(1)\, m\, 2^n\, \mathbf{P}\big[|I| \ge m\big] \\ &\le O(1)\, m\, 2^n\, e^{-C(\kappa) m^2} \end{aligned}$$

for all $m \ge 0$. This shows that in (8.3) the terms with $m \ge n$ have finite sum. For $m \le n$, we use Theorem 3.2 together with (3.17) and obtain

$$\mathbf{E}\Big[\sum_{j=-\infty}^{\infty} \big|\hat{f}_1'(z_{j,n})\; y_n\big|^a\, 1_{\mathcal{A}(j,n,m)} \Big] \le C(h,\kappa,b)\, m^{2b+1}\, 2^{-n\,(\lambda - 2b - 1)},$$

where a, λ, b and κ are related by (3.6) with $\nu = 1$ and $b > 0$. For b between $1/2$ and $2/\kappa$ we have $\lambda - 2b - 1 > 0$ and conclude that $\mathbf{E}\big[\tilde{S}(a)\big] < \infty$ in this range. By (8.2), $\mathbf{E}[S(a)] < \infty$ for $\kappa < 4, a > 1 + \kappa/8$ and for $\kappa > 4, a > 1 + 2/\kappa$.

Now consider $a = \delta(\kappa)$. Suppose that $2^{-n} < h/4$ and $\hat{f}_1(z_{j,n}) \ge h$. We claim that the harmonic measure in H_1 of K_1 from $\hat{f}_1(z_{j,n})$ is bounded from

zero. Indeed, fix a constant $h_1 > 1$ such that $h_1/2 \geq \sup\{\operatorname{Im} z : z \in K_1\}$. (It is easy to see that $h_1 = 4$ works.) Consider the harmonic function $Y(z) := \operatorname{Im} g_1(z)$ on H_1. Then $Y(x + i\,y) - y \to 0$ uniformly in x, as $y \to \infty$, by the hydrodynamic normalization (2.2) of g_1. Since $\mathbb{H} + \frac{i}{2}\,h_1 \subset H_1$, the maximum principle applied to $Y(z) - \operatorname{Im} z$ implies that $Y(x + i\,h_1) \geq h_1/2$ for all $x \in \mathbb{R}$. Consequently, for every $z \in H_1$ with $\operatorname{Im} z < h_1$, the harmonic measure in $H_1 \cap \{y < h_1\}$ from z of the line $y = h_1$ is at most $2\,Y(z)/h_1$. However, the harmonic measure from z of $y = h_1$ in the domain $\mathbb{H} \cap \{y < h_1\}$ is $\operatorname{Im} z/h_1$. Therefore, the harmonic measure from z of K_1 is at least the difference between these two quantities, which is at least $\operatorname{Im} z/h_1 - 2\,Y(z)/h_1$. Applying this to $z := \hat{f}_1(z_{j,n})$ proves our claim that the harmonic measure of K_1 from $\hat{f}_1(z_{j,n})$ in H_1 is bounded from zero. This implies that the harmonic measure of I from $z_{j,n}$ in $\mathbb{H}$ is bounded from zero, which implies that at bounded hyperbolic distance from $z_{j,n}$ we can find a point $z_{j',n'}$ with $x_{j',n'} \in I$. Consequently, to prove that $\mathbf{E}\big[\tilde{S}_h\big(\delta(\kappa)\big)\big] = \infty$ for all sufficiently small $h > 0$, it suffices to prove that for all $n_0 \in \mathbb{N}$ and all sufficiently small $h > 0$

$$\mathbf{E}\Big[\sum_{j=-\infty}^{\infty}\sum_{n=n_0}^{\infty} 1_{\operatorname{Im}\hat{f}_1(z_{j,n})>h}\,|\hat{f}_1'(z_{j,n})\,y_n|^{\delta(\kappa)}\Big] = \infty\,.$$

This follows from Lemma 3.3 with $b = 1/2$ if $\kappa < 4$ or $b = 2/\kappa$ if $\kappa > 4$. An appeal to (8.2) gives $\mathbf{E}\big[S\big(\delta(\kappa)\big)\big] = \infty$. Since $\mathbf{E}\big[S(a)\big] < \infty$ for $a > \delta(\kappa)$, it also follows that $\mathbf{E}\big[S(a)\big] = \infty$ for $a < \delta(\kappa)$, and the proof is complete. □

Corollary 8.5. *For all $\kappa > 0$, $\kappa \neq 4$, and all sufficiently small $h > 0$*

$$\limsup_{\varepsilon\to 0} \big|\log\varepsilon\big|^{1+h}\,\varepsilon^{\delta(\kappa)}\,\mathbf{E}\Big[N\big(\varepsilon, \partial K_1 \cap \{y > h\}\big)\Big] = \infty.$$

Proof. Denote by $W(n)$ the number of $Q \in W$ with $2^{-n-1} < d(Q) \leq 2^{-n}$. Then (8.1) immediately gives

$$W(n) \leq C\,N(2^{-n}, \partial K_1 \cap \{y > h\}). \tag{8.4}$$

The corollary now follows from $S(a) = \sum_{n=0}^{\infty} W(n) 2^{-an}$ and Theorem 8.3. □

Finally we state an upper bound for the covering dimension.

Theorem 8.6. *For all $\kappa, h > 0$ and $a > \delta(\kappa)$, a.s.*

$$\lim_{\varepsilon\to 0} \varepsilon^a\,N\big(\varepsilon, \partial K_1 \cap \{y > h\}\big) = 0.$$

Proof. Suppose first that $\kappa \neq 4$. By Theorem 5.2 we know that $\hat{f}_1$ is a.s. Hölder continuous. There is no lower bound version of (8.4) for general domains. However, by results of Pommerenke [Pom92] and Makarov [Mak98] it is known that for Hölder domains (i.e., the conformal map from $\mathbb{U}$ into the

domain is Hölder) the asymptotics of N can be recovered from the convergence exponent of the Whitney decomposition. Our proof follows this philosophy. We claim that there is a (random) constant $C = C(\omega) > 0$, depending on h and the Hölder norm of $\hat{f}_1$, such that for every small $\varepsilon > 0$ and every point $z_0 \in \partial K_1 \cap \{y \geq 2h\}$

$$\exists Q \in W, \qquad Q \subset D(z_0, \varepsilon), \;\; \varepsilon \geq d(Q) \geq C\varepsilon/|\log \varepsilon| \,. \tag{8.5}$$

In fact, this follows from [KR97], but we include a short proof here, for the convenience of the reader.

Let $x_0 \in \mathbb{R}$ be a point such that $\hat{f}_1(x_0) = z_0$. Clearly, there is such an x_0, because $\hat{f}_1$ extends continuously to $\overline{\mathbb{H}}$. Let $\beta(s) := \hat{f}_1(x_0 + i\, s)$, $s > 0$, let $\hat{Q}(s)$ be the set of Whitney cells $Q \in W$ which intersect $\beta\big((0,s)\big)$, and set

$$q(s) := \sum_{Q \in \hat{Q}(s)} d(Q) \,.$$

Since $\hat{f}_1$ is an isometry between the hyperbolic metric of $\mathbb{H}$ and that of H_1, it follows that for every $s > 0$ the arc $\beta[s, 2s]$ meets a bounded number of Whitney cells, and each of these has Euclidean diameter comparable to the Euclidean diameter of $\beta[s, 2s]$. Suppose that $\hat{f}_1$ is Hölder with exponent $\alpha > 0$. Then, $d(Q) \leq O(1)\, s^\alpha$ for $Q \in \hat{Q}(s)$, and consequently also $q(s) \leq O(1)\, s^\alpha$. Observe that there is a constant c_0 so that $q(s)/\sup\big\{q(s') : s' < s\big\} < c_0$, because there are constant upper bounds on the number of Whitney cells containing a point, and on the ratio $d(Q)/d(Q')$ for neighboring Whitney cells Q and Q'. Let $s_0 := \sup\{s : q(s) < \varepsilon\}$ and $s_1 := \sup\{s : q(s) < c_0^{-1}\, \varepsilon/2\}$. The choice of c_0 ensures that $q(s_0) - q(s_1) \geq \varepsilon/2$. Clearly $\hat{Q}(s_0) \setminus \hat{Q}(s_1)$ has at most $O\big(1 + \log(s_0/s_1)\big)$ cells. Therefore, there is some $Q \in \hat{Q}(s_0) \setminus \hat{Q}(s_1)$ with $d(Q) \geq \varepsilon / O\big(1 + \log(s_0/s_1)\big)$. Since $q(s) \leq O(1)\, s^\alpha$, it follows that $O(1)\, s_1 \geq \varepsilon^{1/\alpha}$, and we get $O(1)\, d(Q) \geq \varepsilon/|\log \varepsilon|$, proving (8.5).

We know that for every set $F \subset \mathbb{C}$ the covering number $N(\varepsilon, F)$ is bounded by a constant times the maximum number of disjoint disks of radius ε centered in F. Therefore, by (8.5), $N\big(\varepsilon, \partial K_1 \cap \{y > 2\, h\}\big)$ is bounded by a constant times the number of Whitney cells in W with size between $O(1)\, \varepsilon$ and $\varepsilon / O\big(|\log \varepsilon|\big)$; that is,

$$N\big(2^{-n}, \partial K_1 \cap \{y > 2\, h\}\big) \leq C(\omega) \sum_{j=n}^{n + O(\log n)} W(j) \,,$$

which implies for all $a' > 0$,

$$N\big(2^{-n}, \partial K_1 \cap \{y > 2\, h\}\big) \leq C(\omega)\, 2^{(n + O(\log n))\, a'}\, S(a') \,.$$

By Theorem 8.3, $S(a') < \infty$ a.s. for $a' > \delta(\kappa)$. Therefore, the case $\kappa \neq 4$ in Theorem 8.6 now follows by taking $a' \in \big(\delta(\kappa), a\big)$.

The following argument covers not only the case $\kappa = 4$, but the whole range $\kappa \in (0, 4]$. For such κ, Theorem 8.1 can be used, since $\partial K_1 = \gamma[0, 1]$ for $\kappa \in (0, 4]$. Although Theorem 8.1 discusses the expected number of balls needed to cover $\gamma[0, \infty) \cap A$, where A is an arbitrary bounded set satisfying $\overline{A} \subset \mathbb{H}$, one can refine the result to bound $N\big(\varepsilon, \partial K_1 \cap \{y > H\}\big)$, as follows. Given a point $z = x + i\, y$, $y > h$, we need to estimate the probability that z is close to K_1. We may certainly restrict attention to the case where y is bounded by some constant. There is a constant $C \geq 1$ such that the probability that z is close to K_1 is bounded by the probability that there is some time $t_1 < 1$ such that $\xi(t_1)$ is within distance C from x times the probability that $Z(z)$ is large, conditioned on the existence of such a t_1. The proof is then completed by applying the estimates in the proof of Theorem 8.1 and the Markov property. The details are left to the reader. □

The upper bound for the Hausdorff dimension follows immediately:

COROLLARY 8.7. *For $\kappa > 4$, the Hausdorff dimension of ∂K_1 is a.s. at most $1 + 2/\kappa$.*

9. Further discussion and some open problems

We believe Theorem 5.1 to hold also in the case $\kappa = 8$:

CONJECTURE 9.1. *The* SLE_8 *trace is a continuous path a.s.*

Update. This is proved in [LSW].

CONJECTURE 9.2. *The Hausdorff dimension of the* SLE_κ *trace is a.s.* $1 + \kappa/8$ *when* $\kappa < 8$.

Update. A proof of this result is announced in [Bef] for $\kappa \neq 4$.

Corollary 8.2 shows that the Hausdorff dimension cannot be larger than $1 + \kappa/8$.

CONJECTURE 9.3 (Richard Kenyon). *The Hausdorff dimension of ∂K_1 is a.s. $1 + 2/\kappa$ when $\kappa \geq 4$.*

Corollary 8.7 shows that the Hausdorff dimension cannot be larger than $1 + 2/\kappa$.

Problem 9.4. Find an estimate for the modulus of continuity for the trace of SLE_4. Improve the estimate for the modulus of continuity for SLE_κ for other κ. What better modulus of continuity can be obtained with a different parametrization of the trace?

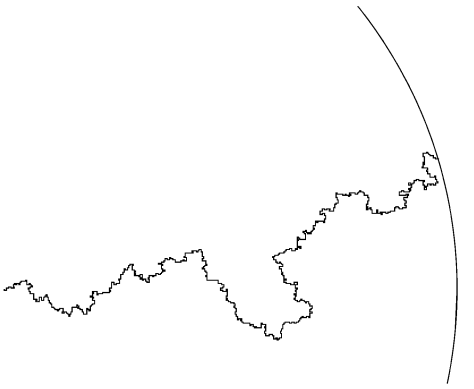


Figure 9.1: A sample of the loop-erased random walk; believed to converge to radial SLE_2.

Based on the conjectures for the dimensions [Dup00], Duplantier suggested that $SLE_{16/\kappa}$ describes the boundary of the hull of SLE_κ when $\kappa > 4$. This has further support, some of which is discussed below. However, at this point there isn't even a precise version for this duality statement.

Problem 9.5. Understand the relation between the boundary of the SLE_κ hull and $SLE_{16/\kappa}$ when $\kappa > 4$.

Update. See [Dub] for recent progress on this problem.

We now list some conjectures about discrete processes tending to SLE_κ for various κ. Let D be a bounded simply connected domain in $\mathbb{C}$ with smooth boundary, and let a and b be two distinct points in ∂D. Let ϕ be a conformal map from $\mathbb{H}$ to D such that $\phi(0) = a$ and $\phi(\infty) = b$. The SLE *trace in* D *from* a *to* b is defined as the image under ϕ of the trace of chordal SLE in $\mathbb{H}$. Similarly, one defines radial SLE in arbitrary simply connected domains strictly contained in $\mathbb{C}$.

In [Sch00] it was shown that if the loop-erased random walk in simply connected domains in $\mathbb{C}$ has a conformally invariant scaling limit, then this must be radial SLE_2. See Figure 9.1 for a sample of the loop-erased random walk in $\mathbb{U}$.

In [Sch00] it was also conjectured that the UST Peano curve, with appropriate boundary conditions (see Figure 9.2), tends to SLE_8 as the mesh goes to zero. This offers additional support to Duplantier's conjectured duality 9.5, Since the outer boundary of the hull of the UST Peano curve consists of two loop-erased random walks, one in the tree and one in the dual tree. See [Sch00] for additional details.

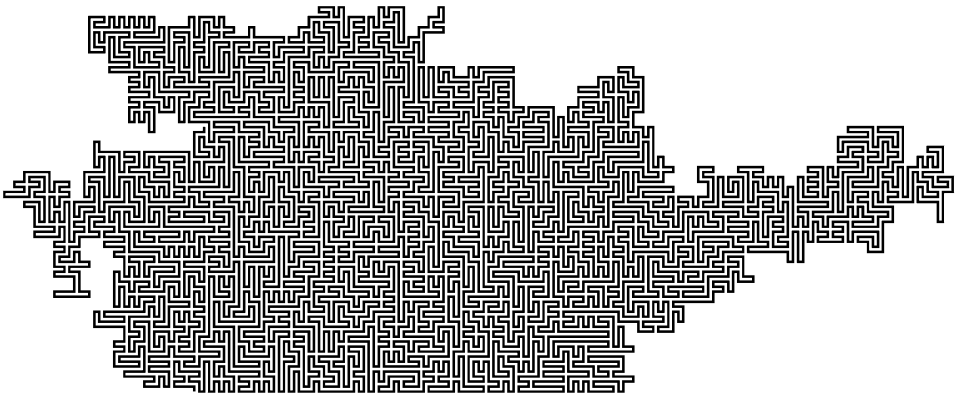


Figure 9.2: A piece of the UST Peano curve; believed to converge to SLE_8.

Update. The convergence of LERW to SLE_2 and of the UST Peano path to SLE_8 is proved in [LSW].

CONJECTURE 9.6. *Consider a square grid of mesh ε in the unit disk $\mathbb{U}$. Then the uniform measure on simple grid paths from 0 to $\partial\mathbb{U}$ on this grid converges weakly to the trace of radial* SLE_8 *as $\varepsilon \searrow 0$, up to reparametrization. The uniform measure on simple paths joining two distinct fixed boundary points converges weakly to the trace of chordal* SLE_8*, up to reparametrization.*

We have not bothered to make this conjecture entirely precise, since there is more than one natural choice for the topology on path space. However, this does not seem to be very important at this point.

Note that we do not require that the grid path be space filling (i.e., visit every vertex). Indeed, it is not hard to see that a uniformly selected path will be essentially filling, in the sense that the probability that a nonempty open set inside $\mathbb{U}$ is unvisited goes to zero as $\varepsilon \searrow 0$.

Consider an $n \times n$ grid in the unit square, and identify the vertices of the right and bottom boundary arcs to a single vertex w_0. Let G_n denote this graph, and note that the dual graph $G_n^\dagger$ is obtained by rotating G_n 180° and shifting appropriately. See Figure 9.3. Let ω be a sample from the critical random-cluster measure on G_n with parameters q and $p = p(q) = \sqrt{q}/(1+\sqrt{q})$ (which is the conjectured value of $p_c = p_c(q)$) and let $\omega^\dagger$ be the set of edges in $G_n^\dagger$ that do not cross edges of ω. Then the law of $\omega^\dagger$ is the random-cluster measure on $G_n^\dagger$ with the same parameters as for ω. Let β be the set of points in the unit square that are at equal distance from the component of ω containing w_0 and the component of $\omega^\dagger$ containing the wired vertex of $G_n^\dagger$.

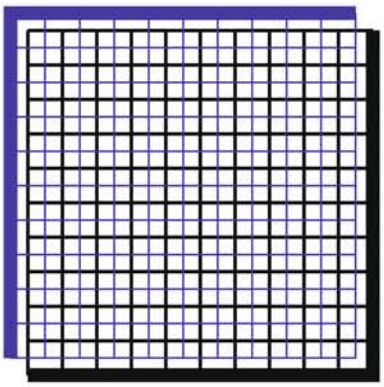


Figure 9.3: A partially wired grid and its dual.

CONJECTURE 9.7. *If $q \in (0,4)$, then as $n \to \infty$, the law of β converges weakly to the trace of* SLE_κ, *up to reparametrization, where* $\kappa = \frac{4\pi}{\cos^{-1}\left(-\sqrt{q}/2\right)}$.

The above value of κ is derived by applying the predictions [SD87] of the dimensions of the perimeters of clusters in the random cluster model.

Consider a random-uniform domino tiling of the unit square, such as in the top right of Figure 9.4. Taking the union with another independent domino tiling with the two lower corner squares removed, one gets a path joining the two removed squares; see Figure 9.4. This model was considered by [RHA97]. Richard Kenyon proposed the following

Problem 9.8. Does this double-domino model converge weakly (up to reparametrization) to SLE_4 as the mesh of the grid tends to zero?

Kenyon also produced some calculations for this model that agree with the corresponding calculations for SLE_4.

Let $a \in \mathbb{R}$, $\varepsilon > 0$, and consider the following random walk S on the grid $\varepsilon\,\mathbb{Z}^2$. Set $S_0 = 0$, and inductively, we construct S_{n+1} given $\{S_0, \ldots, S_n\}$. If $S_n \notin \mathbb{U}$, then stop. Otherwise, let $h_n : \varepsilon\,\mathbb{Z}^2 \to [0,1]$ be the function which is 1 outside of $\mathbb{U}$, 0 on $\{S_0, \ldots, S_n\}$, and discrete-harmonic elsewhere. Now choose S_{n+1} among the neighbors v of S_n such that $h(v) > 0$, with probability proportional to $h(v)^a$. When $a = 1$, this is called *Laplacian random walk*, and is known to be equivalent to loop-erased random walk. When $a = 0$ this walk is believed to be closely related to the boundary of percolation clusters.

The following problem came up in discussions with Richard Kenyon.

Problem 9.9. For what values of a does this process converge weakly to radial SLE? What is the correspondence $\kappa = \kappa(a)$?

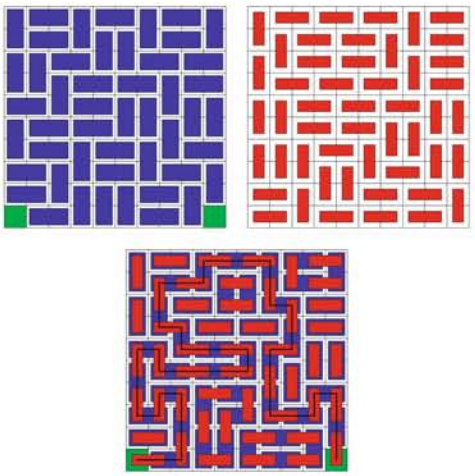


Figure 9.4: The double domino path.

Motivated in part by Duplantier's duality conjecture 9.5, in [LSW03] it is conjectured that chordal $\mathrm{SLE}_{8/3}$ is the scaling limit of the self-avoiding walk in a half-plane. (That is, the limit as $\delta \searrow 0$ of the limit as $n \nearrow \infty$ of the law of a random-uniform n-step simple path from 0 in $\delta\mathbb{Z}^2 \cap \overline{\mathbb{H}}$.) Some strong support to this conjecture is provided there, which in turn can be viewed as further support for the duality in 9.5.

Many of the processes believed to converge to chordal SLE have the following reversibility property. Up to reparametrization, the path from a to b in D $(a, b \in \partial D)$ has the same distribution as the path from b to a. It is then reasonable to conjecture:

CONJECTURE 9.10. *The chordal* SLE *trace is reversible for* $\kappa \in [0, 8]$.

It is not too hard to see that this fails for $\kappa > 8$. In particular, when $\kappa > 8$ the chordal SLE path is more likely to visit 1 before visiting i. Since inversion in the unit circle preserves 1 and i, the probability for visiting i before 1 would have to be $1/2$ if one assumes reversibility. To apply this argument one has to note that a.s. the path visits 1 and i exactly once.

Update. The above conjecture is known to hold when $\kappa = 0, 2, 8/3, 6, 8$. The cases $\kappa = 2, 8$ follow from [LSW] (though some additional work is needed in the case where $\kappa = 2$, since the convergence of the LERW is to radial SLE), the case $\kappa = 6$ follows from Smirnov's Theorem [Smi01], and the case $\kappa = 8/3$ follows from the characterization in [LSW03] of $\mathrm{SLE}_{8/3}$ as the restriction measure with exponent $5/8$.

Some conjectures and simulations concerning the winding numbers of SLE's are described in [WW].

Acknowledgments. The authors thank Richard Kenyon for useful discussions and Jeff Steiff for numerous comments on a previous version of this paper.

University of Washington, Seattle, WA
E-mail address: rohde@math.washington.edu
Web page: http://www.math.washington.edu/∼rohde/

Microsoft Research Corporation, Redmond, WA
E-mail address: schramm@microsoft.com
Web page: http:research.microsoft.com/∼schramm/

References

[Ahl73] L. V. Ahlfors, *Conformal Invariants: Topics in Geometric Function Theory, McGraw-Hill Series in Higher Mathematics*, McGraw-Hill Book Co., New York, 1973.

[Bef] V. Beffara, The dimension of the SLE curves; arXiv:math.PR/0211322.

[Car] J. Cardy, Conformal invariance and percolation; arXiv:math-ph/0103018.

[CM01] L. Carleson and N. Makarov, Aggregation in the plane and Loewner's equation, *Comm. Math. Phys.* **216** (2001), 583–607.

[Dub] J. Dubedat, SLE(κ, ρ) martingales and duality; arXiv:math.PR/0303128.

[Dup00] B. Duplantier, Conformally invariant fractals and potential theory, *Phys. Rev. Lett.*, **84** (2000), 1363–1367.

[Dur96] R. Durrett, *Stochastic Calculus: A Practical Introduction*, CRC Press, Boca Raton, FL, 1996.

[EMOT53] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Higher Transcendental Functions Vol.* I. Based, in part, on notes left by Harry Bateman. McGraw-Hill Book Company, Inc., New York, 1953.

[JM95] P. W. Jones and N. G. Makarov, Density properties of harmonic measure, *Ann. of Math.* **142**(1995), 427–455.

[KR97] P. Koskela and S. Rohde, Hausdorff dimension and mean porosity, *Math. Ann.* **309** (1997), 593–609.

[LSW01a] G. F. Lawler, O. Schramm, and W. Werner, Values of Brownian intersection exponents. I. Half-plane exponents, *Acta Math.* **187** (2001), 237–273.

[LSW01b] ———, Values of Brownian intersection exponents. II. Plane exponents, *Acta Math.* **187** (2001), 275–308.

[LSW02] ———, Values of Brownian intersection exponents. III. Two-sided exponents, *Ann. Inst. H. Poincaré Probab. Statist.* **38** (2002), 109–123.

[LSW03] ———, Conformal restriction: the chordal case, *J. Amer. Math. Soc.* **16** (2003), 917–955 (electronic).

[LSW] ———, Conformal invariance of planar loop-erased random walks and uniform spanning trees, *Ann. Probab.* **32** (2004), 939–995.

[Mak98] N. G. Makarov, Fine structure of harmonic measure, *Algebra i Analiz* **10** (1998), 1–62.

[MR] D. Marshall and S. Rohde, The Loewner differential equation, preprint.

[Pom92] Ch. Pommerenke, *Boundary Behaviour of Conformal Maps*, Springer-Verlag, New York, 1992.

[RHA97] R. Raghavan, C. L. Henley, and S. L. Arouh, New two-color dimer models with critical ground states, *J. Statist. Phys.* **86** (1997), 517–550.

[RY99] D. Revuz and M. Yor, *Continuous martingales and Brownian motion*, Third edition, *Grundlehren der Mathematischen Wissenschaften, Fundamental Principles of Mathematical Sciences* **293**, Springer-Verlag, New York, 1999.

[Sch00] O. Schramm, Scaling limits of loop-erased random walks and uniform spanning trees, *Israel J. Math.* **118** (2000), 221–288.

[SD87] H. Saleur and B. Duplantier, Exact determination of the percolation hull exponent in two dimensions, *Phys. Rev. Lett.* **58** (1987), 2325–2328.

[Smi01] S. Smirnov, Critical percolation in the plane: conformal invariance, Cardy's formula, scaling limits, *C. R. Acad. Sci. Paris Sér.* I *Math.* **333** (2001), 239–244.

[SS] O. Schramm and S. Sheffield, The harmonic explorer and its convergence to SLE(4); arXiv:math.PR/0310210.

[Wer01] W. Werner, Critical exponents, conformal invariance and planar Brownian motion, in *European Congress of Mathematics, Vol.* II (*Barcelona*, 2000), *Progr. Math.* **202**, 87–103, Birkhäuser, Basel, 2001.

[WW] B. Weiland and D. B. Wilson, Winding angle variance of Fortuin-Kasteleyn contours, *Phys. Rev.* E, to appear.

(Received August 8, 2001)

Acta Math., 202 (2009), 21–137
DOI: 10.1007/s11511-009-0034-y
© 2009 by Institut Mittag-Leffler. All rights reserved

Contour lines of the two-dimensional discrete Gaussian free field

by

Oded Schramm

Microsoft Research
Redmond, WA, U.S.A.

Scott Sheffield

New York University
New York, NY, U.S.A.

Contents

1. Introduction 22
   1.1. Main result 22
   1.2. Conditional expectation and the height gap 25
   1.3. GFF definition and background 29
   1.4. SLE background and prior convergence results 32
   1.5. Precise statement of main result 35
   1.6. Outline 38
   1.7. Sequel 39
2. Preliminaries 40
   2.1. A few general properties of the DGFF 40
   2.2. Some assumptions and notation 41
   2.3. Simple random walk background 42
3. The height gap in the discrete setting 44
   3.1. A priori estimates 44
   3.2. Near independence 55
   3.3. Narrows and obstacles 59
   3.4. Barriers 65
   3.5. Meeting of random walk and interface 71
   3.6. Coupling and limit 90
   3.7. Boundary values of the interface 93
4. Recognizing the driving term 103
   4.1. About the definition of SLE($\varkappa;\varrho_1,\varrho_2$) 104
   4.2. A coordinate change 105
   4.3. The Loewner evolution of the DGFF interface 106
   4.4. Approximate diffusions 111
   4.5. Back to chordal 124
   4.6. Loewner driving term convergence 126

During the revision process of this article, Oded Schramm unexpectedly died. I am deeply indebted for all I learned working with him, for his profound personal warmth, for his legendary vision and skill. There was never a better colleague, never a better friend. He will be dearly missed. (Scott Sheffield)

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_32, © Springer Science+Business Media, LLC 2011

4.7. Carathéodory convergence 128
4.8. Improving the convergence topology 129
5. Other lattices 133
References 135

1. Introduction

1.1. Main result

The 2-dimensional massless Gaussian free field (GFF) is a 2-dimensional-time analog of Brownian motion. Just as Brownian motion is a scaling limit of simple random walks and various other 1-dimensional systems, the GFF is a scaling limit of several discrete models for random surfaces. Among these is the discrete Gaussian free field (DGFF), also called the harmonic crystal. We presently discuss the basic definitions and describe the main results of the current work, postponing an overview of the history and general context to §1.3.

Let $G=(V,E)$ be a finite graph and let $V_\partial \subset V$ be some non-empty set of vertices. Let Ω be the set of functions $h: V \to \mathbb{R}$ that are zero on V_∂. Clearly, Ω may be identified with $\mathbb{R}^{V \setminus V_\partial}$. The DGFF on G with zero boundary values on V_∂ is the probability measure on Ω whose density with respect to the Lebesgue measure on $\mathbb{R}^{V \setminus V_\partial}$ is proportional to

$$\exp\Bigg(\sum_{\{u,v\} \in E} -\frac{1}{2}(h(v)-h(u))^2 \Bigg). \tag{1.1}$$

Note that under the DGFF measure, h is a multi-dimensional Gaussian random variable. Moreover, the DGFF is a rather natural discrete model for a random field: the term $-\frac{1}{2}(h(v)-h(u))^2$ corresponding to each edge $\{u,v\}$ penalizes functions h which have a large gradient along the edge.

Now fix some function $h_\partial: V_\partial \to \mathbb{R}$, and let Ω_{h_∂} denote the set of functions $h: V \to \mathbb{R}$ that agree with h_∂ on V_∂. The probability measure on Ω_{h_∂} whose density with respect to the Lebesgue measure on $\mathbb{R}^{V \setminus V_\partial}$ is proportional to (1.1) is the DGFF with boundary values given by h_∂.

Let TG be the usual triangular grid in the complex plane, i.e., the graph whose vertex set is the integer span of 1 and $e^{i\pi/3} = \frac{1}{2}(1+i\sqrt{3})$, with straight edges joining v and w whenever $|v-w|=1$. A TG-*domain* $D \subset \mathbb{R}^2 \cong \mathbb{C}$ is a domain whose boundary is a simple closed curve comprised of edges and vertices in TG. Let $V = V_{\overline{D}}$ be the set of TG-vertices in the closure of D, let $G = G_D$ be the induced subgraph of TG with vertex set $V_{\overline{D}}$, and write $V_\partial = \partial D \cap V_{\overline{D}}$. While introducing our main results, we will focus on graphs G_D and boundary sets V_∂ of this form (though analogous results hold if we replace TG with another doubly periodic planar graph; see §1.5).

We may assume that any function $f: V \to \mathbb{R}$ is interpolated to a continuous function on the closure of D which is affine on each triangle of TG. We often interpret f as a surface embedded in 3 dimensions and refer to $f(v)$ as the *height* of the surface at v.

Let $\partial D = \partial_+ \cup \partial_-$ be a partition of the boundary of a TG-domain D into two disjoint arcs whose endpoints are midpoints of two distinct TG-edges in ∂D. Fix two constants $a, b > 0$. Let h be an instance of the DGFF on (G_D, V_∂), with boundary function h_∂ equal to $-a$ on the vertices in ∂_- and equal to b on the vertices in ∂_+. Then h (linearly interpolated on triangles) almost surely assumes the value zero on a unique piecewise linear path γ_h connecting the two boundary edges containing endpoints of ∂_+.

In §1.4, we will briefly review the definition of SLE(4) (a particular type of random chordal path connecting a pair of boundary points of D whose randomness comes from a 1-dimensional Brownian motion), along with the variants of SLE(4) denoted SLE(4; ϱ_1, ϱ_2). Our main result, roughly stated, is the following.

THEOREM 1.1. *Let D be a* TG*-domain, $\partial D = \partial_+ \cup \partial_-$ and let h and γ_h be as above. There is a constant $\lambda > 0$ such that if $a = b = \lambda$, then as the triangular mesh gets finer, the random path γ_h converges in distribution to* SLE(4). *If $a, b \geqslant \lambda$ are not assumed to equal λ, then the convergence is to* SLE$(4; a/\lambda - 1, b/\lambda - 1)$.

See §1.5 for a more precise version, which describes the topology under which the convergence is attained. As explained there, we can also prove convergence in a weaker form when the conditions $a, b \geqslant \lambda$ are relaxed.

We will elaborate on the role of the constant λ in §1.2. This constant depends only on the lattice used. Although we do not prove it in this paper, for the triangular grid the value of λ is $\lambda_{\mathrm{TG}} := 3^{-1/4}\sqrt{\frac{1}{8}\pi}$ (see §1.7).

Figure 1.1 illustrates a dual perspective on an instance of γ_h. Here, each vertex in the closure of a rhombus-shaped TG-domain D is replaced with a hexagon in the honeycomb lattice. Call hexagons *positive* or *negative* according to the sign of h. Then there is a cluster of positive hexagons that includes the positive boundary hexagons, a similar cluster of negative hexagons, and a path γ forming the boundary between these two clusters. Figure 1.1 depicts a computer generated instance of the DGFF—with $\pm\lambda$ boundary conditions—and the corresponding γ. Followed from bottom to top, the interface γ turns right when it hits a negative hexagon, left when it hits a positive hexagon. It closely tracks the boundary-hitting zero contour line γ_h in the following sense: the edges in γ are the duals of the edges of TG that are crossed by γ_h. This is because h is almost surely non-zero at each vertex in V, so whenever a zero contour line contains a point on an edge of TG, h must be positive on one endpoint of that edge and negative on the other; hence the dual of that edge separates a positive hexagon from a

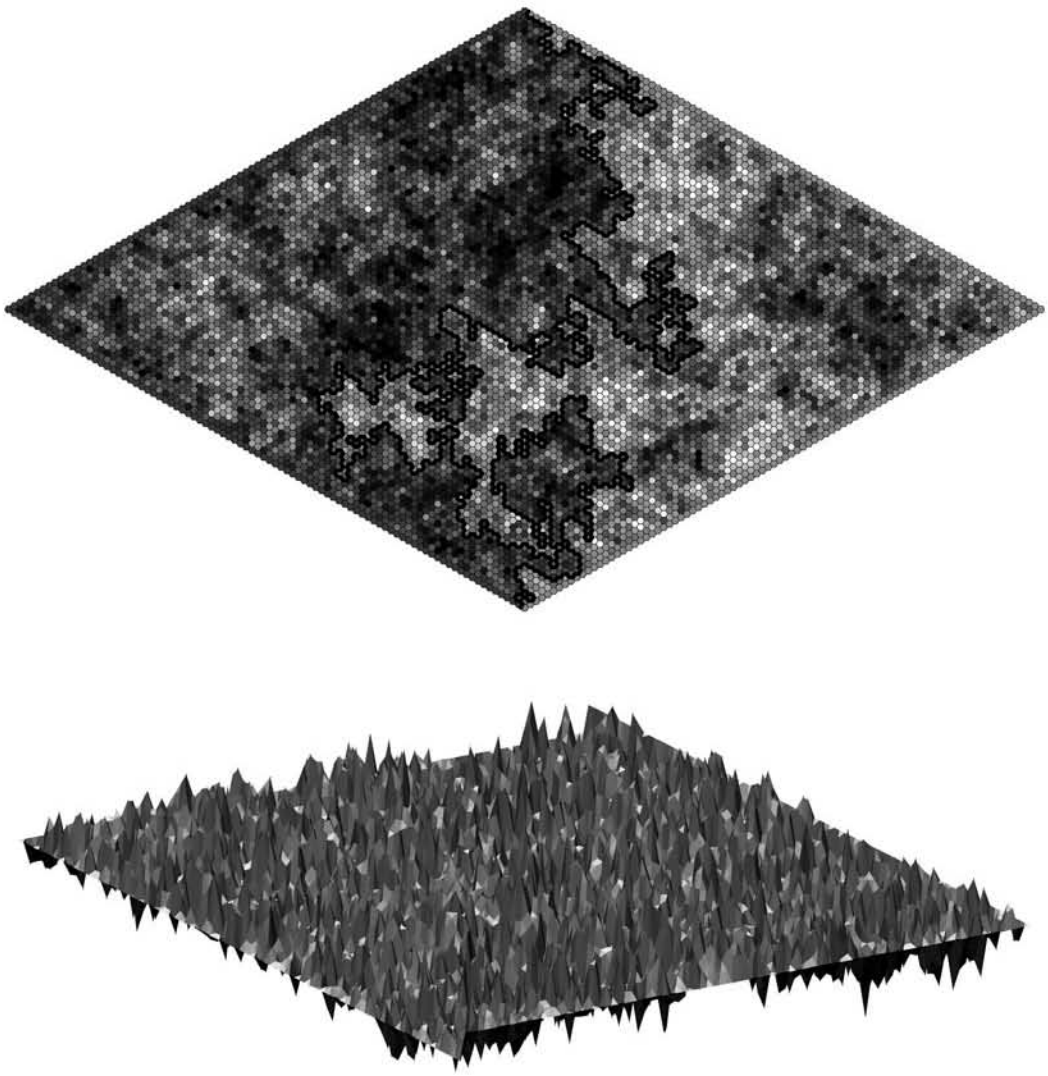


Figure 1.1. (a) DGFF on a 90×90 hexagon array with boundary values λ on the right and $-\lambda$ on the left; faces shaded by height. (b) Surface plot of DGFF.

negative hexagon.

In the fine mesh limit, there will be no difference between γ and γ_h. Thus (by Theorem 1.1) the path in Figures 1.1 (a) and 1.2 (a) approximates SLE(4), while the path of Figure 1.3 (a) approximates SLE$(4; 2, 2)$. We will state and prove most of our results in terms of the dual perspective displayed in the figures.

1.2. Conditional expectation and the height gap

We derive the following well-known facts as a warm-up in §2.1 (see also, e.g., [Gi]).

Boundary influence: The law of the DGFF with boundary conditions $h_\partial: V_\partial \to \mathbb{R}$ is the same as that of the DGFF with boundary conditions 0 *plus* a deterministic function $\tilde{h}_\partial: V \to \mathbb{R}$ which is the unique discrete-harmonic interpolation of h_∂ to V. (By *discrete-harmonic* we mean that for each $v \in V \setminus V_\partial$, the value $h(v)$ is equal to the average value of $h(w)$ over w adjacent to v.) In particular, the expected value of $h(v)$ is discrete-harmonic in $V \setminus V_\partial$.

Markov property: Let $h: V \to \mathbb{R}$ be a random function whose law is the DGFF on G with some boundary values h_∂ on V_∂. Then, given the values of h on a superset $V_0 \supset V_\partial$, the conditional law of h is that of a DGFF on G with boundary set V_0 and with boundary values equal to the given values.

From these facts it follows that conditioned on the path γ described in the previous section and on the values of h on the hexagons adjacent to γ, the expected value of h is discrete-harmonic in the remainder of G_D. Figures 1.2 and 1.3 illustrate the expected value of h conditioned on the values of h on the hexagons adjacent to γ.

The reader may observe in Figure 1.2 that although the expected value of h given the values along γ varies a great deal among hexagons close to γ, the expected value at five or ten lattice spacings away from γ appears to be roughly constant along either side of γ. On the other hand, in Figure 1.3, away from γ, the expected height appears to be a smooth but non-constant function. In a sense we make precise in §3 (see Theorem 3.28), the values $-\lambda$ and λ describe the expected value of h, conditioned on γ, at the vertices near (but not microscopically near) the left and right sides of γ; in the fine mesh limit there is thus an "expected height gap" of 2λ between the two sides of γ. In Figure 1.2 the height expectation appears constant away from γ, because the boundary values of $\pm\lambda$ are the same as the expected values near (but not microscopically near) γ.

Once we have established the height gap result, the proof of Theorem 1.1 (at least for the simplest case that the boundary conditions are $-\lambda$ and λ) is similar to the proof that the harmonic explorer converges to SLE(4), as given by the present authors in [SS],

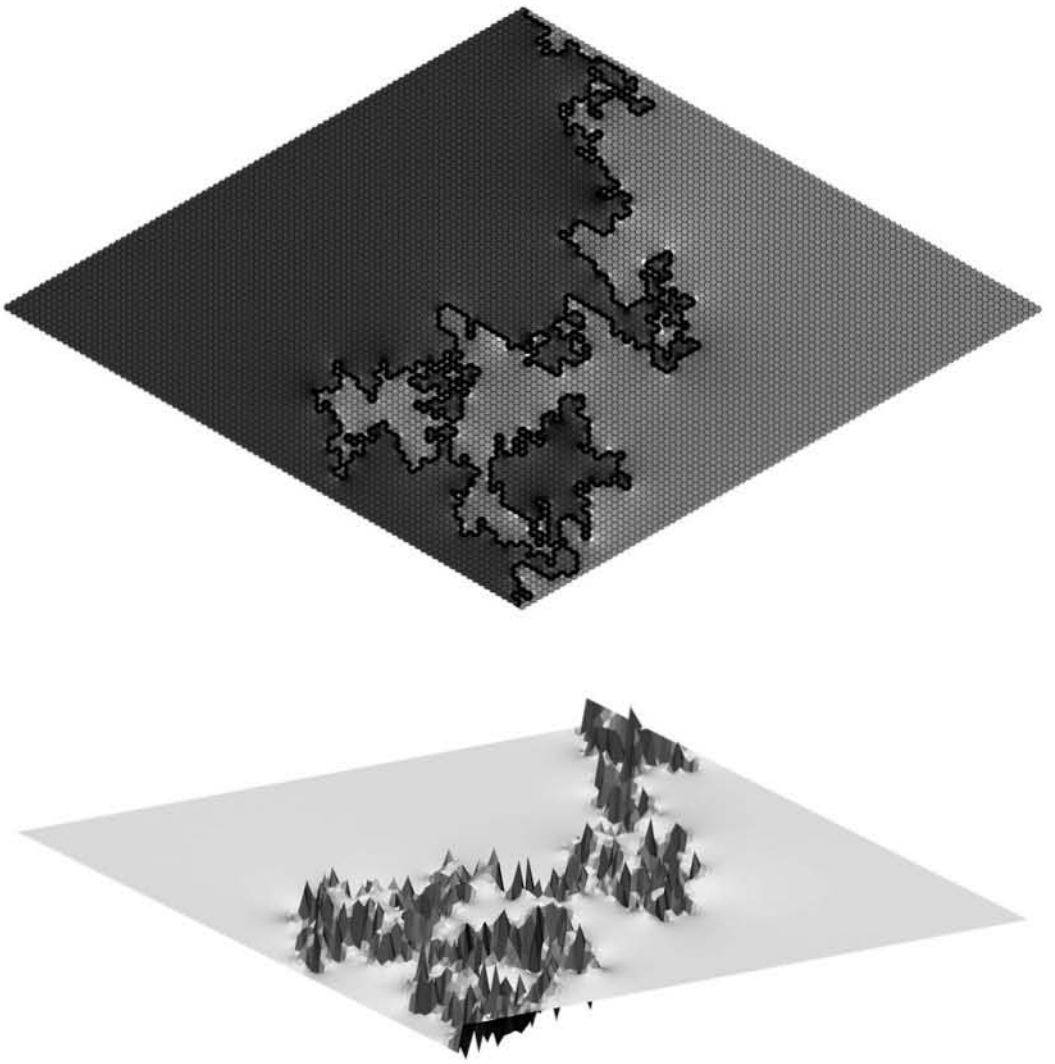


Figure 1.2. (a) Expectation of DGFF with boundary values $\pm\lambda$ given its values at hexagons bordering the interface. (b) Surface plot of the above.

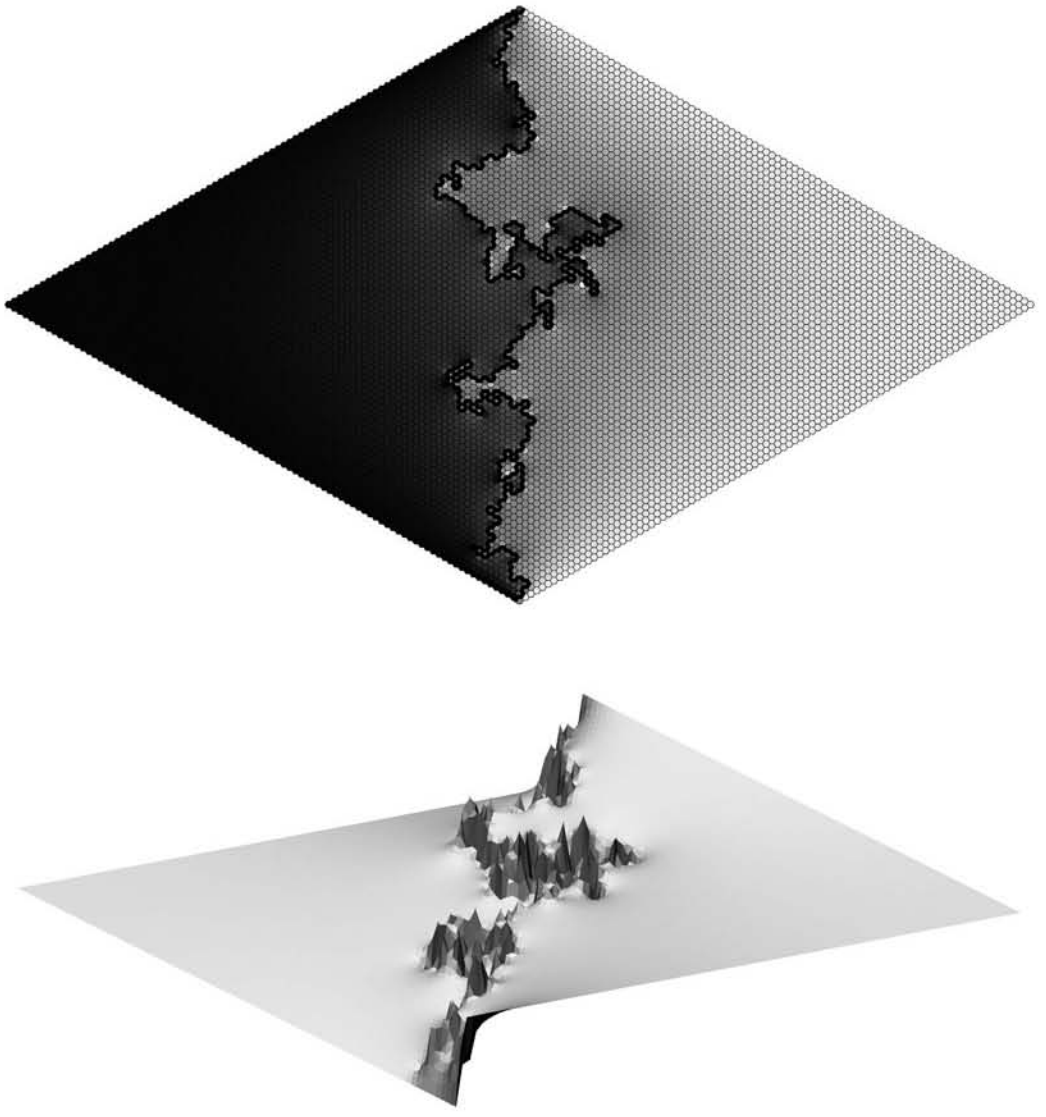


Figure 1.3. (a) Expectation of DGFF given its values at hexagons bordering the interface; exterior boundary values are -3λ on left, and 3λ on right. (b) Surface plot of the above.

which, in turn, follows the same strategy as the proof of convergence of the loop-erased random walk to SLE(2) and the uniform spanning tree Peano curve to SLE(8) in [LSW4].

We will now briefly describe some of the key ideas in the proof of the height gap result. The main step is to show that if one samples a vertex z on γ according to discrete-harmonic measure viewed from a typical point far away from γ, then the absolute value of $h(z)$ is close to independent of the values of h (and the geometry of γ) at points that are *not* microscopically close to z. In other words, if we start a random walk S at a typical point in the interior of D and stop the first time it hits a vertex z which either belongs to V_∂ or corresponds to a hexagon incident to γ, then (conditioned on $z \notin V_\partial$) the random variable $|h(z)|$ (and in particular its conditional expectation) is close to independent of the behavior of γ and h at vertices far away from z.

To prove this, we will actually prove something stronger, namely that (up to multiplication by -1) the collection of all of the values of h (and the geometry of γ) in a microscopic neighborhood of z is essentially independent of the values of h (and the geometry of γ) at points that are not microscopically close to z. One consequence of our analysis is Theorem 3.21, which states that if one takes z to be the origin of a new coordinate system and conditions on the behavior of γ and S outside of a ball of radius R centered at z and S starts outside that ball, then as R tends to infinity the conditional law of the interface γ has a weak limit (which is independent of the sequence of boundary conditions chosen), which is the law of a random infinite path γ on the honeycomb grid TG^* (almost surely containing an edge adjacent to the hexagon centered at the origin $z{=}0$). We will define a function of such infinite paths γ which (in a certain precise sense) describes the expected value of $|h(z)|$ conditioned on γ; the value λ is the expectation of this function when γ is chosen according to the limiting measure described above.

We remark that many important problems in statistical physics involve classifying the measures that can arise as weak limits of Gibbs measures on finite systems. In such problems, showing the uniqueness of the limiting measure often involves proving that properties of a random system near the origin are approximately independent of the properties of the system far away from the origin. In our case, we need to prove that in some sense the behavior of the triple (h, γ, S) near the the origin (i.e., the first point S hits γ) is close to independent of the behavior of (h, γ, S) far from the origin.

Very roughly speaking, our strategy will be to describe the joint law of (h, γ, S) near the origin and (h, γ, S) far from the origin by considering a different measure in which the two are independent and weighting it by the probability that the inside and outside configurations properly "hook up" with one another. To get a handle on these "hook up" probabilities, we will need to develop various techniques to control the probabilities (conditioned on the values of h on certain sets) that certain zero-height level lines hook

up with one another, as well as the probabilities that these level lines avoid certain regions. We will also need bounds on the probability that there exist clusters of positive or negative hexagons crossing certain regions; these are roughly in the spirit of the Russo–Seymour–Welsh theorems for percolation, but the proofs are entirely different. All of the height gap related results are proved in §3.

1.3. GFF definition and background

To help put our DGFF theorems in context and provide further intuition, we now briefly recall the definition of the (continuum) GFF and mention some basic facts described, e.g., in [Sh]. Let $H_s(D)$ be the set of smooth functions supported on compact subsets of a planar domain D, and let $H(D)$ be its Hilbert space completion under the *Dirichlet inner product* $(f,g)_\nabla = \int_D \nabla f \cdot \nabla g \, dx$, where dx refers to area measure. We define an instance of the Gaussian free field to be the formal sum

$$h = \sum_{j=1}^{\infty} \alpha_j f_j,$$

where the α_j's are independent identically distributed 1-dimensional standard (unit variance, zero mean) Gaussians and the f_j's are an orthonormal basis for $H(D)$. Although the sum does not converge pointwise or in $H(D)$, it does converge in the space of distributions [Sh]. In particular, the sum

$$(h,g)_\nabla := \sum_{j=1}^{\infty} \alpha_j (f_j, g)_\nabla$$

is almost surely convergent for every $g \in H_s(D)$.

It is worthwhile to take a moment to compare with the situation where D is 1-dimensional. If D were a bounded open interval in $\mathbb{R}$, then the partial sums of

$$h = \sum_{j=1}^{\infty} \alpha_j f_j$$

would almost surely converge uniformly to a limit, whose law is that of the Brownian bridge, having the value zero at the interval's endpoints. If D were the interval $(0,\infty)$, then the partial sums would converge (uniformly on compact sets) to a function whose law is that of ordinary Brownian motion B_t, indexed by $t \in [0,\infty)$, with $B_0 = 0$ [Sh].

Let g be a *conformal* (i.e., bijective analytic) map from D to another planar domain D'. When g is a rotation, dilation, or translation, it is obvious that

$$\int_{D'} \nabla(f_1 \circ g^{-1}) \cdot \nabla(f_2 \circ g^{-1}) \, dx = \int_D (\nabla f_1 \cdot \nabla f_2) \, dx$$

for any $f_1, f_2 \in H_s(D)$, and an elementary change of variables calculation gives this equality for any conformal g. Taking the completion to $H(D)$, we see that the Dirichlet inner product—and hence the 2-dimensional GFF—is invariant under conformal transformations of D.

Up to a constant, the DGFF on a TG-domain D can be realized as a projection of the GFF on D onto the subspace of $H(D)$ consisting of functions which are continuous and are affine on each triangle of D [Sh]. Note that if f is such a function, then

$$(f,f)_\nabla = \frac{\sqrt{3}}{6}\sum(|f(k)-f(j)|^2+|f(l)-f(j)|^2+|f(l)-f(k)|^2),$$

where the sum is over all triangles (j,k,l) in $V_{\overline{D}}$. This is because the area of each triangle is $\frac{1}{4}\sqrt{3}$ and the norm of the gradient squared in the triangle is

$$\tfrac{2}{3}(|f(k)-f(j)|^2+|f(l)-f(j)|^2+|f(l)-f(k)|^2).$$

Since each interior edge of D is contained in two triangles, for such f,

$$\|f\|_\nabla^2 = \frac{1}{\sqrt{3}}\sum_{\{j,k\}\in E_I}|f(k)-f(j)|^2+\frac{1}{2\sqrt{3}}\sum_{\{j,k\}\in E_\partial}|f(k)-f(j)|^2, \tag{1.2}$$

where E_I and E_∂ are the interior and boundary (undirected) edges of TG in $\overline{D}$. We will refer to the sum $\sum_{E_I}|f(k)-f(j)|^2$ as the *discrete Dirichlet energy* of f. It is equivalent—up to the constant factor $3^{-1/2}$ and an additive term depending only on the boundary values of f—to the *Dirichlet energy* $(f,f)_\nabla$ of the piecewise affine interpolation of f to D.

The above analysis suggests a natural coupling between the GFF and a sequence of DGFF approximations to the GFF (obtained by taking finer mesh approximations of the same domain). The GFF can also be obtained as a scaling limit of other discrete random surface models (e.g., solid-on-solid, dimer-height-function, and $\nabla\phi$-interface models) [Ke] [NS], [Sp]. Its Laplacian is a scaling limit of some Coulomb gas models, which describe random electrostatic charge densities in 2-dimensional domains [F], [FS], [Ko], [KT], [Sp]. Physicists often use heuristic connections to the GFF to predict properties of 2-dimensional statistical physics models that are not obviously random surfaces or Coulomb gases (e.g., Ising and Potts models, $O(n)$ loop models) [dN], [DMS], [D], [Ka], [KN], [N1], [N2]. As a model for the field theory of non-interacting massless bosons, the GFF is a starting point for many constructions in quantum field theory, conformal field theory, and string theory [BPZ], [DMS], [Ga], [GJ].

Because of the conformal invariance of the GFF, physicists and mathematicians have hypothesized that discrete random surface models that are believed or known to converge

to the Gaussian free field (e.g., the discrete Gaussian free field, the height function of the oriented $O(n)$ loop model with $n=2$, height functions for domino and lozenge tilings) have level sets with conformally invariant scaling limits [Co], [DS1], [DS2], [SD], [KDH], [HK], [Ke], [KH], [KHS], [N1]. Our results confirm this hypothesis for the discrete Gaussian free field.

Various properties of the DGFF contour lines (such as winding exponents and the fact that the fractal dimension is $\frac{3}{2}$) have been predicted correctly in the physics literature [Co], [DS2], [DS1], [SD], [HK], [KDH], [KH], [KHS], [N1]. The techniques used to make these predictions are also described in detail in the survey papers [D], [KN], [N2]. Analogous results about winding exponents and fractal dimension have now been proved rigorously for SLE [Sch], [RS], [B].

The study of level lines of the DGFF and related random surfaces is also related to the study of equipotential lines of random charge distributions in statistical physics. The so-called 2-*dimensional Coulomb gas* is a model for electrostatics in which the force between charged particles is inversely proportional to the distance between them. In this model, a continuous function $f \in H_s(D)$ is the *Coulomb gas electrostatic potential function* ("grounded" at the boundary of D) of $-\Delta f$, when Δf is interpreted as a charge density function. The value $(f,f)_\nabla$ is then the total potential energy—also called the *energy of assembly* of the charge distribution $-\Delta f$. In the Coulomb gas model, this is the amount of energy required to move from a configuration in which the charge density is zero throughout D to a configuration in which the charge density is given by $-\Delta f$.

In statistical physics, it is often natural to consider a probability distribution on configurations in which the probability of a configuration with potential energy H is proportional to e^{-H}. If ϱ is a smooth charge distribution, then its energy of assembly is given by $(-\Delta^{-1}\varrho, -\Delta^{-1}\varrho)_\nabla = (\varrho, -\Delta^{-1}\varrho)$; if we define ϱ to be the standard Gaussian in $\Delta H(D)$ determined by this quadratic form, then ϱ is the Laplacian of the Gaussian free field (which, like the GFF itself, is well defined as a random distribution but not as a function). In other words, the Laplacian of a Gaussian free field is a random distribution that we may interpret as a model for random charge density in a statistical physical Coulomb gas.

However, we stress that when physicists refer to the Coulomb gas method for $O(n)$ model computations, they typically have in mind a more complicated Coulomb gas model in which the charges are required to be discrete (i.e., ϱ is required to be a sum of unit positive and negative point masses) and hard core constraints may be enforced.

The surveys [BEF], [Gi], [Sp] contain additional references on lattice spin models that have the GFF as a scaling limit and Coulomb gas models that have its Laplacian as a scaling limit—for example, the *harmonic crystal* (also known as the *discrete Gaussian free*

field) with quadratic nearest-neighbor potential, the more general *anharmonic crystal*, the *discrete-height Gaussian* (where h is a function on a lattice, with values restricted to integers), the *Villain gas* (where h is a function on a lattice and the values of its discrete Laplacian $\mathfrak{p}=-\Delta h$ are restricted to integers), and the *hard core Coulomb gas* (where h is a function on a lattice and its discrete Laplacian $\mathfrak{p}=-\Delta h$ is ± 1 valued).

The physics literature on applications of the GFF to field theory and statistical physics is large, and the authors themselves are only familiar with parts of it. Outside of these areas, there is a body of experimental and computational research on contour lines of random topographical surfaces, such as the surface of the earth. Mandelbrot's famous *How Long Is the Coast of Britain?* [M], which prefigured the notion of "fractal" introduced by Mandelbrot years later, is an early example. The results about contour lines in these studies (including fractal dimension computations) are less detailed than the ones provided here and are not all mathematically rigorous. However, some of the models are similar in spirit to the GFF, involving functions whose Fourier coefficients are independent Gaussians. An eclectic overview of this literature appears in [I].

1.4. SLE background and prior convergence results

We now give a brief definition of (chordal) SLE($\varkappa$) for $\varkappa>0$. See also the surveys [W], [KN], [L4], [Ca] or [L3]. The discussion below along with further discussion of the special properties of SLE(4) appears in another paper by the current authors [SS]. That paper shows that SLE(4) is the scaling limit of a random interface called the harmonic explorer (designed in part to be a toy model for the DGFF contour line addressed here).

Let $T>0$. Suppose that $\gamma:[0,T]\to\overline{\mathbb{H}}$ is a continuous simple path in the closed upper half-plane $\overline{\mathbb{H}}$ which satisfies $\gamma[0,T]\cap\mathbb{R}=\{\gamma(0)\}=\{0\}$. For every $t\in[0,T]$, there is a unique conformal homeomorphism $g_t:\mathbb{H}\setminus\gamma[0,t]$ which satisfies the so-called *hydrodynamic* normalization at infinity

$$\lim_{z\to\infty}(g_t(z)-z)=0.$$

The limit

$$\operatorname{cap}_\infty(\gamma[0,t]):=\lim_{z\to\infty}\frac{z(g_t(z)-z)}{2}$$

is real and monotone increasing in t. It is called the (half-plane) *capacity* of $\gamma[0,t]$ from ∞, or just *capacity*, for short. Since $\operatorname{cap}_\infty(\gamma[0,t])$ is also continuous in t, it is natural to reparameterize γ so that $\operatorname{cap}_\infty(\gamma[0,t])=t$. Loewner's theorem states that in this case the maps g_t satisfy his differential equation

$$\partial_t g_t(z)=\frac{2}{g_t(z)-W_t},\quad g_0(z)=z, \tag{1.3}$$

where $W_t=g_t(\gamma(t))$. (Since $\gamma(t)$ is not in the domain of definition of g_t, the expression $g_t(\gamma(t))$ should be interpreted as a limit of $g_t(z)$ as $z\to\gamma(t)$ inside $\mathbb{H}\setminus\gamma[0,t]$. This limit does exist.) The function $t\mapsto W_t$ is continuous in t, and is called the *driving parameter* for γ.

One may also try to reverse the above procedure. Consider the Loewner evolution defined by the ordinary differential equation (ODE) (1.3), where W_t is a continuous, real-valued function. For a fixed z, the evolution defines $g_t(z)$ as long as $|g_t(z)-W_t|$ is bounded away from zero. For $z\in\overline{\mathbb{H}}$ let τ_z be the first time $t\geqslant 0$ in which $g_t(z)$ and W_t collide, or set $\tau_z=\infty$ if they never collide. Then $g_t(z)$ is well defined on $\{z\in\overline{\mathbb{H}}:\tau_z\geqslant t\}$. The set $K_t:=\{z\in\overline{\mathbb{H}}:\tau_z\leqslant t\}$ is sometimes called the *evolving hull* of the evolution. In the case discussed above where the evolution is generated by a simple path γ parameterized by capacity and satisfying $\gamma(t)\in\mathbb{H}$ for $t>0$, we have $K_t=\gamma[0,t]$.

The path of the evolution is defined as $\gamma(t)=\lim_{z\to W_t} g_t^{-1}(z)$, where z tends to W_t from within the upper half-plane $\mathbb{H}$, provided that the limit exists and is continuous. However, this is not always the case. The process (chordal) SLE($\varkappa$) in the upper half-plane, beginning at 0 and ending at ∞, is the path $\gamma(t)$ when W_t is $B_t\sqrt{\varkappa}$, where $B_t=B(t)$ is a standard 1-dimensional Brownian motion. (*Standard* means $B(0)=0$ and $\mathbf{E}[B(t)^2]=t$, $t\geqslant 0$. Since $(B_t\sqrt{k}:t\geqslant 0)$ has the same distribution as $(B_{\varkappa t}:t\geqslant 0)$, taking $W_t=B_{\varkappa t}$ is equivalent.) In this case, almost surely $\gamma(t)$ does exist and is a continuous path. See [RS] ($\varkappa\neq 8$) and [LSW4] ($\varkappa=8$).

We now define the processes SLE($\varkappa;\varrho_1,\varrho_2$). Given a Loewner evolution defined by a continuous W_t, we let x_t and y_t be defined by $x_t:=\sup\{g_t(x):x<0 \text{ and } x\notin K_t\}$ and $y_t:=\inf\{g_t(x):x>0 \text{ and } x\notin K_t\}$. When the Loewner evolution is generated by a simple path $\gamma(t)$ satisfying $\gamma(t)\in\mathbb{H}$ for $t>0$, these points x_t and y_t can be thought of as the two images of 0 under g_t. Note that, by (1.3),

$$\partial_t x_t=\frac{2}{x_t-W_t}\quad\text{and}\quad\partial_t y_t=\frac{2}{y_t-W_t} \tag{1.4}$$

for all t such that $x_t<W_t<y_t$. Beginning from an initial time r for which $x_r<W_r<y_r$, we define SLE($\varkappa;\varrho_1,\varrho_2$) to be the evolution that makes (x_t,W_t,y_t) a solution to the stochastic differential equation (SDE) system

$$dW_t=\sqrt{\varkappa}\,dB_t+\frac{\varrho_1\,dt}{W_t-x_t}+\frac{\varrho_2\,dt}{W_t-y_t},\quad dx_t=\frac{2\,dt}{x_t-W_t},\quad dy_t=\frac{2\,dt}{y_t-W_t}, \tag{1.5}$$

noting that existence and uniqueness of solutions to this SDE (at least from the initial time r until the first $s>r$ for which either $x_s=W_s$ or $W_s=y_s$) follow easily from standard results in [RY]. (The x_t and y_t are called *force points* because they apply a "force" affecting the drift of the process W_t by an amount inversely proportional to their distance from W_t.)

Some subtlety is involved in extending the definition of SLE($\varkappa; \varrho_1, \varrho_2$) beyond times when W_t hits the force points, and in starting the process from the natural initial values $x_0{=}W_0{=}y_0{=}0$. This is closely related to the issues which come up when defining the Bessel processes of dimension less than 2 and will be discussed in more detail in §4.

Although many random self-avoiding lattice paths from the statistical physics literature are conjectured to have forms of SLE as scaling limits, rigorous proofs have thus far appeared only for a few cases: site percolation cluster boundaries on the hexagonal lattice (SLE(6), [Sm]; see also [CN]), branches (loop-erased random walk) and outer boundaries (random Peano curves) of uniform spanning trees (forms of SLE(2) and SLE(8), respectively, [LSW4]), the harmonic explorer (SLE(4), [SS]), and boundaries of simple random walks (forms of SLE$\left(\frac{8}{3}\right)$, [LSW3]).

In the latter case, conformal invariance properties follow almost immediately from the conformal invariance of 2-dimensional Brownian motion. In each of the other cases listed above, the initial step of the proof is to show that a certain function of the partially generated paths $\gamma([0,t])$, which is a martingale in t when γ is SLE($\varkappa$) for the appropriate $\varkappa$, has a discrete analog which is (approximately or exactly) a martingale for the discrete paths and is approximately equivalent to the continuous version in the fine mesh limit. For loop-erased random walk, harmonic explorer, and uniform spanning tree Peano curves, this initial step is the easy part of the argument; it follows almost immediately from the fact that simple random walk converges to Brownian motion. The analogous step for site percolation on the hexagonal lattice, as given by [Sm], is an ingenious but nonetheless short and simple argument.

By contrast, the analogous step in this paper (which requires the proof of the height gap lemma, as given in §3) is quite involved; it is the most technically challenging part of the current work and includes many new techniques and lemmas about the geometry of DGFF contours that we hope are interesting for their own sake.

Another way in which the DGFF differs from percolation, the harmonic explorer, and the uniform spanning tree is that it has a natural continuum analog (the GFF) which can be easily rigorously constructed without any reference to SLE, and which is itself (like Brownian motion) an object of great significance. It becomes natural to ask whether the DGFF results enable us to define the "contour lines" of the continuum GFF in a canonical way; we plan to answer this question (affirmatively) in a subsequent work (see §1.7).

A final difference is that, for the DGFF, there is a continuum of choices for left and right boundary conditions (a and b) which are equally natural a priori, so we are led to consider a family of paths SLE($4; a/\lambda{-}1, b/\lambda{-}1$) instead of simply SLE(4). (The case $a{=}b{=}0$ is particularly natural; see Figure 1.4.) In these processes, the driving

parameters W_t are generally no longer Brownian motions (rather, they are continuous semimartingales with constant quadratic variation and a drift term that can become singular on a fractal set). Proving driving parameter convergence to these processes requires some rather general convergence infrastructure (§4.4), which we hope will be useful in other settings as well.

1.5. Precise statement of main result

Let $\mathbb{H}$ be the upper half-plane. Let D be any TG-domain and let $\partial D = \partial_+ \cup \partial_-$ be a partition of the boundary of D into two disjoint arcs whose endpoints are midpoints of two TG-edges contained in ∂D. As before, let V denote the vertices of TG in $\overline{D}$. Let $h_\partial = -a$ on $\partial_- \cap V$ and $h_\partial = b$ on $\partial_+ \cap V$, where a and b are positive constants.

Let $h: V \to \mathbb{R}$ be an instance of the DGFF with boundary conditions h_∂. Let ϕ_D be any conformal map from D to $\mathbb{H}$ that maps ∂_+ bijectively onto the positive real ray $(0, \infty)$. (Note that ϕ_D is unique up to positive scaling.)

There is almost surely a unique interface $\gamma \subset \overline{D}$ between hexagons in the dual grid TG^* containing TG-vertices where h is positive and such hexagons where h is negative, such that the endpoints of γ are on ∂D. In fact, the endpoints of γ are the same as the endpoints of ∂_+. (As mentioned above, this interface γ stays within a bounded distance from the zero-height contour line γ_D of the affine interpolation of h.) Now, $\phi_D \circ \gamma$ is a random path on $\mathbb{H}$ connecting 0 to ∞. We will show that this path converges to a form of SLE(4). Rather than considering a fixed domain $\widehat{D}$ and a sequence of discrete domains D_n approximating $\widehat{D}$, with the mesh tending to 0, we will employ a setup that is more general in which the mesh is fixed (the triangular lattice will not be rescaled), and we consider domains D that become "larger". The correct sense of "large" is measured by

$$r_D = r_{D,\phi} := \operatorname*{rad}_{\phi_D^{-1}(i)}(D),$$

where $\mathrm{rad}_x(D)$ denotes the radius of D viewed from x, i.e., $\inf_{y \notin D} |x-y|$. Of course, if $\phi_D^{-1}(i)$ is at a bounded distance from ∂D, then the image of the triangular grid under ϕ_D is not fine near i, and there is no hope for approximating SLE by $\phi_D \circ \gamma$.

We have chosen to use $\mathbb{H}$ as our canonical domain (mapping all other paths into $\mathbb{H}$), because it is the most convenient domain in which to define chordal SLE. However, to make the completion of $\mathbb{H}$ a compact metric space, we will endow $\mathbb{H}$ with the metric it inherits from its conformal map onto the unit disk $\mathbb{U}$. Namely, we let $d_*(\cdot, \cdot)$ be the metric on $\overline{\mathbb{H}} \cup \{\infty\}$ given by $d_*(z,w) = |\Psi(z) - \Psi(w)|$, where $\Psi(z) := (z-i)/(z+i)$ maps $\overline{\mathbb{H}} \cup \{\infty\}$ onto $\overline{\mathbb{U}}$. If $z \in \overline{\mathbb{H}}$, then $d_*(z_n, z) \to 0$ is equivalent to $|z_n - z| \to 0$, and $d_*(z_n, \infty) \to 0$ is equivalent to $|z_n| \to \infty$.

If γ_1 and γ_2 are distinct unparameterized simple paths in $\overline{\mathbb{H}}$, then we define $d_{\mathcal{U}}(\gamma_1,\gamma_2)$ to be the infimum over all pairs (η_1,η_2) of parameterizations of γ_1 and γ_2 in $[0,1]$ (i.e., $\eta_j\colon[0,1]\to\overline{\mathbb{H}}$ is a simple path satisfying $\eta_j([0,1])=\gamma_j$ for $j=1,2$) of the uniform distance $\sup\{d_*(\eta_1(t),\eta_2(t)):t\in[0,1]\}$ with respect to the metric d_*.

Our strongest result is in the case where $a,b\geqslant\lambda$. We prove the following result.

THEOREM 1.2. *There is a constant $\lambda>0$ such that if $a,b\geqslant\lambda$, then, as $r_D\to\infty$, the random paths $\phi_D\circ\gamma$ described above converge in distribution to* $\mathrm{SLE}(4;a/\lambda-1,b/\lambda-1)$ *with respect to the metric $d_{\mathcal{U}}$.*

In other words, for every $\varepsilon>0$ there is some $R=R(\varepsilon)$ such that if $r_D>R$, then there is a coupling of $\phi_D\circ\gamma$ and a path γ_{SLE} whose law is that of $\mathrm{SLE}(4;a/\lambda-1,b/\lambda-1)$ *such that*

$$\mathbf{P}[d_{\mathcal{U}}(\phi_D\circ\gamma,\gamma_{\mathrm{SLE}})>\varepsilon]<\varepsilon.$$

We first comment that it follows that when r_D is large, γ is "close" to $\phi_D^{-1}\circ\gamma_{\mathrm{SLE}}$. For example, if $r_D\to\infty$ and $r_D^{-1}D$ tends to a bounded domain $\widehat{D}$ whose boundary is a simple closed path in such a way that the boundaries of the domains may be parameterized to give uniform convergence of parameterized paths and if $r_D^{-1}\partial_+$ converges, then $r_D^{-1}\gamma$ converges in law to the corresponding SLE in $\widehat{D}$. To prove this from Theorem 1.2, we only need to note that in this case the maps $r_D^{-1}\phi_D^{-1}$ converge uniformly in $\overline{\mathbb{H}}\cup\{\infty\}$ (see, e.g., [P, Proposition 2.3]).

When we relax the assumption $a,b\geqslant\lambda$ to $a,b>0$, we still prove some sort of convergence to $\mathrm{SLE}(4;a/\lambda-1,b/\lambda-1)$, but with respect to a weaker topology. In fact, we can allow a and b to be zero or even slightly negative, but in this case we need to appropriately adjust the above definition of the interface γ. Say that a hexagon in the hexagonal grid TG^* dual to TG is positive if either the center v of the hexagon is in D and $h(v)>0$, or $v\in\partial_+$. Likewise, say that the hexagon is negative if $v\in D$ and $h(v)<0$, or $v\in\partial_-$. Let γ be the unique oriented path in TG^* that joins the two endpoints of ∂_+, has only positive hexagons adjacent to its right-hand side and only negative hexagons adjacent to its left-hand side. (If $a,b>0$, this definition clearly agrees with the previous definition of γ.) We prove the following.

THEOREM 1.3. *For every constant $\bar{\Lambda}>0$ there is a constant $\Lambda_0=\Lambda_0(\bar{\Lambda})>0$ such that if $a,b\in[-\Lambda_0,\bar{\Lambda}]$ and γ is the* DGFF *interface defined above, then as $r_D\to\infty$ the Loewner driving term $\widehat{W}_t$ of $\phi_D\circ\gamma$, parameterized by capacity from ∞, converges in law to the driving term W_t of* $\mathrm{SLE}(4;a/\lambda-1,b/\lambda-1)$ *with respect to the topology of locally uniform convergence. That is, for every $T,\varepsilon>0$ there is some $R>0$ such that if $r_D>R$, then γ and* $\mathrm{SLE}(4;a/\lambda-1,b/\lambda-1)$ *may be coupled so that with probability at least $1-\varepsilon$,*

$$\sup\{|\widehat{W}_t-W_t|:t\in[0,T]\}<\varepsilon.$$

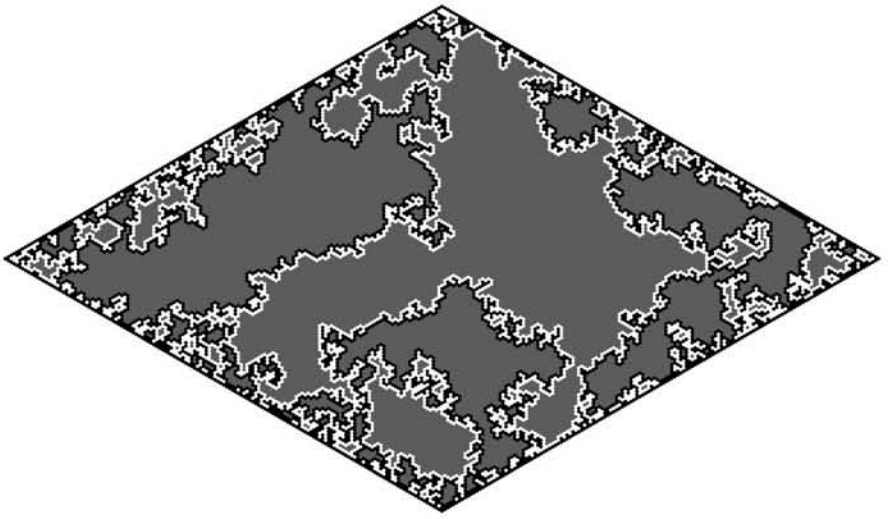


Figure 1.4. Zero-height interfaces starting and ending on the boundary, shown for the discrete GFF on a 150×150 hexagonal array with zero boundary. Interior white hexagons have height greater than zero; interior black hexagons have height less than zero; boundary hexagons (of height zero) are black. Hexagons that are not incident to a zero-height interface that reaches the boundary are grey.

Some (essentially well-known) geometric consequences of this kind of convergence are proved in §4.7.

A particularly interesting case of Theorem 1.3 is the case $a=b=0$, corresponding to the DGFF with zero boundary values. In this case, when h is interpolated linearly to triangles, its zero level set will almost surely include a finite number of piecewise linear arcs in D whose endpoints on ∂D are vertices of TG. A dual representation of this set of arcs is shown in Figure 1.4. For any fixed choice of endpoints on the boundary, the interface connecting those endpoints will converge to $\mathrm{SLE}(4;-1,-1)$. The limit of the complete set of arcs in Figure 1.4 is in some sense a coupling of $\mathrm{SLE}(4;-1,-1)$ processes, one for each pair of boundary points.

Finally, we discuss the generalizations: replace TG with an arbitrary weighted doubly periodic planar lattice—i.e., a connected planar graph $\mathfrak{G}\subset\mathbb{R}^2$ invariant under two linearly independent translations, T_1 and T_2, such that every compact subset of $\mathbb{R}^2$ meets only finitely many vertices and edges, together with a map w from the edges of $\mathfrak{G}$ to the positive reals, which is invariant under T_1 and T_2.

A $\mathfrak{G}$-*domain* $D\subset\mathbb{R}^2$ is a domain whose boundary is a simple closed curve comprised of edges and vertices in $\mathfrak{G}$. Let $V=V_{\overline{D}}$ be the set of $\mathfrak{G}$-vertices in the closure of D, let $G=G_D=(V,E)$ be the induced subgraph of $\mathfrak{G}$ with vertex set $V_{\overline{D}}$, and write $V_\partial=\partial D\cap V_{\overline{D}}$. Given a boundary value function $h_\partial\colon V_\partial\to\mathbb{R}$, the edge weighted DGFF on G has a density

with respect to the Lebesgue measure on $\mathbb{R}^{V\setminus V_\partial}$ which is proportional to

$$\exp\Bigg(\sum_{\{u,v\}\in E} -\frac{1}{2}w(\{u,v\})(h(v)-h(u))^2\Bigg).$$

If every face of $\mathfrak{G}$ has three edges, then every vertex in the dual graph is an endpoint of exactly three edges, and the boundary between positive and negative faces can be defined as a simple path in this dual lattice, similar to the one shown in Figure 1.1 (a). If not every face of $\mathfrak{G}$ has three edges, then we may "triangulate" $\mathfrak{G}$ by adding additional edges to $\mathfrak{G}$, while maintaining the invariance under T_1 and T_2, to make this the case (and set w to zero on these edges so that their presence does not affect the law of the DGFF).

We define the weighted random walk on $\mathfrak{G}$ to be the Markov chain with transition probability $w(\{u,v\})/\sum_{v'} w(\{u,v'\})$ from u to v, where we take $w(\{u,v\})=0$ unless u and v are neighbors in $\mathfrak{G}$. It is well known and easy to prove that such a walk on the rescaled lattice $\varepsilon\mathfrak{G}$ converges to a linear transformation of time-scaled 2-dimensional Brownian motion when ε tends to zero (but since we could not find a reference, we very briefly explain this in §5). It is convenient to replace the embedding of $\mathfrak{G}$ into $\mathbb{R}^2$ described above with a linear transformation of that embedding that causes this limit to be standard Brownian motion.

THEOREM 1.4. *Both Theorems* 1.2 *and* 1.3 *continue to hold if* TG *is replaced by a general weighted doubly periodic planar lattice* $\mathfrak{G}$, *as described above, provided that* $\mathfrak{G}$ *is embedded in* $\mathbb{R}^2$ *in such a way that the weighted random walk converges to Brownian motion.*

If $\mathfrak{G}$ is the grid $\mathbb{Z}^2$, then one natural way to triangulate $\mathfrak{G}$ is to add all the edges of the form $\{(x,y),(x+1,y+1)\}$. Another would be to add the edges $\{(x,y),(x+1,y-1)\}$. The above theorem implies, perhaps surprisingly, that the limiting law of the zero-height interface is the same in either case, with no need for a linear change of coordinates.

1.6. Outline

In §2 we introduce the basic notation and assumptions that are necessary for the height gap results proved in §3. In §§3.1–3.4 we develop bounds and estimates related to the geometry of zero-height interfaces. The random walk S comes into the picture in §3.5, where we develop results about the near-independence of the triple (h,γ,S) on microscopic and macroscopic scales. In §3.6 we apply these results to prove uniqueness of the limiting measure, namely Theorem 3.21, and in §3.7 we apply this to prove our main height gap result, Theorem 3.28.

Based on Theorem 3.28, the convergence in the case where the boundary values are $\pm\lambda$ is not too hard, using the method from [LSW4]. However, to prove the convergence to SLE$(4; a/\lambda-1, b/\lambda-1)$ in §4, we need to contend with a few other issues which stem from the fact that the driving parameter of SLE$(4; a/\lambda-1, b/\lambda-1)$ is the solution to an SDE with a drift term that blows up to infinity on a fractal set of times. To overcome these difficulties, we change coordinates to a coordinate system in which the drift terms stay bounded. In §4.4 we define and study *approximate diffusions*. These are random processes that are not necessarily Markov, but satisfy an approximate discrete version of an SDE. The Loewner driving term of the DGFF interface (before going to the scaling limit) is the approximate diffusion we are interested in. The main point is that an approximate diffusion of an SDE is shown to be close to the corresponding true diffusion satisfying the same SDE, under appropriate regularity conditions. This is how the convergence of the driving term of the DGFF interface to the driving term of the corresponding SLE$(\varkappa; \varrho_1, \varrho_2)$ is established. In §4.7 and §4.8 more geometric convergence results are deduced from the convergence of the Loewner driving term of the interface.

Finally, the rather brief §5 then describes the (very minor) modifications required for the generalization to other lattices, Theorem 1.4.

1.7. Sequel

This paper is actually the first of two papers the current authors are writing about this subject. In the second paper we will make sense of the "contour lines" of the continuum GFF. An instance h of the continuum GFF is a random distribution, not a random function; however, given an instance of the GFF on a domain D, we can project h onto the space of functions which are piecewise linear on a triangulation of D to yield an instance of the DGFF which is, in some sense, a piecewise linear approximation to h. We can then define the level lines of the GFF to be the limits of the level lines of its piecewise linear approximations (after proving that these limits exist). We will also characterize these random paths directly—without reference to discrete approximations—by showing that they are the unique path-valued functions of h which satisfy a simple Markov property. Similar techniques allow us to describe the contour lines of h that form loops (instead of starting and ending at points on the boundary of D).

The determination of the value of λ for a given lattice is not too hard, but fits better with the general spirit of our next paper on the subject, in which we will prove, in particular, that $\lambda_{\mathrm{TG}}=3^{-1/4}\sqrt{\frac{1}{8}\pi}$. If the DGFF is scaled so that its fine mesh limit is the ordinary GFF, we have $\lambda=\sqrt{\frac{1}{8}\pi}$.

Acknowledgments. We wish to thank John Cardy, Julien Dubédat, Bertrand Duplantier, Richard Kenyon, Jané Kondev, and David Wilson for inspiring and useful conversations, and thank Alan Hammond and an anonymous referee for numerous remarks on an earlier version of this paper.

2. Preliminaries

2.1. A few general properties of the DGFF

In this subsection, we recall a few well-known properties of the DGFF that are valid on any finite graph. Let (V,E) be a finite graph, and let $V_\partial \subset V$ be a non-empty set of vertices. Let h denote a sample from the DGFF with boundary values given by some function $h_\partial : V_\partial \to \mathbb{R}$.

When f is a function on V, the discrete gradient ∇f is the function on the set of ordered pairs (v,u) such that $\{v,u\}\in E$ defined by $\nabla f((u,v))=f(v)-f(u)$. When defining the norm of the gradient we sum over undirected edges, i.e., we write

$$\|\nabla f(v)\|^2 = \sum_{\{u,v\}\in E} (f(v)-f(u))^2. \tag{2.1}$$

Thus, the probability density of h is proportional to $e^{-\|\nabla h(v)\|^2/2}$. Therefore, when $h_\partial = 0$ on V_∂, h is a standard Gaussian with respect to the norm $\|\nabla h\|$ on Ω. The (discrete) Dirichlet inner product that defines this norm can be written

$$(f,g)_\nabla = \sum_{\{u,v\}\in E} (f(v)-f(u))(g(v)-g(u)). \tag{2.2}$$

Now write $\Delta f(v)=\sum_{u\sim v}(f(u)-f(v))$, where the sum is over all neighbors u of v. By expanding and rearranging the summands in (2.2), we find

$$(f,g)_\nabla = -(\Delta f, g). \tag{2.3}$$

Let $V_0 \subset V_D$. We claim that the vector space of functions $f: V\to\mathbb{R}$ that are zero on $V\setminus V_0$, and the vector space of functions $f: V\to\mathbb{R}$ that are discrete-harmonic on V_0 (i.e., $\Delta f = 0$ on V_0) are orthogonal to each other with respect to the inner product $(\cdot,\cdot)_\nabla$, and together they span $\mathbb{R}^V$. This basic observation will be used frequently below. Indeed, that they are orthogonal follows immediately from (2.3), and a dimension count now shows that the two spaces together span $\mathbb{R}^V$.

The following consequence of this orthogonality property will be used below. Let $V_0 \subset V$ satisfy $V_0 \supset V_\partial$ and let h_0 denote the function that is discrete-harmonic in $V\setminus V_0$

and equal to h in V_0. Then $h-h_0$ and h_0 are independent random variables, because $h=h_0+(h-h_0)$ is the corresponding orthogonal decomposition of h. It also follows that

$$h-h_0 \text{ is the DGFF in } V\setminus V_0 \text{ with zero boundary values on } V_0. \tag{2.4}$$

Observe that the Markov property and the effect of boundary conditions that were mentioned in §1.2 are immediate consequences of (2.4). An additional useful consequence is that the law of h_0 is proportional to $e^{-(h_0,h_0)_\nabla^2/2}$ times the Lebesgue measure on $\mathbb{R}^{V_0}$.

We now derive a useful well-known expression for the expectation of $h(v)h(u)$:

$$\mathbf{E}[h(v)h(u)]-\mathbf{E}[h(v)]\mathbf{E}[h(u)]=\frac{G(u,v)}{\deg(v)}, \tag{2.5}$$

where $G(u,v)$ is the expected number of visits to v by a simple random walk started at u before it hits V_∂ and $\deg(v)$ is the degree of v; that is, the number of edges incident with it. (The function G is known as the Green function.) As we have noted in the introduction, h is the sum of the discrete-harmonic extension of h_∂ and a DGFF with zero boundary values. It therefore suffices to prove (2.5) in the case where $h_\partial=0$ on V_∂. In this case, $\mathbf{E}[h(v)]=0=\mathbf{E}[h(u)]$. Setting $G_v(u)=G(u,v)$, we observe (or recall) that $\Delta G_v(u)=-\deg(v)1_v(u)$. Thus,

$$h(v)=(h,1_v)=-\frac{(h,\Delta G_v)}{\deg(v)}\overset{(2.3)}{=}\frac{(h,G_v)_\nabla}{\deg(v)}.$$

If X is a standard Gaussian in $\mathbb{R}^n$, and $x,y\in\mathbb{R}^n$, then $\mathbf{E}[(X\cdot x)(X\cdot y)]=x\cdot y$. Consequently, when $h_\partial=0$, we have

$$\begin{aligned}\mathbf{E}[h(v)h(u)]&=\frac{\mathbf{E}[(h,G_v)_\nabla(h,G_u)_\nabla]}{\deg(v)\deg(u)}=\frac{(G_v,G_u)_\nabla}{\deg(v)\deg(u)}\\&=\frac{-(G_v,\Delta G_u)}{\deg(v)\deg(u)}=\frac{(G_v,1_u)}{\deg(v)}=\frac{G(u,v)}{\deg(v)}.\end{aligned}$$

This proves (2.5).

2.2. Some assumptions and notation

We will make frequent use of the following notation and assumptions:

(h) A bounded domain (non-empty, open, connected set) $D\subset\mathbb{R}^2$ whose boundary ∂D is a subgraph of TG is fixed. The set of vertices of TG in D is denoted by V_D and V_∂ denotes the set of vertices in ∂D. A constant $\bar\Lambda>0$ is fixed, as well as a function $h_\partial: V_\partial\to\mathbb{R}$ satisfying $\|h_\partial\|_\infty\leqslant\bar\Lambda$. The DGFF on D with boundary values given by h_∂ is denoted by h. Also set $V=V_{\overline{D}}=V_D\cup V_\partial$.

We denote by TG^* the hexagonal grid which is dual to the triangular lattice TG—so that each hexagonal face of TG^* is centered at a vertex of TG. Generally, a TG^*-hexagon will mean a closed hexagonal face of TG^*. Denote by $\mathfrak{B}_R$ the union of all TG^*-hexagons that intersect the ball $B(0,R)$.

Sometimes, in addition to (h) we will need to assume:

(D) The domain D is simply connected, and (to avoid minor but annoying trivialities) ∂D is a simple closed curve. We fix two distinct midpoints of TG-edges x_∂ and y_∂ on ∂D. Let the counterclockwise (respectively, clockwise) arc of ∂D from x_∂ to y_∂ be denoted by ∂_+ (respectively, ∂_-).

If $H \subset D$ is a TG^*-hexagon, we write $h(H)$ as a shorthand for the value of h on the center of H (which is a vertex of TG). Assuming (h) and (D), let $\mathfrak{D}^+$ denote the union of all TG^*-hexagons contained in D where h is positive together with the intersection of $\overline{D}$ with TG^*-hexagons centered at vertices in ∂_+. Let $\mathfrak{D}^-$ be the closure of $\overline{D}\backslash\mathfrak{D}^+$ (which almost surely consists of TG^*-hexagons in D where $h<0$ and the intersection of $\overline{D}$ with TG^*-hexagons whose center is in ∂_-). Then $\partial\mathfrak{D}^- \cap \partial\mathfrak{D}^+$ necessarily consists of the interface we previously called γ, and a collection of disjoint simple closed paths. We use the term *interface* (or *zero-height interface*) to describe a simple (or simple closed) path in $\partial\mathfrak{D}^- \cap \partial\mathfrak{D}^+$ oriented so that $\mathfrak{D}^+$ is on its right (that is, oriented clockwise around $\mathfrak{D}^+$).

Throughout, the notation $O(s)$ represents any quantity f such that $|f| \leqslant Cs$ for some absolute constant C. We use the notation $O_{\bar{\Lambda}}(s)$ if the constant also depends on $\bar{\Lambda}$. When introducing a constant c, we often write $c=c(a,b)$ as shorthand to indicate that c may depend on a and b.

2.3. Simple random walk background

We need to recall a very useful property of the discrete-harmonic measure of simple random walk.

LEMMA 2.1. (Hit near) *Let v be a vertex of the grid* TG, *and let H be a connected subgraph of* TG. *Set $d=\operatorname{diam} H$. The probability that a simple random walk on* TG *started from v exits the ball $B(v,d)$ before hitting H is at most $c(\operatorname{dist}(v,H)/d)^{\zeta_1}$, where c and $\zeta_1 \in (0,1)$ are absolute constants.*

Likewise, the same bound applies to the probability that a simple random walk started at some vertex outside $B(v,d)$ will hit $B(v,\operatorname{dist}(v,H))$ before H.

In fact, we may take $\zeta_1 = \frac{1}{2}$. The continuous version of this statement is known as the Beurling projection theorem (the extremal case is when H is a line segment). The above statement can probably be deduced from the discrete Beurling theorem as given

in [L1, Theorem 2.5.2], though the setting there is slightly different. In any case, since we do not require any particular value for ζ_1, the lemma is rather easily proved directly (see [Sch, Lemma 2.1]).

We will also use the (well-known) discrete Harnack principle, in the following form.

LEMMA 2.2. (Harnack principle) *Let D, V, V_∂ and V_D be as in* (h), *let $v,u\in V_D$, and let $f\colon V\to\mathbb{R}$ be discrete-harmonic in V_D.*

(1) *If v and u are neighbors, then*

$$|f(v)-f(u)|\leqslant O(1)\frac{\|f\|_\infty}{\operatorname{dist}(v,\partial D)}.$$

(2) *If we assume that $f\geqslant 0$ and that there is a path of length ℓ from v to u whose minimal distance to ∂D is ϱ, then*

$$f(u)\geqslant f(v)\exp\biggl(-\frac{O(\ell+\varrho)}{\varrho}\biggr).$$

The proof of statement (1) can be obtained by noting that $f(v)$ is the expected value of f evaluated at the first hitting point on V_∂ of a random walk started at v, and observing that a random walk started at v and a random walk started at u may be coupled so that they meet before reaching a distance of $r:=\operatorname{dist}(v,\partial D)$ with probability $1-O(1/r)$ and walk together after they meet. (See also the more general [LSW4, Lemma 6.2], for example.)

To prove statement (2), let $W:=\{w\in V\colon f(w)\geqslant f(v)\}$, and observe that the maximum principle implies that W contains a path from v to ∂D. If we assume that $\ell\leqslant\frac{1}{2}\varrho$, then a random walk started at u has probability bounded away from zero to hit W before ∂D, which by the optional sampling theorem implies (2) in this case. The case $\ell>\frac{1}{2}\varrho$ now follows by induction on $\lceil 2\ell/\varrho\rceil$.

We also need the following well-known estimate on the Green function $G(u,v)$.

LEMMA 2.3. *Let D, V_D and V be as in* (h), *let $u,v\in V_D$, and let $\varepsilon>0$. Set $r:=\operatorname{dist}(u,\partial D)$, and suppose that within distance r/ε from u there is a connected component of ∂D of diameter at least εr. If $|u-v|\leqslant\frac{2}{3}r$, say, then $G_D(v,u)$ (the expected number of visits to u by a simple random walk starting at v before hitting ∂D) satisfies*

$$G_D(v,u)=\exp(O_\varepsilon(1))\log\frac{r+1}{|u-v|+1}.$$

The probability $H_D(v,u)$ that a random walk started from v hits u before ∂D can be expressed as $G_D(v,u)/G_D(u,u)$ and hence by the lemma

$$H_D(v,u)=\exp(O_\varepsilon(1))\biggl(1-\frac{\log(|u-v|+1)}{\log(r+1)}\biggr). \tag{2.6}$$

Let p_R denote the probability that a simple random walk started at 0 does not return to 0 before exiting the ball $B(0,R)$. We now show that the lemma follows from the well-known estimate

$$p_R = \frac{\exp(O(1))}{\log(R+2)}. \tag{2.7}$$

In the setting of the lemma, consider some vertex $w \in V_D$ such that $|w-u|>r$. It is easy to see that with probability bounded away from zero (by a function of ε), a simple random walk started from w will hit ∂D before u. Hence, $q := \min\{1-H_D(w,u) : w \in V_D \setminus B(u,r)\}$ is bounded away from 0 by a positive function of ε. Clearly,

$$\frac{1}{qp_r} \geqslant G_D(u,u) \geqslant \frac{1}{p_r}.$$

This proves the case $u=v$ of the lemma from (2.7). Now start a simple random walk at a vertex $v \neq u$ satisfying $|v-u| \leqslant \frac{2}{3}r$, and let the walk stop when it hits ∂D. It is easy to see that the probability that the walk visits v after exiting the ball $B\big(v, \frac{1}{4}|v-u|\big)$, say, is within a bounded factor from $H_D(v,u)$. Therefore, the expected number of visits to v after exiting $B\big(v, \frac{1}{4}|v-u|\big)$ is within a bounded factor of $G_D(v,u)$. But the former is the same as $G_D(v,v) - G_{B(v,|v-u|/4)}(v,v)$. The lemma follows.

We have not found a reference proving (2.7) in a way that generalizes to the setting of Theorem 1.4, though the result is well known. In fact, it easily follows from Rayleigh's method, as explained in [DoSn, §2.2]. In that book, the goal is to show that $p_R \to 0$, as $R \to \infty$, for lattices in the plane but not in $\mathbb{R}^3$. However, the method easily yields the quantitative bounds (2.7).

3. The height gap in the discrete setting

3.1. A priori estimates

This subsection contains some technical (and uninspiring) estimates that are necessary to carry out the technical (but hopefully interesting) coupling argument of the later parts of the section.

Suppose that β is some path in the hexagonal grid TG^*, which has a positive probability to be a subset of a zero-height interface h. We will need to understand rather well the behavior of h conditioned on β being a subset of a contour line. This conditioning amounts to conditioning h to be positive on vertices adjacent to β on one side and negative on vertices adjacent to β on the other side. (Here and in the following, a vertex v of TG is *adjacent* to β if β intersects the interior of one of the six boundary edges of the TG^*-hexagon centered at v.) Thus, the following lemma will be rather useful.

LEMMA 3.1. (Expectation bounds) *There is a finite $c=c(\bar{\Lambda})>0$ such that the following holds. Assume* (h). *Let V_+ and V_- be non-empty disjoint subsets of V_D, and set $U:=V_\partial\cup V_+\cup V_-$. Suppose that every vertex in V_+ has a neighbor in $V_-\cup V_\partial$ and every vertex in V_- has a neighbor in $V_+\cup V_\partial$. Let $\mathcal{K}$ be the event that $h>0$ on V_+ and $h<0$ on V_-. Then for every $v\in V_+\cup V_-$,*

$$\mathbf{E}[e^{|h(v)|}\,|\,\mathcal{K}]<c, \tag{3.1}$$

and

$$\frac{1}{c}<\mathbf{E}[|h(v)|\,|\,\mathcal{K}]<c. \tag{3.2}$$

Moreover,

$$\mathbf{E}[|h(v)|^{-1/2}\,|\,\mathcal{K}]<c. \tag{3.3}$$

Let $B\subset D$ be a disk whose radius r is smaller than its distance to U and assume that $B\cap V_D\neq\varnothing$. Then

$$\mathbf{E}[\max\{|\bar{h}(v)|:v\in V_D\cap B\}\,|\,\mathcal{K}]<c, \tag{3.4}$$

where $\bar{h}$ denotes the discrete-harmonic extension of the restriction of h to U (which is also the conditional expectation of h given its restriction to U). Moreover, if $\varepsilon>0$ and U has a connected component whose distance from B is R and whose diameter is at least εR, then

$$E\left[\left(\frac{1}{|V_D\cap B|}\sum_{v\in V_D\cap B}h(v)\right)^2\,\middle|\,\mathcal{K}\right]\leqslant c+c'\log\frac{R}{r}, \tag{3.5}$$

where $c'=c'(\varepsilon)$.

Proof. The proof of (3.1) is a maximum principle type argument. Suppose that v_1 maximizes $\mathbf{E}[e^{|h(v)|}|\mathcal{K}]$ among $v\in U$, and let $M:=\mathbf{E}[e^{|h(v_1)|}|\mathcal{K}]$. Clearly, $M<\infty$. Assume first that $v_1\in V_+$. Set $\widetilde{V}:=U\setminus\{v_1\}$, and for $v\in\widetilde{V}$ let p_v be the probability that the simple random walk starting at v_1 first hits $\widetilde{V}$ in v. We may write

$$h(v_1)=X+Y,$$

where

$$Y:=\sum_{v\in\widetilde{V}}p_v h(v) \tag{3.6}$$

and X is a centered Gaussian independent of $(h(v):v\in\widetilde{V})$. Moreover, by (2.5), $\mathbf{E}[X^2]$ is $\frac{1}{6}$ times the expected number of visits to v_1 by a simple random walk starting from v_1 until it first hits $\widetilde{V}$. Set $a:=\mathbf{E}[X^2]$. Since v_1 has a neighbor in $\widetilde{V}$, it follows that $a=O(1)$.

Also, clearly, a is bounded away from zero, since this random walk has at least one visit to v_1.

For every $y\in\mathbb{R}$, we have that

$$\mathbf{E}[e^{h(v_1)} \mid Y=y,\mathcal{K}] = \mathbf{E}[e^{X+y} \mid Y=y,\mathcal{K}] = e^y\mathbf{E}[e^X \mid X+y>0] = e^y \frac{\int_{-y}^{\infty} \exp(x-x^2/2a)\,dx}{\int_{-y}^{\infty} \exp(-x^2/2a)\,dx}.$$

If $y>0$, the right-hand side is bounded by $O(e^y)$. If $y\leqslant 0$, then both integrals on the right are comparable to their value when the upper integration limit is reduced to $-y+1$. But then the ratio between the corresponding integrands is bounded by e^{1-y}, which implies the same for the ratio of the integrals. Consequently, the right-hand side is $O(1)$ when $y\leqslant 0$. We take expectations conditioned on $\mathcal{K}$ and get

$$M = O(1)\mathbf{E}[e^Y \mid \mathcal{K}] + O(1). \tag{3.7}$$

Repeated use of Hölder's inequality shows that for non-negative random variables $x_1,\ldots,x_n$ and non-negative constants $p_1,\ldots,p_n$ such that $\sum_{j=1}^n p_j\leqslant 1$ we have

$$\mathbf{E}\biggl[\prod_{j=1}^n x_j^{p_j}\biggr] \leqslant \prod_{j=1}^n \mathbf{E}[x_j]^{p_j}.$$

Thus, (3.6) and (3.7) give

$$M \leqslant O(1)\prod_{v\in\widetilde{V}} \mathbf{E}[e^{h(v)} \mid \mathcal{K}]^{p_v} + O(1).$$

Clearly,

$$\mathbf{E}[e^{h(v)} \mid \mathcal{K}] \leqslant \begin{cases} 1, & \text{if } v\in V_-,\\ M, & \text{if } v\in V_+,\\ e^{\bar{\Lambda}}, & \text{if } v\in V_\partial. \end{cases}$$

Setting $p_+ := \sum_{v\in\widetilde{V}\cap V_+} p_v$, we therefore obtain

$$M \leqslant O(1)M^{p_+}e^{(1-p_+)\bar{\Lambda}} + O(1). \tag{3.8}$$

Since v_1 has a neighbor in $U\setminus V_+$, we have $p_+\leqslant\frac{5}{6}$ and $M\leqslant O(1)e^{\bar{\Lambda}}$ follows. A symmetric argument applies if $v_1\in V_-$. If $v_1\in V_\partial$, then obviously $M\leqslant e^{\bar{\Lambda}}$. Thus (3.1) holds with $c=O(e^{\bar{\Lambda}})$. The right-hand inequality in (3.2) is an immediate consequence of (3.1) (possibly with a different c).

We now prove (3.4). Given $v\in V_D$ and $u\in U$, let $H(v,u)$ denote the probability that a simple random walk started at v first hits U at u. Suppose that $v,v'\in B\cap V_D$ and $u\in U$. The discrete Harnack principle (Lemma 2.2) then gives $H(v',u)\leqslant H(v,u)O(1)$. Thus,

$$|\bar{h}(v')| = \biggl|\sum_{u\in U} H(v',u)h(u)\biggr| \leqslant \sum_{u\in U} O(1)H(v,u)|h(u)|.$$

The right-hand side therefore bounds $\max\{|\bar{h}(v')|:v'\in V_D\cap B\}$. Now, (3.4) follows from the right-hand inequality of (3.2) and $\sum_{u\in U} H(v,u)=1$.

We now prove (3.3). Consider some $v\in V_+$. We first claim that if w neighbors with v and $w\notin U$, then $\mathbf{E}[h(w)^2|\mathcal{K}]=O_{\bar{\Lambda}}(1)$. To see this, first note that $\bar{h}(w)$ (where as above $\bar{h}$ is the discrete-harmonic extension of the values of h on U) is a linear combination of the values $\bar{h}(v)$ for $v\in U$ (which are the same as the values of h there); since each of these values has mean and variance which are $O_{\bar{\Lambda}}(1)$ (by (3.1)), the mean and variance of $\bar{h}(w)$ are also $O_{\bar{\Lambda}}(1)$. By (2.5), conditioned on $\bar{h}$, the value $h(w)$ is Gaussian with variance $\frac{1}{6}$ times the expected number of times a random walk started at w visits w before hitting U, which is $O(1)$ because U contains a neighbor of w.

Let Z denote the average of h on the neighbors of v, and let $Z':=h(v)-Z$. The above implies that $\mathbf{E}[Z^2|\mathcal{K}]=O_{\bar{\Lambda}}(1)$. Since Z' is a centered Gaussian with variance $\frac{1}{6}$, for every $z\in\mathbb{R}$,

$$\mathbf{E}[h(v)^{-1/2}\,|\,\mathcal{K},Z=z]=\mathbf{E}[(Z'+z)^{-1/2}\,|\,Z'+z>0]=\frac{\int_0^\infty x^{-1/2}e^{-6(x-z)^2/2}\,dx}{\int_0^\infty e^{-6(x-z)^2/2}\,dx}.$$

If $z\geqslant -2$, then this is clearly bounded. Assume therefore that $z<-2$. It is easy to verify that the integrals in the numerator and denominator are comparable to the same integrals restricted to the range $x\in[0,|z|^{-1}]$. But in this range, the maximum value of

$$\exp\bigl(-\tfrac{6}{2}(x-z)^2\bigr)$$

is comparable to the minimum value of the same quantity. Consequently, when $z<-2$,

$$\begin{aligned}\mathbf{E}[h(v)^{-1/2}\,|\,\mathcal{K},Z=z]&=O(1)\frac{\int_0^{|z|^{-1}} x^{-1/2}e^{-6(x-z)^2/2}\,dx}{\int_0^{|z|^{-1}} e^{-6(x-z)^2/2}\,dx}\\&=O(1)\frac{\int_0^{|z|^{-1}} x^{-1/2}\,dx}{\int_0^{|z|^{-1}}\,dx}=O(1)|z|^{1/2},\end{aligned}$$

which gives

$$\mathbf{E}[h(v)^{-1/2}\,|\,\mathcal{K}]=O(1)+O(1)\mathbf{E}[|Z|^{1/2}\,|\,\mathcal{K}].$$

Since $\mathbf{E}[Z^2|\mathcal{K}]=O_{\bar{\Lambda}}(1)$, we certainly have $\mathbf{E}[|Z|^{1/2}|\mathcal{K}]=O_{\bar{\Lambda}}(1)$. This proves (3.3). Now the left-hand inequality in (3.2) is an immediate consequence.

We now prove (3.5). Consider two vertices $v,u\in B\cap V_D$. Assume that within distance R from B there is a connected component of U of diameter at least εR. Since

$$\mathbf{E}[h(v)\,|\,\mathcal{K},\bar{h}]=\bar{h}(v),$$

we have that

$$\mathbf{E}[h(v)h(u)\,|\,\mathcal{K}] = \mathbf{E}[(h(v)-\bar{h}(v))(h(u)-\bar{h}(u))\,|\,\mathcal{K}] + \mathbf{E}[\bar{h}(v)\bar{h}(u)\,|\,\mathcal{K}]. \tag{3.9}$$

Now, $\bar{h}(v)$ is just a weighted average of $h(w)$ with $w \in U$ (according to the discrete-harmonic measure from v). Consequently,

$$\mathbf{E}[\bar{h}(v)^2\,|\,\mathcal{K}] \leqslant \max_{w\in U} \mathbf{E}[h(w)^2\,|\,\mathcal{K}] \leqslant \max_{w\in U} 2\mathbf{E}[e^{|h(w)|}\,|\,\mathcal{K}] \leqslant 2c$$

and the Cauchy–Schwarz inequality implies that the last summand in (3.9) is also bounded by $2c$. Since $h-\bar{h}$ is the Gaussian free field with zero boundary values on U and is independent of the restriction of h to U, by (2.5) we have that

$$\mathbf{E}[(h(v)-\bar{h}(v))(h(u)-\bar{h}(u))\,|\,\mathcal{K}]$$

is $\frac{1}{6}$ times the expected number of visits to u by a random walker started at v which stops when it hits U. Thus

$$\sum_{u\in B\cap V_D} \mathbf{E}[(h(v)-\bar{h}(v))(h(u)-\bar{h}(u))\,|\,\mathcal{K}]$$

is $\frac{1}{6}$ times the expected number of steps that the walker spends in B, which is

$$O_\varepsilon(1)|B\cap V_D| \log \frac{R}{r},$$

by Lemma 2.3. The estimate (3.5) is now obtained by averaging (3.9) over all $v, u \in B\cap V_D$. □

The following lemma provides a variant of the right-hand bound in (3.2) in the case where instead of looking for a zero-height interface of h, we consider instead a zero-height interface of $h-g$, for some fixed function g.

LEMMA 3.2. (Further expectation bounds) *In the setting of Lemma* 3.1, *suppose that* $g\colon \overline{D}\to\mathbb{R}$ *is zero on* V_∂ *and Lipschitz in* $\overline{D}$. *Let* $\mathcal{K}_g$ *be the event that* $h>g$ *on* V_+ *and* $h<g$ *on* V_-. *Then for every* $v\in V_+\cup V_-$,

$$\mathbf{E}[|h(v)-g(v)|\,|\,\mathcal{K}_g] \leqslant c(1+\bar{\Lambda}) + c\frac{\|g\|_\infty}{\log(q+2)}, \tag{3.10}$$

where $c>0$ *is a universal constant and* $q=\|g\|_\infty/\|\nabla g\|_\infty$.

Proof. The proof is similar to the proof of Lemma 3.1. Suppose that v_1 maximizes $\mathbf{E}[h(v)-g(v)|\mathcal{K}_g]$ among $v\in V_+$, and let $M:=\mathbf{E}[h(v_1)-g(v_1)|\mathcal{K}_g]$. By symmetry, it is enough to get a bound on M. We define $\widetilde{V}$, X, Y and p_v as in the proof of Lemma 3.1.

For every $y\in\mathbb{R}$ we have

$$\mathbf{E}[h(v_1)-g(v_1)\,|\,Y=y,\mathcal{K}_g]=\mathbf{E}[X+y-g(v_1)\,|\,X+y-g(v_1)>0]\leqslant O(1)+(y-g(v_1))_+,$$

where $(x)_+:=\max\{0,x\}$. Consequently,

$$\begin{aligned}M&\leqslant O(1)+\mathbf{E}[(Y-g(v_1))_+\,|\,\mathcal{K}_g]\\&\leqslant O(1)+\sum_{v\in\widetilde{V}}p_v\mathbf{E}[(h(v)-g(v_1))_+\,|\,\mathcal{K}_g]\\&\leqslant O(1)+\sum_{v\in\widetilde{V}}p_v\mathbf{E}[(h(v)-g(v))_+\,|\,\mathcal{K}_g]+\sum_{v\in\widetilde{V}}p_v|g(v)-g(v_1)|\\&\leqslant O(1)+\bar{\Lambda}+M\sum_{v\in V_+}p_v+\sum_{v\in\widetilde{V}}p_v|g(v)-g(v_1)|.\end{aligned}$$

Consider a simple random walk started at v_1 and stopped when it first hits $\widetilde{V}$. Denote the vertex where it first hits $\widetilde{V}$ by w. (Then, $\mathbf{P}[v=w]=p_v$.) Set $r:=q/\log(q+2)$. Since v_1 has a neighbor in $\widetilde{V}$, we have $\mathbf{P}[|v-v_1|>r]\leqslant O(1)/\log(r+2)$, by standard random walk estimates. When $|v-v_1|\leqslant r$, we have $|g(v)-g(v_1)|\leqslant O(1)r\|\nabla g\|_\infty$. Therefore,

$$\sum_{v\in\widetilde{V}}p_v|g(v)-g(v_1)|\leqslant O(1)r\|\nabla g\|_\infty+O(1)\frac{\|g\|_\infty}{\log(r+2)}\leqslant O(1)\frac{\|g\|_\infty}{\log(q+2)}.$$

This gives

$$M\sum_{v\in\widetilde{V}\setminus V_+}p_v\leqslant O(1)+\bar{\Lambda}+O(1)\frac{\|g\|_\infty}{\log(q+2)}.$$

Because $\sum_{v\in\widetilde{V}\setminus V_+}p_v$ is bounded away from zero, the proof is now complete. □

Our next result establishes the continuity of the conditional distribution of h in the specified data. More precisely, the following proposition holds.

PROPOSITION 3.3. (Heights interface continuity) *For every $\varepsilon>0$ there is some $R=R(\varepsilon,\bar{\Lambda})>1/\varepsilon$ such that the following holds. Let D, V_D, V_∂, h_∂, V_+, V_-, U and $\mathcal{K}$ be as in Lemma 3.1, and let $\widehat{D}$, $\widehat{V}_D$, $\widehat{V}_\partial$, $\hat{h}_\partial$, $\widehat{V}_+$, $\widehat{V}_-$, $\widehat{U}$ and $\widehat{\mathcal{K}}$ be another such system, which is also assumed to satisfy $\|\hat{h}_\partial\|_\infty\leqslant\bar{\Lambda}$. Let $h_\mathcal{K}$ be a DGFF in D with boundary values given by h_∂ conditioned on $\mathcal{K}$, and let $h_{\widehat{\mathcal{K}}}$ be a DGFF in $\widehat{D}$ with boundary values given by $\hat{h}_\partial$ conditioned on $\widehat{\mathcal{K}}$. Suppose that within $\mathfrak{B}_R$, the two systems are the same; that is, $D\cap\mathfrak{B}_R=\widehat{D}\cap\mathfrak{B}_R$, $h_\partial|_{\mathfrak{B}_R}=\hat{h}_\partial|_{\mathfrak{B}_R}$, $V_+\cap\mathfrak{B}_R=\widehat{V}_+\cap\mathfrak{B}_R$ and $V_-\cap\mathfrak{B}_R=\widehat{V}_-\cap\mathfrak{B}_R$. Further suppose that $0\in U$. Then there is a coupling of $h_\mathcal{K}$ and $h_{\widehat{\mathcal{K}}}$ such that for every vertex $v\in\mathfrak{B}_{1/\varepsilon}$ we have $\mathbf{E}[|h_\mathcal{K}(v)-h_{\widehat{\mathcal{K}}}(v)|]<\varepsilon$.*

The following lemma will be needed in the proof.

LEMMA 3.4. *Let X be a 1-dimensional Gaussian of zero mean and unit variance. Let $x,\hat{x}\in\mathbb{R}$, let Z be a random variable whose distribution is the same as that of $X+x$ conditioned on $X+x>0$, and let $\widehat{Z}$ be a random variable whose distribution is the same as that of $X+\hat{x}$ conditioned on $X+\hat{x}>0$. Then there is a coupling of Z and $\widehat{Z}$ such that $|Z-\widehat{Z}|<|x-\hat{x}|$ almost surely if $x\neq\hat{x}$. Moreover, there is a continuous function $\delta(x,\hat{x})$ satisfying $\delta(x,\hat{x})<1$ such that $\mathbf{E}[|Z-\widehat{Z}|]\leqslant\delta(x,\hat{x})|x-\hat{x}|$ under this coupling.*

The coupling that we use is what is known as the *quantile coupling* of Z and $\widehat{Z}$.

Proof. Let $F(s)=\mathbf{P}[X<s]$, and let $G=F^{-1}$. Set $t:=F(-x)$ and $\hat{t}:=F(-\hat{x})$. Let p be a random variable uniformly distributed in $[0,1]$. Then

$$Z(t):=x+G(t+p(1-t))=G(t+p-tp)-G(t)$$

has the same distribution as Z. Therefore, $(Z(t),Z(\hat{t}))$ is a coupling of Z and $\widehat{Z}$. Consequently, to verify the first claim it is sufficient to show that $|\partial_t Z(t)|<\partial_t G(t)$. In fact, we will prove the stronger statement, $-\partial_t G(t)<\partial_t Z(t)<0$ for all $p\in(0,1)$, which is equivalent to

$$0<\partial_t G(t+p-tp)<\partial_t G(t).$$

The left-hand inequality is immediate, because $G'>0$ on $(0,1)$. The right-hand inequality translates to $(1-p)G'(t+p-tp)<G'(t)$, which we rewrite as

$$(1-(t+p-tp))G'(t+p-tp)<(1-t)G'(t).$$

This is equivalent to $(1-t_p)/F'(G(t_p))<(1-t)/F'(G(t))$, where $t_p:=t+p-tp>t$. Now, note that

$$\frac{1-t}{F'(G(t))}=\frac{\int_{-x}^{\infty}\exp(-s^2/2)\,ds}{\exp(-x^2/2)}=\int_{-x}^{\infty}e^{(x^2-s^2)/2}\,ds=\int_0^{\infty}e^{xs-s^2/2}\,ds$$

is strictly decreasing in t, because x is strictly decreasing in t. This proves the first claim. The second claim follows with

$$\delta(x,\hat{x}):=\int_0^1\frac{Z(t)-Z(\hat{t})}{x-\hat{x}}\,dp\quad\text{ when } x\neq\hat{x}$$

and

$$\delta(x,x):=\int_0^1\frac{Z'(t)}{-G'(t)}\,dp.$$

□

Proof of Proposition 3.3. Fix a coupling of $h_{\mathcal{K}}$ and $h_{\widehat{\mathcal{K}}}$ that minimizes

$$\sum_{v\in V_D\cap\widehat{V}_D}\mathbf{E}[|h_{\mathcal{K}}(v)-h_{\widehat{\mathcal{K}}}(v)|].$$

Standard continuity and compactness arguments show that there is such a coupling. Set $f(v):=\mathbf{E}[|h_{\mathcal{K}}(v)-h_{\widehat{\mathcal{K}}}(v)|]$ for vertices $v\in(V_D\cup V_\partial)\cap(\widehat{V}_D\cup\widehat{V}_\partial)$.

First, we claim that f is discrete-subharmonic on vertices in $\mathfrak{B}_R\setminus U$. Indeed, fix a vertex $w\in\mathfrak{B}_R\setminus U$. The conditional distribution of $h_{\mathcal{K}}(w)$ given the value of $h_{\mathcal{K}}$ at every vertex but w is that of $x+AX$, where x is the average of $h_{\mathcal{K}}$ on the neighbors of w, X is a standard Gaussian, and A is the lattice-dependent constant $1/\sqrt{6}$ (since each vertex has six neighbors in TG). Similarly, $h_{\widehat{\mathcal{K}}}(w)=\hat{x}+A\widehat{X}$. By the choice of coupling, when we fix the values of $h_{\mathcal{K}}$ and $h_{\widehat{\mathcal{K}}}$ off of w, the corresponding conditioned coupling of $h_{\mathcal{K}}(w)$ and $h_{\widehat{\mathcal{K}}}(w)$ minimizes the conditioned expectation of $|h_{\mathcal{K}}(w)-h_{\widehat{\mathcal{K}}}(w)|$. But one such conditioned coupling is obtained by taking $X=\widehat{X}$. Thus, $f(w)\leqslant\mathbf{E}[|x-\hat{x}|]$, which implies that f is discrete-subharmonic at w, since

$$\sum_{u\sim w}|h_{\mathcal{K}}(u)-h_{\widehat{\mathcal{K}}}(u)|\geqslant\Big|\sum_{u\sim w}h_{\mathcal{K}}(u)-\sum_{u\sim w}h_{\widehat{\mathcal{K}}}(u)\Big|,\tag{3.11}$$

where the sums are over the neighbors of w.

Next, consider any vertex $v\in V_+\cap\mathfrak{B}_R$. As before, we may write $h_{\mathcal{K}}(v)=x+AX$, where x is the average of $h_{\mathcal{K}}$ on the neighbors of v and X is a random variable whose conditional law given x is that of a standard Gaussian conditional on $x+AX>0$. Lemma 3.4 applied with x/A instead of x and $\hat{x}/A$ instead of $\hat{x}$ implies that f is also discrete-subharmonic at v. We claim that there is a constant $b=b(\Lambda,\varepsilon)>0$ such that

$$\Delta f(v)\geqslant b,\quad\text{if } f(v)\geqslant\tfrac{1}{2}\varepsilon,\tag{3.12}$$

where Δ denotes the discrete Laplacian on TG. Indeed, the optimality of the coupling gives

$$\mathbf{E}[|h_{\mathcal{K}}(v)-h_{\widehat{\mathcal{K}}}(v)|\,|\,x,\hat{x}]\leqslant\delta\Big(\frac{x}{A},\frac{\hat{x}}{A}\Big)|x-\hat{x}|,$$

where $\delta(\cdot,\cdot)<1$ is as in the lemma. Thus,

$$f(v)\leqslant\mathbf{E}\Big[\delta\Big(\frac{x}{A},\frac{\hat{x}}{A}\Big)|x-\hat{x}|\Big].$$

By (3.11) with v in place of w, we have

$$\Delta f(v)\geqslant\mathbf{E}[|x-\hat{x}|]-f(v)\geqslant\mathbf{E}\Big[\Big(1-\delta\Big(\frac{x}{A},\frac{\hat{x}}{A}\Big)\Big)|x-\hat{x}|\Big].\tag{3.13}$$

For every neighbor u of v we have, by (3.5), that $\mathbf{E}[h_{\mathcal{K}}(u)^2]<O_{\bar{\Lambda}}(1)$. It easily follows that $\mathbf{E}[x^2]<O_{\bar{\Lambda}}(1)$. Hence, the Cauchy–Schwarz inequality implies that there is a constant $b_0(\varepsilon,\bar{\Lambda})>0$ such that $\mathbf{E}[|x|1_{\mathcal{A}}]<\frac{1}{8}\varepsilon$ for every event $\mathcal{A}$ satisfying $\mathbf{P}[\mathcal{A}]<b_0$. There is a constant $b_1=b_1(\varepsilon,\bar{\Lambda})$ such that $\mathbf{P}[|x|>b_1]<\frac{1}{2}b_0$. The same inequalities will hold with $\hat{x}$ and $h_{\widehat{\mathcal{K}}}$ in place of x and $h_{\mathcal{K}}$. Therefore,

$$\begin{aligned}\mathbf{E}[|x-\hat{x}|]&=\mathbf{E}[|x-\hat{x}|1_{|x|\vee|\hat{x}|\leqslant b_1}]+\mathbf{E}[|x-\hat{x}|1_{|x|\vee|\hat{x}|>b_1}]\\&\leqslant\mathbf{E}[|x-\hat{x}|1_{|x|\vee|\hat{x}|\leqslant b_1}]+\mathbf{E}[(|x|+|\hat{x}|)1_{|x|\vee|\hat{x}|>b_1}]\\&\leqslant\mathbf{E}[|x-\hat{x}|1_{|x|\vee|\hat{x}|\leqslant b_1}]+\tfrac{1}{4}\varepsilon.\end{aligned}$$

Thus, if we assume that $f(v)\geqslant\frac{1}{2}\varepsilon$, then also $\mathbf{E}[|x-\hat{x}|]\geqslant\frac{1}{2}\varepsilon$, and therefore

$$\mathbf{E}[|x-\hat{x}|1_{|x|\vee|\hat{x}|\leqslant b_1}]\geqslant\tfrac{1}{4}\varepsilon.$$

Therefore, (3.13) gives (3.12) with

$$b:=\frac{\varepsilon}{4}\min\Big\{1-\delta\Big(\frac{x}{A},\frac{\hat{x}}{A}\Big):x,\hat{x}\in[-b_1,b_1]\Big\}.$$

Clearly, f is also discrete-subharmonic on $V_-\cap\mathfrak{B}_R$ and (3.12) also holds for $v\in V_-\cap\mathfrak{B}_R$ and (trivially) for $v\in V_\partial\cap\mathfrak{B}_R$.

Next, we prove that for all vertices $w\in\mathfrak{B}_R$ we have

$$f(w)\leqslant O_{\bar{\Lambda}}(1)\sqrt{\log R}.\tag{3.14}$$

Fix such a w, and assume that $w\notin U$. We may decompose $h_{\mathcal{K}}(w)$ as a sum $h_{\mathcal{K}}(w)=y+Y$, where y is the value at w of the discrete-harmonic extension of the restriction of $h_{\mathcal{K}}$ to U, and Y is a centered Gaussian whose variance is $\frac{1}{6}$ times the expected number of visits to w by a simple random walk started at w that is stopped when it hits U. A simple random walk on TG started at w has probability at least a positive constant times $1/\log R$ to reach distance R from w before returning to w, and once it does reach this distance, it has probability bounded away from zero to hit 0 before returning to w. Since $0\in U$, it follows that $\mathbf{E}[Y^2]=O(\log R)$. Thus, $\mathbf{E}[|Y|]=O(\sqrt{\log R})$. As y is the average of the value of $h_{\mathcal{K}}$ on U with respect to the discrete-harmonic measure from w, it follows from (3.2) that $\mathbf{E}[|y|]=O_{\bar{\Lambda}}(1)$. Thus, we have $\mathbf{E}[|h_{\mathcal{K}}(w)|]\leqslant O_{\bar{\Lambda}}(\sqrt{\log R})$. This certainly also holds if $w\in U$, and a similar estimate holds for $h_{\widehat{\mathcal{K}}}(w)$. Now (3.14) follows, since $f(w)\leqslant\mathbf{E}[|h_{\mathcal{K}}(w)|]+\mathbf{E}[|h_{\widehat{\mathcal{K}}}(w)|]$.

We now show that the established properties of f imply that $f\leqslant\varepsilon$ on $\mathfrak{B}_{1/\varepsilon}$ if R is sufficiently large. Fix some vertex $w\in\mathfrak{B}_{1/\varepsilon}$, and let S_t be a simple random walk on TG

started at w. Let t_1 be the first time t such that $|S_t|>\frac{1}{2}R$ or $S_t=0$. Since f is discrete-subharmonic on $V_D\cap\mathfrak{B}_R$, we have that $t\mapsto f(S_{t\wedge t_1})$ is a submartingale. The optional sampling theorem implies that $f(w)\leqslant\mathbf{E}[f(S_{t_1})]$. By standard random walk estimates, $\mathbf{P}\big[|S_{t_1}|>\frac{1}{2}R\big]\leqslant O(1)|\log\varepsilon|/\log R$. (We assume, with no loss of generality, that $\varepsilon<\frac{1}{2}$, say.) Consequently,

$$f(w)\leqslant\mathbf{E}[f(S_{t_1})]\leqslant f(0)+O(1)\frac{|\log\varepsilon|}{\log R}\max\{f(u):u\in V_D\cap\mathfrak{B}_R\}\leqslant f(0)+O_{\bar\Lambda}(1)\frac{|\log\varepsilon|}{\sqrt{\log R}}.$$

This proves that $f(w)\leqslant\varepsilon$ if $f(0)\leqslant\frac{1}{2}\varepsilon$ and R is sufficiently large.

Now assume that $f(0)\geqslant\frac{1}{2}\varepsilon$. Let $\widetilde{S}_t$ be a simple random walk starting at 0. Let

$$t_*:=\min\big\{t:|\widetilde{S}_t|>\tfrac{1}{2}R\big\}$$

and let n_s be the number of $t\in\{0,...,s-1\}$ such that $\widetilde{S}_t=0$. By (3.12) and our assumption that $f(0)\geqslant\frac{1}{2}\varepsilon$, we have that $t\mapsto f(\widetilde{S}_{t\wedge t_*})-bn_{t\wedge t_*}$ is a submartingale. Thus,

$$\begin{aligned}0\leqslant f(0)\leqslant\mathbf{E}[f(\widetilde{S}_{t_*})]-b\mathbf{E}[n_{t_*}]&\leqslant\max\{f(u):u\in V_D\cap\mathfrak{B}_R\}-b\mathbf{E}[n_{t_*}]\\&\leqslant O_{\bar\Lambda}(1)\sqrt{\log R}-b\mathbf{E}[n_{t_*}].\end{aligned}$$

Now note that as $R\to\infty$, while ε is fixed, $\mathbf{E}[n_{t_*}]$ grows at least as fast as a positive constant times $\log R$, because the probability for $\widetilde{S}_t$ not to return to 0 after any specific visit to 0 is bounded by $O(1/\log R)$. Thus, the above rules out the possibility that $f(0)\geqslant\frac{1}{2}\varepsilon$ if R is sufficiently large. This completes the proof. □

As a corollary of the proposition, we now show that the correlation in the values of h at two vertices in U decays when the distance between them tends to infinity.

COROLLARY 3.5. (Correlation decay) *For every $\varepsilon>0$ there is some $R=R(\varepsilon,\bar\Lambda)$ such that the following holds. Let D, V_D, V_∂, h_∂, V_+, V_-, U and $\mathcal{K}$ be as in Lemma* 3.1, *and let $v_1,v_2\in U$ satisfy $|v_1-v_2|>R$. Then*

$$\big|\mathbf{E}[h(v_1)h(v_2)\,|\,\mathcal{K}]-\mathbf{E}[h(v_1)\,|\,\mathcal{K}]\mathbf{E}[h(v_2)\,|\,\mathcal{K}]\big|<\varepsilon.$$

Proof. Suppose, without loss of generality, that $v_2\in V_+$. Fix some $a>0$ and let $X:=1_{\{0<h(v_2)\leqslant a\}}$. We may apply Proposition 3.3 to our present setup and to the setup where the value of $h(v_2)$ is fixed at some constant $y\in(0,a]$ and $v_2\in\partial D$. Thus, the proposition would apply, provided that $\bar\Lambda$ is replaced by $\bar\Lambda\vee a$. Consequently, we find that there is an $R'=R'(\varepsilon,\bar\Lambda,a)$ such that if $|v_1-v_2|>R'$, then

$$\big|\mathbf{E}[h(v_1)\,|\,h(v_2),\mathcal{K}]-\mathbf{E}[h(v_1)\,|\,\mathcal{K}]\big|X\leqslant\frac{\varepsilon}{2a}.$$

Since $h(v_2)X \leqslant a$, $X^2 = X$ and $h(v_2)X$ is $h(v_2)$-measurable, this gives

$$\begin{aligned}\tfrac{1}{2}\varepsilon &\geqslant \bigl|\mathbf{E}[h(v_1) \mid h(v_2), \mathcal{K}]h(v_2)X - \mathbf{E}[h(v_1) \mid \mathcal{K}]h(v_2)X\bigr| \\ &= \bigl|\mathbf{E}[h(v_1)h(v_2)X \mid h(v_2), \mathcal{K}] - \mathbf{E}[h(v_1) \mid \mathcal{K}]h(v_2)X\bigr|.\end{aligned}$$

Taking expectations conditioned on $\mathcal{K}$ now gives

$$\bigl|\mathbf{E}[h(v_1)h(v_2)X \mid \mathcal{K}] - \mathbf{E}[h(v_1) \mid \mathcal{K}]\mathbf{E}[h(v_2)X \mid \mathcal{K}]\bigr| \leqslant \tfrac{1}{2}\varepsilon. \tag{3.15}$$

Since

$$(1-X)h(v_2)^2 \leqslant \frac{(1-X)|h(v_2)|^3}{a} \leqslant \frac{6e^{|h(v_2)|}}{a}$$

and $h(v_1)^2 \leqslant 2e^{|h(v_1)|}$, if c denotes the constant satisfying (3.1), then the Cauchy–Schwarz inequality gives

$$\mathbf{E}[h(v_1)h(v_2)(1-X) \mid \mathcal{K}]^2 \leqslant \mathbf{E}[h(v_1)^2 \mid \mathcal{K}]\mathbf{E}[h(v_2)^2(1-X) \mid \mathcal{K}] \leqslant 2c\frac{6c}{a}.$$

Similarly,

$$0 \leqslant \mathbf{E}[|h(v_2)|(1-X) \mid \mathcal{K}] \leqslant E\left[\frac{2e^{|h(v_2)|}}{a} \,\middle|\, \mathcal{K}\right] \leqslant \frac{2c}{a}.$$

Consequently, if a is chosen sufficiently large then

$$\bigl|\mathbf{E}[h(v_1)h(v_2)X \mid \mathcal{K}] - \mathbf{E}[h(v_1)h(v_2) \mid \mathcal{K}]\bigr| = \bigl|\mathbf{E}[h(v_1)h(v_2)(1-X) \mid \mathcal{K}]\bigr| < \tfrac{1}{4}\varepsilon$$

and

$$\bigl|\mathbf{E}[h(v_1) \mid \mathcal{K}]\mathbf{E}[h(v_2)X \mid \mathcal{K}] - \mathbf{E}[h(v_1) \mid \mathcal{K}]\mathbf{E}[h(v_2) \mid \mathcal{K}]\bigr| < \tfrac{1}{4}\varepsilon.$$

The corollary now follows from (3.15). □

Next, we provide a simple lemma which bounds the amount in which adding a function to h affects its distribution.

LEMMA 3.6. (DGFF distortion) *Assume* (h). *Let* $f: V_D \cup V_\partial \to \mathbb{R}$ *satisfy* $f = 0$ *on* V_∂. *Let* μ *be the law of* h, *and let* μ_f *be the law of* $\tilde{h} := h + f$. *Then, for every event* $\mathcal{A}$,

$$\mu_f[\mathcal{A}] \leqslant \exp\left(\frac{\|\nabla f\|^2}{2}\right)\mu[\mathcal{A}]^{1/2}.$$

Proof. Suppose that X is a standard Gaussian in $\mathbb{R}^n$, $y \in \mathbb{R}^n$ is some fixed vector, and $A \subset \mathbb{R}^n$ is measurable. Then

$$\begin{aligned}\mathbf{P}[X + y \in A] &= c_n \int_{\mathbb{R}^n} 1_A \exp\left(-\frac{\|x-y\|^2}{2}\right) dx \\ &= c_n \int_{\mathbb{R}^n} 1_A \exp\left(x \cdot y - \frac{\|y\|^2}{2}\right) \exp\left(-\frac{\|x\|^2}{2}\right) dx,\end{aligned}$$

where $c_n^{-1}=\int_{\mathbb{R}^n}\exp(-\|x\|^2/2)\,dx$ and the integrals are with respect to the Lebesgue measure in $\mathbb{R}^n$. (This is the Cameron–Martin formula.) We may think of the right-hand side as the inner product of 1_A and $\exp(x\cdot y-\|y\|^2/2)$ with respect to the Gaussian measure. Hence, the Cauchy–Schwarz inequality gives

$$\mathbf{P}[X+y\in A]\leqslant\mathbf{P}[X\in A]^{1/2}\mathbf{E}[\exp(2X\cdot y-\|y\|^2)]^{1/2}=\mathbf{P}[X\in A]^{1/2}\exp\biggl(\frac{\|y\|^2}{2}\biggr).$$

Let $\bar{h}$ denote the discrete-harmonic extension of h_∂. Then $h-\bar{h}$ is the DGFF with zero boundary values, and hence is a standard Gaussian on $\mathbb{R}^{V_D}$ with respect to the norm $g\mapsto\|\nabla g\|_2$. The lemma follows. □

3.2. Near independence

In this subsection we build on the infrastructure developed above to prove that under appropriate assumptions the shape of an interface inside a ball does not depend too strongly on the shape of an interface outside a slightly larger ball. More precisely, we have the following result.

PROPOSITION 3.7. (Near independence) *Let $C>1$ and let $R>10^3C$. Assume* (h) *and $\mathfrak{B}_{5R}\subset D$. Let $R_1,R_2,R_3\in[R,5R]$ satisfy*

$$R_1+\frac{R}{C}<R_2,\quad R_2+\frac{R}{C}<R_3\quad\text{and}\quad R_3+\frac{R}{C}<5R.$$

Let V_+^3 and V_-^3 be disjoint sets of vertices in $D\setminus\mathfrak{B}_{R_3}$ and let V_+^1 and V_-^1 be disjoint sets of vertices in $\mathfrak{B}_{R_1}$. Suppose that every vertex of V_+^1 neighbors with a vertex in V_-^1, every vertex of V_-^1 neighbors with a vertex in V_+^1, and similarly for V_-^3 and V_+^3. Also suppose that a random walk started at 0 has probability at least $1/C$ to hit $V_-^3\cup V_+^3$ before exiting $\mathfrak{B}_{5R}$. Let $\mathcal{K}_1$ be the event that $h>0$ on V_+^1 and that $h<0$ on V_-^1, and let $\mathcal{K}_3$ be the corresponding event for V_-^3 and V_+^3. Let $a(V_+^1,V_-^1)$ be the probability of $\mathcal{K}_1$ for the DGFF on $\mathfrak{B}_{R_2}$ with zero boundary values outside $\mathfrak{B}_{R_2}$. Then there is a constant $c=c(\bar{\Lambda},C)>0$ such that

$$\frac{1}{c}a(V_+^1,V_-^1)\leqslant\mathbf{P}[\mathcal{K}_1\mid\mathcal{K}_3]\leqslant ca(V_+^1,V_-^1).$$

Proof. For $j=0,\dots,8$ set $r_j=R_1+\frac{1}{8}j(R_2-R_1)$. Then $r_8=R_2$ and $r_{j+1}\geqslant r_j+\frac{1}{8}C^{-1}R$. We set $W^j:=(V_D\cup V_\partial)\cap\mathfrak{B}_{r_j}=V_D\cap\mathfrak{B}_{r_j}$, $W_j:=(V_D\cup V_\partial)\setminus W^j$ and $W_j^k:=W_j\cap W^k$. Let $\tilde{h}$ denote the discrete-harmonic extension of the restriction of h to $W^1\cup W_7$. We may identify $\tilde{h}$ with a point in $\mathbb{R}^{W^1\cup W_7}$; namely, its restriction to $W^1\cup W_7$. As we have noted

after (2.4), the probability density of $\tilde{h}$ with respect to the Lebesgue measure on $\mathbb{R}^{W^1\cup W_7}$ is proportional to $\exp(-\|\nabla\tilde{h}\|^2/2)$. Hence,

$$\mathbf{P}[\mathcal{K}_1 \,|\, \mathcal{K}_3] = \frac{\int_{\mathcal{K}_1\cap\mathcal{K}_3} \exp(-\|\nabla\tilde{h}\|^2/2)\,dx}{\int_{\mathcal{K}_3} \exp(-\|\nabla\tilde{h}\|^2/2)\,dx}, \tag{3.16}$$

where the integrals are with respect to the Lebesgue measure on $\mathbb{R}^{W^1\cup W_7}$. Let h_1 be the function that agrees with h on W^1, is discrete-harmonic on W_1^7, and is zero in W_7. Let h_3 be the function that agrees with h on W_7, is discrete-harmonic on W_1^7, and is zero in W^1. Clearly, $\tilde{h}=h_1+h_3$.

We claim that

$$\int_{\mathcal{K}_1\cap\mathcal{K}_3} \exp\biggl(-\frac{\|\nabla\tilde{h}\|^2}{2}\biggr)\,dx \asymp \int_{\mathcal{K}_1\cap\mathcal{K}_3} \exp\biggl(-\frac{\|\nabla h_1\|^2}{2}\biggr)\exp\biggl(-\frac{\|\nabla h_3\|^2}{2}\biggr)\,dx, \tag{3.17}$$

where $\asymp$ means equivalence up to multiplicative constants depending on $\bar{\Lambda}$ and C. Let $V_+:=V_+^1\cup V_+^3$ and $V_-:=V_-^1\cup V_-^3$. Fix some $v\in W_2^6$. Let $\bar{h}$ denote the discrete-harmonic extension of the restriction of h to $V_+\cup V_-$. Then $h(v)-\tilde{h}(v)$ and $\tilde{h}(v)-\bar{h}(v)$ are independent Gaussian random variables, and both are independent of $\bar{h}$ (by the orthogonality property noted in §2.1). By (2.5), the variance of $h(v)-\tilde{h}(v)$ is $\frac{1}{6}$ times the expected number of visits to v by a random walk started at v, which is stopped when it hits $W^1\cup W_7$, and the variance of $h(v)-\bar{h}(v)$ is $\frac{1}{6}$ times the expected number of visits to v by the same random walk stopped when it hits $V_+\cup V_-$. Consequently, the variance of $\tilde{h}(v)-\bar{h}(v)$ is $\frac{1}{6}$ times the expected number of visits to v after the first hit of $W^1\cup W_7$ and before the first hit of $V_+\cup V_-$. Note that the probability to hit v by a random walk started in $W^1\cup W_7$ before exiting $\mathfrak{B}_{5R}$ is $O(1/\log R)$ and conditioned on hitting v before exiting $\mathfrak{B}_{5R}$ the number of visits to v prior to exiting $\mathfrak{B}_{5R}$ is $O(\log R)$. Our assumption on the probability to hit $V_-^3\cup V_+^3$ therefore easily implies that $\mathbf{E}[(\tilde{h}(v)-\bar{h}(v))^2]=O_C(1)$ and hence $\mathbf{E}[|\tilde{h}(v)-\bar{h}(v)|]=O_C(1)$. Since $\mathcal{K}_1\cap\mathcal{K}_3$ is determined by $\bar{h}$, it is independent of $\tilde{h}(v)-\bar{h}(v)$ and, consequently, $\mathbf{E}[|\tilde{h}(v)-\bar{h}(v)|\,|\mathcal{K}_1,\mathcal{K}_3]=O_C(1)$. Now, by (3.4), we have $\mathbf{E}[|\bar{h}(v)|\,|\mathcal{K}_1,\mathcal{K}_3]=O_{\bar{\Lambda}}(1)$. Combining these estimates, we get $\mathbf{E}[|\tilde{h}(v)|\,|\mathcal{K}_1,\mathcal{K}_3]=O_{C,\bar{\Lambda}}(1)$.

We will now apply the argument used to prove (3.4) in order to establish

$$\mathbf{E}[\max\{|\tilde{h}(v)| : v\in W_3^5\} \,|\, \mathcal{K}_1,\mathcal{K}_3] = O_{C,\bar{\Lambda}}(1). \tag{3.18}$$

Indeed, let A denote the set of vertices in W_2^6 neighboring with some vertex outside W_2^6, and let $H(v,u)$ denote the probability that a simple random walk started at v first hits A in u. As in the proof of (3.4), $H(v,u)\leqslant O_C(1)H(v',u)$ for $v,v'\in W_3^5$. Now (3.18) follows as in the proof of (3.4).

Next, we want to show that (3.18) holds with h_1 and h_3 replacing $\tilde{h}$; that is,

$$\mathbf{E}[M \,|\, \mathcal{K}_1, \mathcal{K}_3] = O_{C,\tilde{\Lambda}}(1), \tag{3.19}$$

where $M := \max\{|h_j(v)| : v \in W_3^5, j = 1, 3\}$. Let v_1 be the vertex $v \in W_3^5$ where $|h_1(v)|$ is maximized, and let v_3 be the vertex $v \in W_3^5$ where $|h_3(v)|$ is maximized. Then $M = \max\{|h_1(v_1)|, |h_3(v_3)|\}$. Assume, for now, that $M = |h_3(v_3)|$. The maximum principle for discrete-harmonic functions implies that v_3 neighbors with a vertex outside $\mathfrak{B}_{r_5}$ and v_1 neighbors with a vertex in $\mathfrak{B}_{r_3}$. Let p be the probability that a simple random walk started at v_3 exits $\mathfrak{B}_{R_3}$ before hitting a vertex neighboring with a vertex in $\mathfrak{B}_{r_3}$. Then p is bounded away from 0 by a function of C. Since h_1 composed with a simple random walk is a martingale while the walk stays in W_1^7, we get $|h_1(v_3)| \leqslant (1-p)|h_1(v_1)| \leqslant (1-p)M$. As $\tilde{h} = h_1 + h_3$, we get $|\tilde{h}(v_3)| \geqslant |h_3(v_3)| - |h_1(v_3)| = M - |h_1(v_3)| \geqslant pM$. The case $M = |h_1(v_1)|$ is similarly treated. Using (3.18), we then get (3.19).

Next, we want to prove that

$$|\nabla h_1 \cdot \nabla h_3| = O(M^2). \tag{3.20}$$

Since h_1 is discrete-harmonic in W_1^7, if $v \in W_1^7$ we have

$$\sum_{u \sim v} (h_1(v) - h_1(u)) h_3(v) = 0,$$

where the sum is over the neighbors of v. This is also true for $v \in W^1$, since h_3 is zero there. Consequently,

$$\sum_{v \in W^4} \sum_{u \sim v} (h_1(v) - h_1(u)) h_3(v) = 0. \tag{3.21}$$

Similarly, we find that

$$\sum_{u \in W_4} \sum_{v \sim u} (h_3(u) - h_3(v)) h_1(u) = 0. \tag{3.22}$$

Set $\partial W^4 := \{(v, u) \in W^4 \times W_4 : u \sim v\}$. By considering the contribution of each edge $[v, u]$ to $\nabla h_1 \cdot \nabla h_3$, we compare the sum of the left-hand sides of (3.21) and (3.22) to $\nabla h_1 \cdot \nabla h_3$ and conclude that

$$\begin{aligned} \nabla h_1 \cdot \nabla h_3 &= \sum_{(v,u) \in \partial W^4} (h_3(v) h_1(u) - h_1(v) h_3(u)) \\ &= \sum_{(v,u) \in \partial W^4} ((h_1(u) - h_1(v)) h_3(u) + (h_3(v) - h_3(u)) h_1(u)). \end{aligned} \tag{3.23}$$

The number of summands is clearly $O(R)$. Note that for every $v \in W^4$ neighboring with a vertex $u \in W_4$, there is a disk of radius proportional to R/C such that all the vertices

in that disk are in W_3^5. Consequently, the discrete Harnack principle (Lemma 2.2) gives $|h_1(u)-h_1(v)|=O(M/R)$ and $|h_3(u)-h_3(v)|=O(M/R)$. Hence, (3.23) gives (3.20).

Now, (3.19) implies that the expectation of $|\nabla h_1\cdot\nabla h_3|^{1/2}$ conditioned on $\mathcal{K}_1\cap\mathcal{K}_3$ is bounded by a function of C and $\bar{\Lambda}$. In particular, there is a constant $c_1=c_1(\bar{\Lambda},C)>0$ such that

$$\mathbf{P}[|\nabla h_1\cdot\nabla h_3|<c_1\,|\,\mathcal{K}_1,\mathcal{K}_3]>\frac{1}{c_1}.$$

In terms of Lebesgue measure, this may be written as

$$\int_{\mathcal{K}_1\cap\mathcal{K}_3}\exp\biggl(-\frac{\|\nabla\tilde{h}\|^2}{2}\biggr)\,dx<c_1\int_{\mathcal{K}_1\cap\mathcal{K}_3}\exp\biggl(-\frac{\|\nabla\tilde{h}\|^2}{2}\biggr)1_{\{|\nabla h_1\cdot\nabla h_3|<c_1\}}\,dx.$$

Since $\tilde{h}=h_1+h_3$, this implies that

$$\int_{\mathcal{K}_1\cap\mathcal{K}_3}\exp\biggl(-\frac{\|\nabla\tilde{h}\|^2}{2}\biggr)\,dx<c_1e^{c_1}\int_{\mathcal{K}_1\cap\mathcal{K}_3}\exp\biggl(-\frac{\|\nabla h_1\|^2+\|\nabla h_3\|^2}{2}\biggr)\,dx,$$

which gives one side of (3.17).

The other direction is proved in essentially the same way. Under the probability measure weighted by $\exp(-\frac{1}{2}(\|\nabla h_1\|^2+\|\nabla h_3\|^2))$ (with respect to the Lebesgue measure on $\mathbb{R}^{W^1\cup W_7}$), h_1 restricted to W^1 has the law of the DGFF with zero boundary values on W_7 restricted to W^1. Similarly, with this weighting, h_3 restricted to W_7 has the law of the DGFF with zero boundary values on W^1 and with boundary values given by h_∂ on ∂D, restricted to W_7. Moreover, under this measure, h_1 and h_3 are clearly independent. The above arguments show that under this measure (3.19) holds (where $\mathcal{K}_1$ refers to h_1 while $\mathcal{K}_3$ refers to h_3). Since (3.20) is still valid, the opposite inequality in (3.17) is then easily established.

We may also apply (3.17) in the case where $V_+^1=V_-^1=\varnothing$, and hence $\mathcal{K}_1$ has full measure. Since

$$\begin{aligned}\int_{\mathcal{K}_1\cap\mathcal{K}_3}&\exp\biggl(-\frac{\|\nabla h_1\|^2}{2}\biggr)\exp\biggl(-\frac{\|\nabla h_3\|^2}{2}\biggr)\,dx\\&=\int_{\mathbb{R}^{W^1}}1_{\mathcal{K}_1}\exp\biggl(-\frac{\|\nabla h_1\|^2}{2}\biggr)\,dx\int_{\mathbb{R}^{W_7}}1_{\mathcal{K}_3}\exp\biggl(-\frac{\|\nabla h_3\|^2}{2}\biggr)\,dx,\end{aligned}$$

from (3.16) we get

$$\mathbf{P}[\mathcal{K}_1\,|\,\mathcal{K}_3]\asymp\frac{\int_{\mathcal{K}_1}\exp(-\|\nabla h_1\|^2/2)\,dx}{\int_{\mathbb{R}^{W_1}}\exp(-\|\nabla h_1\|^2/2)\,dx},$$

which completes the proof. □

3.3. Narrows and obstacles

The present subsection and the next will use the infrastructure developed in §3.1 to prove some bounds on the probabilities that contour lines cross certain regions in specified ways. This is roughly in the spirit of the Russo–Seymour–Welsh theorem for percolation, though the proofs are entirely different.

The following lemma is an estimate for having a crossing by TG^*-hexagons where h is negative between two arcs on the boundary of some subset of the domain, conditioned on some zero-height interface paths. The statement below is slightly complicated, because we need to keep the geometric assumptions quite general. In percolation, boundary values do not play a role, of course. But in our case we need the crossing estimate in the case where one boundary arc of the domain is conditioned to be an interface.

LEMMA 3.8. (Narrows) *For every $\varepsilon>0$ there is a $\delta=\delta(\bar{\Lambda},\varepsilon)>0$ such that the following crossing estimate holds. Assume* (h) *and* (D). *Let $\mathcal{K}$ be the event that a fixed collection $\{\gamma_1,...,\gamma_k\}$ of oriented paths in TG^* are contained in oriented zero-height interfaces of h, and suppose that $\mathbf{P}[\mathcal{K}]>0$. Let $\alpha\subset D\setminus(\gamma_1\cup...\cup\gamma_k)$ be a simple path that has both its endpoints on the right-hand side (positive side) of γ_1. Let A be the domain bounded by α and a subarc of γ_1, and assume that A does not meet the left side of γ_1 and that $\bar{A}\cap(\gamma_2\cup\gamma_3\cup...\cup\gamma_k\cup\partial D)=\varnothing$. Let α_1, α_2 and α' be three disjoint subarcs of α, where α_1 contains one endpoint of α, and α_2 contains the other endpoint of α. (See Figure* 3.1.) *Suppose that each point in α' is contained in a TG^*-hexagon whose center is outside A. Set $d_1:=\sup_{z\in\alpha'}\operatorname{dist}(z,\gamma_1)$. Let $\mathcal{C}$ be the event that there is a path crossing from α_1 to α_2 in A inside hexagons where h is negative. Let d^* be the infimum diameter of any path connecting α' to $\gamma_1\cup...\gamma_k\cup\partial D$ which does not contain a subpath connecting $\alpha\setminus(\alpha_1\cup\alpha_2)$ to γ_1 in A. If*

$$d_1+1\leqslant\delta\min\{d^*,\operatorname{dist}(\alpha_1,\alpha'),\operatorname{dist}(\alpha_2,\alpha'),\operatorname{diam}\alpha'\},\tag{3.24}$$

then

$$\mathbf{P}[\mathcal{C}\,|\,\mathcal{K}]<\varepsilon.$$

The idea of the proof is to observe the effect that such a crossing would have on certain averages of heights of vertices, and thereby conclude that it is unlikely. The challenge in the implementation of this strategy is to condition on a crossing in such a way that the expected heights are easy to estimate.

Proof. Set $N:=\lfloor\frac{1}{2}\operatorname{diam}\alpha'\rfloor$. Assume that (3.24) holds. Note that

$$\operatorname{diam}\gamma_1\geqslant\operatorname{diam}\alpha'-2d_1.$$

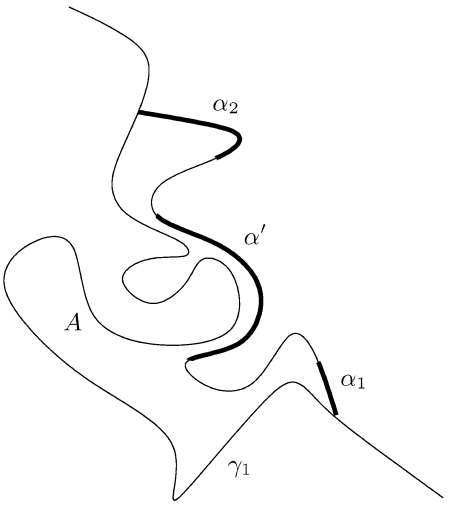


Figure 3.1. Setup in the narrows lemma.

Choose some point $z_0 \in \alpha'$. For $j=1,...,N$ let z_j be a point in α' at distance j from z_0. Now for each z_j we let s_j be some center of a hexagon that contains z_j satisfying $s_j \notin A$. Then $|s_j - s_{j'}| \geqslant |j-j'| - O(1)$. Let U be the union of V_∂ and the set of vertices adjacent to any one of the paths $\gamma_1, ..., \gamma_k$. (These are precisely the vertices v where h takes boundary values, or the sign of $h(v)$ is determined by $\mathcal{K}$.) Set

$$X := \frac{1}{N} \sum_{j=1}^{N} h(s_j),$$

$b := \mathbf{E}[X \,|\, \mathcal{K}]$ and fix some $\varepsilon' > 0$. We first claim that

$$\mathbf{E}[(X-b)^2 \,|\, \mathcal{K}] < \varepsilon' \tag{3.25}$$

if $\delta = \delta(\varepsilon', \bar{\Lambda}) > 0$ is sufficiently small. Let h_U denote the discrete-harmonic extension of the restriction of h to U and set $X_U := N^{-1} \sum_{j=1}^{N} h_U(s_j)$. Note that $\mathbf{E}[X - X_U \,|\, \mathcal{K}, h_U] = 0$, and hence $\mathbf{E}[X_U \,|\, \mathcal{K}] = b$. For each $u \in U$ and $j \in \{1, ..., N\}$ let $p(j,u)$ denote the probability that a simple random walk started from s_j first hits U at u. Also set $p(u) := N^{-1} \sum_{j=1}^{N} p(j,u)$. Then $X_U = \sum_{u \in U} p(u) h(u)$. Consequently,

$$\mathbf{E}[(X_U - b)^2 \,|\, \mathcal{K}] = \sum_{u,u' \in U} p(u)p(u')(\mathbf{E}[h(u)h(u') \,|\, \mathcal{K}] - \mathbf{E}[h(u) \,|\, \mathcal{K}]\mathbf{E}[h(u') \,|\, \mathcal{K}]).$$

Let $Z(u,u')$ denote the term in parentheses corresponding to the summand involving u and u'. Then $\mathbf{E}[(X_U - b)^2 | \mathcal{K}]$ is just the average of the conditioned covariances $Z(u,u')$

weighted by $p(u)p(u')$. We know from (3.1) that $\mathbf{E}[h(u)^2|\mathcal{K}]$ is bounded by a constant depending only on $\bar{\Lambda}$. It then follows by the Cauchy–Schwarz inequality that the same is true for $Z(u,u')$. Consequently, to prove that $\mathbf{E}[(X_U-b)^2|\mathcal{K}]$ is small, it suffices to show that when (u,u') is chosen with probability $p(u)p(u')$ it is very likely that $|Z(u,u')|$ is small. Suppose that we select j from $\{1,...,N\}$ uniformly at random and given j select $u\in U$ with probability $p(j,u)$. Independently, we also select (j',u') with the same distribution. It suffices to show that $|Z(u,u')|$ is likely to be small, and by Corollary 3.5 it suffices to show that the distance between u and u' is likely to be large. Since $|s_j-s_{j'}|=|j-j'|+O(1)$ is unlikely to be much smaller than $\operatorname{diam}\alpha'$, which is larger than $\delta^{-1}(d_1+1)$, it follows from Lemma 2.1 (hit near) that for any fixed R the probability that $|u-u'|<R$ tends to zero as $\delta\to 0$. Consequently,

$$\mathbf{E}[(X_U-b)^2\,|\,\mathcal{K}]<\tfrac{1}{2}\varepsilon',$$

provided that δ is sufficiently small.

Set $X_j:=h(s_j)-h_U(s_j)$. Recall from §2.1 that, given the restriction of h to U, the function $h-h_U$ is the DGFF on $V_D\setminus U$ with zero boundary values on U. Therefore, by (2.5), $\mathbf{E}[X_iX_j|\mathcal{K},h_U]=\frac{1}{6}G(s_i,s_j)$, where $G(v,u)$ is the expected number of visits to u by a random walker started at v and stopped when it hits U. From Lemmas 2.1 and 2.3,

$$G(s_j,s_{j'})\leqslant\begin{cases}O(1)\log(d_1/(|s_j-s_{j'}|\vee 1)), & \text{if } |s_j-s_{j'}|<\frac{1}{2}d_1,\\ O(1)(d_1/(|s_j-s_{j'}|\vee 1))^{\zeta_1}, & \text{if } |s_j-s_{j'}|\geqslant\frac{1}{2}d_1.\end{cases}$$

Since $|s_j-s_{j'}|\geqslant|j-j'|-O(1)$ and $\zeta_1\in(0,1)$, these estimates give

$$\mathbf{E}[(X-X_U)^2\,|\,\mathcal{K}]=\frac{1}{N^2}\sum_{j,j'=1}^{N}\frac{G(s_j,s_{j'})}{6}=O(1)\biggl(\frac{d_1}{N}\biggr)^{\zeta_1}.$$

Now (3.25) follows for sufficiently small $\delta=\delta(\varepsilon',\bar{\Lambda})>0$, since $X-X_U$ is independent of X_U and they are also independent given $\mathcal{K}$.

We now claim that $b\geqslant 0$ if δ is sufficiently small. Let c be the constant given by Lemma 3.1. If u is a fixed vertex adjacent to γ_1 on the right, then $\mathbf{E}[h(u)|\mathcal{K}]>1/c$, by (3.2). On the other hand, $\mathbf{E}[|h(u)||\mathcal{K}]<c$ for every $u\in U$. By (3.24) and Lemma 2.1, it follows that when δ is small with high probability a random walk starting at any s_j is likely to first hit U at a vertex adjacent to the right-hand side of γ_1. Thus, when δ is small, we have $b=\mathbf{E}[X_U|\mathcal{K}]>0$. Also, clearly, $b\leqslant c$.

Now set $a=b+1$. Let Q denote the union of the closed hexagons in TG^* for which $h(H)\in[0,a]$ and let Q' denote the union of the edges in TG^* that are on the common boundary of two hexagons H_1 and H_2 satisfying $h(H_1)<0$ and $a<h(H_2)$. Let Q_0 denote

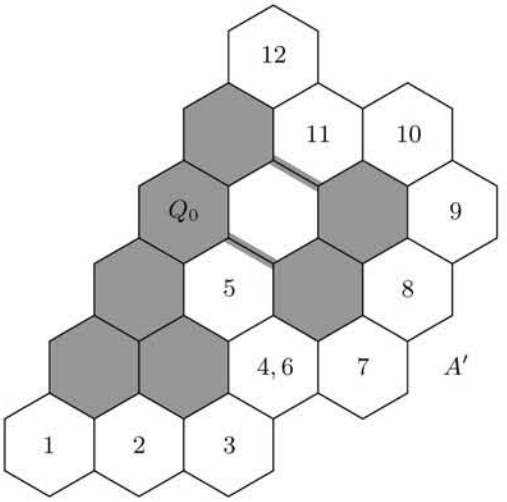


Figure 3.2. A portion of the sequence of hexagons adjacent to $\partial A' \setminus \partial A$.

the connected component of $(Q \cup Q') \cap \bar{A}$ that contains $\gamma_1 \cap \partial A$. Let $\mathcal{Q}$ denote the event $Q_0 \cap (\alpha \setminus (\alpha_1 \cup \alpha_2)) = \varnothing$. If $\mathcal{C}$ holds, then the corresponding crossing by hexagons where h is negative separates $\gamma_1 \cap \partial A$ from $\alpha \setminus (\alpha_1 \cup \alpha_2)$ in A, and hence $\mathcal{Q}$ holds as well. Thus, $\mathcal{C} \subset \mathcal{Q}$.

On the event $\mathcal{Q}$, let A' be the connected component of $\bar{A} \setminus Q_0$ that contains α', and let U_Q denote the set of centers of hexagons H such that $\overline{H \cap A'} \cap Q_0 \neq \varnothing$. Clearly, $h(v) < 0$ or $h(v) > a$ for each $v \in U_Q$. Since A' and Q_0 are connected, it is immediate to verify (using the Jordan planar curve theorem) that A' is simply connected and $\partial A' \setminus \partial A \subset Q_0$ is connected. The closed hexagons of TG^* with centers in U_Q form a sequence (possibly with repetitions) with each pair H, H' of consecutive hexagons along the sequence satisfying $H \cap H' \setminus Q_0 \neq \varnothing$ and $H \cap H' \cap Q_0 \neq \varnothing$. (See Figure 3.2.) If $v, u \in U_Q$ are centers of consecutive hexagons in this sequence, then it is impossible that $h(v) < 0$ and $h(u) > a$ (otherwise, the boundary between the hexagons would be in Q'). Thus, either $h(U_Q) \subset (a, \infty)$, or $h(U_Q) \subset (-\infty, 0)$. Let $\mathcal{Q}_+$ be the event that $\mathcal{Q}$ occurs and $\min h(U_Q) > a$, and let $\mathcal{Q}_-$ be the event that $\mathcal{Q}$ occurs and $\max h(U_Q) < 0$.

We now want to estimate $\mathbf{E}[X \,|\, \mathcal{K}, \mathcal{Q}_-]$ and $\mathbf{E}[X \,|\, \mathcal{K}, \mathcal{Q}_+]$. Let U' be the set of vertices that are either in hexagons adjacent to Q_0 or in U. Since $a = O_{\bar{\Lambda}}(1)$, it is clear that the proof of (3.2) gives $\mathbf{E}[|h(u)| \,|\, \mathcal{K}, U', \mathcal{Q}_\pm] = O_{\bar{\Lambda}}(1)$ for $u \in U'$. On the other hand, Lemma 2.1 shows that at least $1 - O(\delta^{\zeta_1})$ of the discrete-harmonic measure on U' starting from every s_j is in U_Q. If $\mathcal{Q}_-$ holds and $u \in U_Q$, then $\mathbf{E}[h(u) \,|\, \mathcal{K}, \mathcal{Q}_-, U_Q]$ is negative and bounded away from zero, by the corresponding analog of the left-hand side of (3.2). Thus, we find that $\mathbf{E}[X \,|\, \mathcal{K}, \mathcal{Q}_-]$ is negative and bounded away from zero when $\delta = \delta(\varepsilon', \bar{\Lambda}) > 0$ is small.

Since $\mathbf{E}[(X-b)^2|\mathcal{K}]<\varepsilon'$ and $b\geqslant 0$, we conclude that by choosing $\varepsilon'>0$ small it can be guaranteed that $\mathbf{P}[\mathcal{Q}_-|\mathcal{K}]<\frac{1}{2}\varepsilon$.

On $\mathcal{Q}_+$, we clearly have $h(u)>a\geqslant b+1$ on every $u\in U_Q$. Thus, as above, it follows that when δ is small $\mathbf{E}[X|\mathcal{K},\mathcal{Q}_+]>b+\frac{1}{2}$. This again implies that $\mathbf{P}[\mathcal{Q}_+|\mathcal{K}]$ can be made smaller than $\frac{1}{2}\varepsilon$. Since $\mathcal{C}\subset\mathcal{Q}_+\cup\mathcal{Q}_-$, this completes the proof. □

Next, we formulate an analogous lemma for crossings near the boundary of the domain.

LEMMA 3.9. (Domain boundary narrows) *There is a constant $\Lambda_0=\Lambda_0(\bar{\Lambda})>0$ such that for every $\varepsilon>0$ there is a $\delta=\delta(\bar{\Lambda},\varepsilon)>0$ such that the following crossing estimate holds. Assume* (h) *and* (D), *assume that ∂_+ is a simple path contained in ∂D and that $h_\partial\geqslant-\Lambda_0$ on $\partial_+\cap V_\partial$. Set $\partial_-:=\partial D\setminus\partial_+$. Let $\mathcal{K}$ be the event that a fixed collection $\{\gamma_1,\ldots,\gamma_k\}$ of oriented paths in* TG^* *are contained in oriented zero-height interfaces of h, and suppose that $\mathbf{P}[\mathcal{K}]>0$. Let $\alpha\subset D\setminus(\gamma_1\cup\ldots\cup\gamma_k)$ be a simple path that has both its endpoints on ∂_+. Let A be the domain bounded by α and a subarc of ∂_+, and assume that $\bar{A}\cap(\gamma_1\cup\gamma_2\cup\ldots\cup\gamma_k\cup\partial_-)=\varnothing$. Let $\alpha'\subset\alpha$ be a subarc. Suppose that each point in α' is contained in a hexagon whose center is outside A. Set $d_1:=\sup_{z\in\alpha'}\operatorname{dist}(z,\partial_+)$. Let α_1 be a subarc of α that contains one of the endpoints of α, and let α_2 be a subarc of α that contains the other endpoint of α. Let $\mathcal{C}$ be the event that there is a path crossing from α_1 to α_2 in A inside hexagons where h is negative. Let d^* be the infimum diameter of any path connecting α' to $\gamma_1\cup\ldots\gamma_k\cup\partial D$ which does not contain a subpath connecting $\alpha\setminus(\alpha_1\cup\alpha_2)$ to ∂_+ in A. If*

$$d_1+1\leqslant\delta\min\bigl\{d^*,\operatorname{dist}(\alpha_1,\alpha'),\operatorname{dist}(\alpha_2,\alpha'),\operatorname{diam}\alpha'\bigr\},\tag{3.26}$$

then

$$\mathbf{P}[\mathcal{C}\,|\,\mathcal{K}]<\varepsilon.$$

Proof. The proof is slightly simpler but essentially the same as that of Lemma 3.8 (narrows). We use the same notation as in that lemma, and only indicate the few differences in the proof. In the present setting $b=\mathbf{E}[X|\mathcal{K}]$ can be made larger than $-2\Lambda_0$ by taking $\delta>0$ small. Here, we define Q_0 as the connected component of $(Q\cup Q'\cup\partial_+)\cap\bar{A}$ that contains $\partial_+\cap\partial A$. Observe that $U_Q\cap\partial D=\varnothing$ on $\mathcal{Q}_-\cap\mathcal{K}$. It follows that $\mathbf{E}[X|\mathcal{K},\mathcal{Q}_-]$ is negative and bounded away from zero (by a function of $\bar{\Lambda}$) when $\delta>0$ is small. By taking $\Lambda_0>0$ sufficiently small, we can make sure that $\mathbf{E}[X|\mathcal{K},\mathcal{Q}_-]<-3\Lambda_0$. But since $\mathbf{E}[(X-b)^2|\mathcal{K}]$ is arbitrarily small and $b\geqslant-2\Lambda_0$, this makes $\mathbf{P}[\mathcal{Q}_-|\mathcal{K}]$ small. The rest of the argument is essentially the same. □

The previous lemmas will help us control the behavior of the continuation of contours near existing contours or the boundary of the domain. The next lemma will help us control the behavior in the interior away from existing contours.

LEMMA 3.10. (Obstacle) *For every $\varepsilon>0$ there is some constant $c=c(\varepsilon,\bar{\Lambda})>0$ such that the following estimate holds. Assume* (h) *and* (D). *Let $\mathcal{K}$ be the event that a fixed collection $\{\gamma_1,...,\gamma_k\}$ of oriented paths in* TG* *are contained in oriented zero-height interfaces of h, and suppose that $\mathbf{P}[\mathcal{K}]>0$. Let U be the union of V_∂ and the vertices of* TG *adjacent to $\bar{\gamma}:=\gamma_1\cup...\cup\gamma_k$. Let g be a function defined on the vertices of* TG *that is 0 on U. Let $\widehat{\gamma}_g$ denote the union of the interfaces of $h+g$ that contain any one of the paths $\gamma_1,...,\gamma_k$. Let $B(z_0,r)$ be a disk of radius r that is centered at some vertex z_0 satisfying $|g(z_0)|\geqslant\frac{1}{2}\|g\|_\infty$. Let $d>0$ and suppose that at distance at most $\varepsilon^{-1}d$ from z_0 there is a connected component of $\bar{\gamma}\cup\partial D$ whose diameter is at least εd. Also assume that $\|g\|_\infty/\|\nabla g\|_\infty>cr>c$. Then*

$$\mathbf{P}[\widehat{\gamma}_g\cap B(z_0,r)\neq\varnothing\,|\,\mathcal{K}]\leqslant c\|g\|_\infty^{-2}\log\frac{d}{r}.$$

Proof. With no loss of generality, we assume that $g(z_0)>0$. Set $q:=\|g\|_\infty/\|\nabla g\|_\infty$ and $r_1:=\frac{1}{10}q$. Since between any two vertices z and z' in TG there is a path in TG whose length is at most $2|z-z'|$,

$$\min\{g(z):z\in B(z_0,r_1)\}\geqslant g(z_0)-2r_1\|\nabla g\|_\infty=g(z_0)-\tfrac{1}{5}\|g\|_\infty\geqslant\tfrac{1}{4}\|g\|_\infty.$$

As $g=0$ on U, it follows that $\varepsilon^{-1}d\geqslant r_1$. Since we are assuming that $q>cr$, and we may assume that c is a large constant which may depend on ε, it follows that $d/r>100$, say. Thus, we also assume, with no loss of generality, that $\|g\|_\infty\geqslant\sqrt{c}$, since the required inequality is trivial otherwise.

Let X denote the average value of h on the vertices in $B(z_0,r)$. The inequality (3.5) and $d/r>100$ give

$$\mathbf{E}[X^2\,|\,\mathcal{K}]\leqslant O_{\varepsilon,\bar{\Lambda}}(1)\log\frac{d}{r}. \tag{3.27}$$

If γ_1 is not a closed path, we start exploring the interface of $h+g$ containing γ_1 starting from one of the endpoints of γ_1 until that interface is completed or $B(z_0,r)$ is hit, whichever occurs first. (This may entail going through several of the interfaces γ_j, $j>1$.) If that interface is completed before we hit $B(z_0,r)$, we continue and explore the interface of $h+g$ containing γ_2, and so forth, until finally either all of $\widehat{\gamma}_g$ is explored or $B(z_0,r)$ is hit. Let $\mathcal{Q}$ denote the event that $B(z_0,r)$ is hit, and let β be the interfaces explored up to the time when the exploration terminates.

Let U' be the union of U with the vertices adjacent to β. Since we are assuming that $q>cr$, $r>1$ and $\|g\|_\infty\geqslant\sqrt{c}$, and since c may be chosen arbitrarily large, Lemma 3.2 shows that for every vertex $v\in U'$, we have that

$$\mathbf{E}[|h(v)+g(v)|\,|\,\mathcal{K},\mathcal{Q},\beta]\leqslant\frac{\|g\|_\infty}{100}. \tag{3.28}$$

(Note that conditioning on $\mathcal{Q}$ and β amounts to conditioning that $h+g>0$ on vertices adjacent to the right-hand side of β and $h+g<0$ on vertices adjacent to the left-hand side. Consequently, the lemma applies.)

Now let x be any vertex in $B(z_0,r)$. For each $u\in U'$, let p_u denote the probability that a simple random walk started at x will first hit U' in u; that is, the discrete-harmonic measure from x. Then (3.28) gives

$$\mathbf{E}[h(x)\,|\,\mathcal{K},\mathcal{Q},\beta]\leqslant\frac{\|g\|_\infty}{100}-\sum_{u\in U'}p_u g(u). \tag{3.29}$$

Since $\beta\cap B(z_0,r)\neq\varnothing$ and β intersects the complement of $B(z_0,r_1)$, Lemma 2.1 gives

$$\sum_{u\in U'\setminus B(z_0,r_1)}p_u\leqslant O(1)\Big(\frac{r}{r_1}\Big)^{\zeta_1}.$$

Since we are assuming that $q\geqslant cr$, we may assume that the right-hand side is less than $\frac{1}{10}$. Recall that $g\geqslant\frac14\|g\|_\infty$ inside $B(z_0,r_1)$. Outside $B(z_0,r_1)$, the trivial estimate $g\geqslant-\|g\|_\infty$ applies. When these estimates are applied to (3.29), one gets

$$\begin{aligned}\mathbf{E}[h(x)\,|\,\mathcal{K},\mathcal{Q},\beta]-\frac{\|g\|_\infty}{100}&\leqslant-\sum_{u\in U'\cap B(z_0,r_1)}p_u g(u)-\sum_{u\in U'\setminus B(z_0,r_1)}p_u g(u)\\&\leqslant-\Big(\frac{\|g\|_\infty}{4}\Big)\Big(1-\frac{1}{10}\Big)+\frac{\|g\|_\infty}{10}=-\frac{\|g\|_\infty}{8}.\end{aligned}$$

We may take expectation with respect to β and average with respect to x to conclude that $\mathbf{E}[X\,|\,\mathcal{K},\mathcal{Q}]\leqslant-\frac19\|g\|_\infty$, which implies, by Jensen's inequality, that

$$\mathbf{E}[X^2\,|\,\mathcal{K},\mathcal{Q}]\geqslant\frac{\|g\|_\infty^2}{81}.$$

Since

$$\mathbf{P}[\mathcal{Q}\,|\,\mathcal{K}]\leqslant\frac{\mathbf{E}[X^2\,|\,\mathcal{K}]}{\mathbf{E}[X^2\,|\,\mathcal{Q},\mathcal{K}]},$$

the lemma now follows from the above and (3.27). □

3.4. Barriers

In this subsection we apply Lemmas 3.8, 3.10 and 3.6 to get a flexible (though slightly complicated) criterion giving lower bounds for the probability that contours avoid certain sets. The complications arise from the need to handle pre-existing contours that are highly non-smooth on large scales.

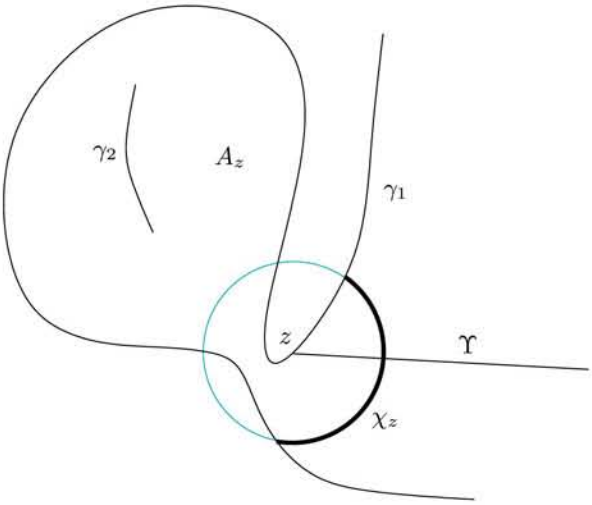


Figure 3.3. The domain A_z in a situation where condition (3) fails. (The figure does not show detail on the scale of the lattice; that is, $D(\bar{\gamma})$ appears as $D\backslash\bar{\gamma}$.)

The following relative notion of distance will sometimes be used below:

$$\operatorname{dist}(A,B;X) := \inf\{\operatorname{diam}\alpha : \alpha \text{ is a path in } \overline{X} \text{ connecting } A \text{ and } B\}. \tag{3.30}$$

If $\bar{\gamma}$ is a collection of paths in $\overline{D}$, let $D(\bar{\gamma})$ denote the complement in D of the union of the closed triangles of TG meeting $\bar{\gamma}$.

We now define the notion of barrier. Assume (h). Let $\gamma_1, \gamma_2, ..., \gamma_k$ be disjoint simple paths or simple closed paths in $\overline{D}$. Set $\bar{\gamma}:=\gamma_1\cup...\cup\gamma_k$. Let $\Upsilon\subset\overline{D(\bar{\gamma})}$ be a path, which is contained in $D(\bar{\gamma})$, except possibly for its two endpoints. Fix some $R>0$ and $\varepsilon>0$. We call Υ an (ε, R)-*barrier* for the configuration $(D,\bar{\gamma})$ if the following conditions hold:

(1) $\varepsilon R<\operatorname{diam}\Upsilon\leqslant R$;
(2) within distance $\varepsilon^{-1}R$ from Υ there is a connected component of $\bar{\gamma}\cup\partial D$ whose diameter is at least εR;
(3) if $z\in\partial D(\bar{\gamma})$ is an endpoint of Υ and χ_z is the connected component of $\partial B(z,\varepsilon R)\cap D(\bar{\gamma})$ first encountered when traversing Υ from z (which exists by (1)), then the connected component A_z of $D(\bar{\gamma})\backslash\chi_z$ that contains points of Υ arbitrarily close to z satisfies (a) ∂A_z consists of χ_z and a simple path contained in $\partial D(\bar{\gamma})$ and (b) $\Upsilon\cap A_z\cap\partial B(z,r)$ consists of a single point for every $r\in(0,\varepsilon R)$;
(4) for every point $w\in\Upsilon$ such that $\operatorname{dist}(w,\partial D(\bar{\gamma}))\leqslant\frac{1}{5}\varepsilon R$ there is an endpoint $z\in\Upsilon\cap\partial D(\bar{\gamma})$ such that $w\in A_z$ and $|w-z|<\frac{1}{2}\varepsilon R$ (roughly, Υ does not get close to $\partial D(\bar{\gamma})$ except near its endpoints on $\partial D(\bar{\gamma})$).

One example where condition (3) fails is given in Figure 3.3. If we remove the strand of γ_2 from that figure, we get an example that illustrates that $A_z\subset\overline{B(z,\varepsilon R)}$ does not

follow from the above conditions. Note that it may happen that $\partial A_z \setminus \chi_z$, which is a simple path, by (3), consists of an arc in ∂D together with one or two arcs in $\bar{\gamma}$ that have endpoints in ∂D.

We now use barriers to "manipulate" contours of the DGFF.

THEOREM 3.11. (Barriers) *For every $\varepsilon>0$, $m\in\mathbb{N}_+$ and $\bar{\Lambda}>0$ there is a*

$$p = p(\varepsilon, \bar{\Lambda}, m) > 0$$

such that the following estimate holds. Assume (h) *and* (D). *Let $\mathcal{K}$ be the event that a fixed collection $\{\gamma_1, ..., \gamma_k\}$ of oriented paths in* TG^* *is contained in the oriented zero-height interfaces of h, and suppose that $\mathbf{P}[\mathcal{K}]>0$. Set*

$$\bar{\gamma} := \bigcup_{j=1}^{k} \gamma_j.$$

Let V_+ (respectively, V_-) be the set of vertices adjacent to $\bar{\gamma}$ on the right-hand (respectively, left-hand) side. Let $R>0$ and let $Y=Y_+\cup Y_-$ be a collection of m (ε, R)-barriers for the configuration $(D, \bar{\gamma})$. Assume that the endpoints of these barriers are not on ∂D and that for every $\Upsilon\in Y_+$ (respectively, $\Upsilon\in Y_-$) and every endpoint $z\in\Upsilon\cap\partial D(\bar{\gamma})$, the vertices in $\overline{A_z(\Upsilon)}\cap\partial D(\bar{\gamma})$ are in V_+ (respectively, V_-), where $A_z(\Upsilon)$ is as in condition (3). *Also assume that* $\operatorname{dist}(\bigcup Y_+, \bigcup Y_-; D(\bar{\gamma})) \geqslant 2\varepsilon R$. *In the situation where $\varepsilon R = O(1)$, we also need to assume that there is no hexagon in* TG^* *meeting both $V_+\cup(\bigcup Y_+)$ and $V_-\cup(\bigcup Y_-)$ and there is no hexagon meeting $\bigcup Y$ and ∂D. Let $\widehat{\gamma}$ denote the union of the zero-height interfaces of h which contain any one of the arcs $\gamma_1, ..., \gamma_k$. Then*

$$\mathbf{P}[(\widehat{\gamma}\setminus\bar{\gamma})\cap(\textstyle\bigcup Y) = \varnothing \mid \mathcal{K}] > p.$$

The basic idea of the proof is as follows. We define a function g that is large (positive) near $\bigcup Y_+$ away from $\partial D(\bar{\gamma})$ and is negative and large in absolute value near $\bigcup Y_-$ away from $\partial D(\bar{\gamma})$. The obstacle lemma (Lemma 3.10) will then imply that $\widehat{\gamma}_g$, as defined there, is unlikely to hit Υ, except near endpoints of barriers. The narrows lemma (Lemma 3.8) will be used to show that $\widehat{\gamma}_g$ is also unlikely to hit $\bigcup Y$ near endpoints. Finally, the distortion lemma (Lemma 3.6) will be used to conclude that with probability bounded away from zero, $\widehat{\gamma}$ will not hit Υ.

Proof. Let

$$\hat{A}_+^1 := \bigcup\{A_z(\Upsilon)\cap B(z, \varepsilon R) : \Upsilon \in Y_+ \text{ and } z \in \Upsilon\cap\bar{\gamma}\},$$
$$\hat{A}_+^2 := \big\{z \in D(\bar{\gamma}) : \operatorname{dist}(z, \textstyle\bigcup Y_+; D(\bar{\gamma})) \leqslant \frac{1}{10}\varepsilon R\big\},$$

and $\hat{A}_+ := \hat{A}_+^1 \cup \hat{A}_+^2$. Similarly, define $\hat{A}_-$, with Y_- replacing Y_+. We fix constants $c_0 > 0$, large, and $\delta > 0$ much smaller than ε, and set

$$g(z) := \begin{cases} c_0 \delta^{-1} R^{-1} \min\{\delta R, \operatorname{dist}(z, \mathbb{R}^2 \setminus \hat{A}_+)\}, & \text{if } z \in \hat{A}_+, \\ -c_0 \delta^{-1} R^{-1} \min\{\delta R, \operatorname{dist}(z, \mathbb{R}^2 \setminus \hat{A}_-)\}, & \text{if } z \in \hat{A}_-, \\ 0, & \text{otherwise.} \end{cases}$$

Note that $\hat{A}_+ \cap \hat{A}_- = \varnothing$ and $\|g\|_\infty = c_0$.

Let $Z := \{z \in \mathbb{R}^2 : |g(z)| = c_0\}$. Let $c_1 > 1$ be some large constant and set $r := \delta R / c_1$. Let W be a maximal collection of vertices in $\{v : |g(v)| \geqslant \frac{1}{2} c_0\}$ such that the distance between any two distinct vertices in W is at least $\frac{1}{3} r$, and for $a \in W$ let B_a denote the disk of radius r centered at a. Then the disks B_a, $a \in W$, cover Z, assuming that $r > 1$ and $c_1 > 100$, say. Note that

$$|W| = O(m) \left(\frac{R}{r} \right)^2 = O(m) \left(\frac{c_1}{\delta} \right)^2.$$

Fix some $a \in W$. We wish to invoke Lemma 3.10 to get a good upper bound on $\mathbf{P}[\widehat{\gamma}_g \cap B_a \neq \varnothing | \mathcal{K}]$. We now verify the assumptions of the lemma. We note that in our case $\|g\|_\infty = c_0$ and $\|\nabla g\|_\infty = c_0 / \delta R$. Thus, $q := \|g\|_\infty / \|\nabla g\|_\infty = \delta R$. Consequently, we set c_1 to be the maximum of 100 and twice the constant c in the lemma, and the assumption $q > cr$ is satisfied. The assumption $r > 1$ will hold once R is large enough, which we assume for now (we promise that δ will be a constant depending only on ε, m and $\bar{\Lambda}$). For the d in the lemma we may take $d = R$. Thus, $\log(d/r) = \log(c_1/\delta)$, and the lemma gives

$$\mathbf{P}[\widehat{\gamma}_g \cap B_a \neq \varnothing \,|\, \mathcal{K}] \leqslant O_{\varepsilon, \bar{\Lambda}}(1) \frac{1}{c_0^2} \log \frac{c_1}{\delta}.$$

Since $\bigcup_{a \in W} B_a \supset Z$ and $|W| = O(m)(c_1/\delta)^2$, we conclude that

$$\mathbf{P}[\widehat{\gamma}_g \cap Z \neq \varnothing \,|\, \mathcal{K}] \leqslant O_{\varepsilon, \bar{\Lambda}}(m) \frac{1}{c_0^2} \left(\frac{c_1}{\delta} \right)^2 \log \frac{c_1}{\delta}.$$

Although we have not specified δ yet, we choose $c_0 = c_0(c_1, \delta, \varepsilon, m, \bar{\Lambda})$ so that

$$\mathbf{P}[\widehat{\gamma}_g \cap Z \neq \varnothing \,|\, \mathcal{K}] \leqslant \tfrac{1}{10}. \tag{3.31}$$

Now that we have established that it is unlikely that $\widehat{\gamma}_g$ intersects Z, we need to worry about the case in which $\widehat{\gamma}_g$ circumvents Z but hits $\bigcup Y \setminus Z$. This can only happen near endpoints of barriers. Let us fix some $\Upsilon \in Y_+$ that has an endpoint, say z_1, on $\partial D(\bar{\gamma})$. Let $\widetilde{\chi}_{z_1}$ be the arc $\partial A_{z_1}(\Upsilon) \setminus \chi_{z_1}$. (It is an arc, by condition (3) of the definition of barrier.) We will now prepare the geometric setup that will enable the use of Lemma 3.8 to prove that $\mathbf{P}[\widehat{\gamma}_g \cap A_{z_1}(\Upsilon) \cap \Upsilon \setminus Z \neq \varnothing | \mathcal{K}]$ is small.

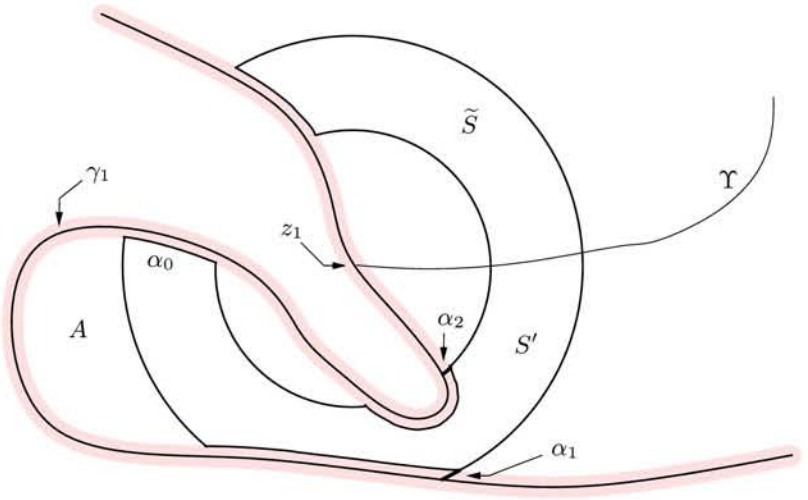


Figure 3.4. The set S, and the paths α_0 and $\widetilde{\alpha}_0$. The shaded region is Q.

Let Q be the set of all hexagons of the grid TG^* whose distance from $\partial D(\bar{\gamma})$ is at most $2\delta R$. Let S be the connected component of

$$A_{z_1} \cap B\bigl(z_1, \tfrac{7}{8}\varepsilon R\bigr) \setminus \bigl(Q \cup B\bigl(z_1, \tfrac{5}{8}\varepsilon R\bigr)\bigr)$$

that intersects Υ. (See Figure 3.4.) Condition (4) in the definition of barriers and our assumption that $\delta \ll \varepsilon$ guarantees that there is a unique such component S. We have $S \subset Z$, and so $\widehat{\gamma}_g$ is unlikely to hit S.

Let S' and $\widetilde{S}$ be the two connected components of $S \setminus \Upsilon$. Consider the connected component M_1 of $\partial B\bigl(\frac{7}{8} z_1 \varepsilon R\bigr) \setminus \widetilde{\chi}_{z_1}$ that intersects Υ. Let α_1 be the arc in $M_1 \setminus \overline{S}$ that has one endpoint in $\overline{S}'$ and the other in $\widetilde{\chi}_{z_1}$. Likewise, let M_2 be the connected component of $\partial B\bigl(\frac{5}{8} z_1 \varepsilon R\bigr) \setminus \widetilde{\chi}_{z_1}$ that intersects Υ, and let α_2 be the arc in $M_2 \setminus \overline{S}$ that has one endpoint in $\overline{S}'$ and the other in $\widetilde{\chi}_{z_1}$. Let α denote the union of $\alpha_1 \cup \alpha_2$ with the arc of $\partial S' \setminus \Upsilon$ that connects the endpoint of α_1 with the endpoint of α_2. Let $A \subset A_{z_1}(\Upsilon)$ be the domain whose boundary consists of $\alpha_1 \cup \alpha_2$, an arc of $\partial S'$ connecting α_1 and α_2 and an arc of $\widetilde{\chi}_{z_1}$ connecting α_1 and α_2. Let α_0 be the unique arc of $\partial A \setminus \bigl(\partial B\bigl(\frac{5}{8} z_1 \varepsilon R\bigr) \cup \partial B\bigl(\frac{7}{8} z_1 \varepsilon R\bigr)\bigr)$ connecting $\partial B\bigl(\frac{5}{8} z_1 \varepsilon R\bigr)$ with $\partial B\bigl(\frac{7}{8} z_1 \varepsilon R\bigr)$. Finally, let α' be any subarc of α_0 whose diameter is $\frac{1}{16}\varepsilon R$ that intersects the circle $\partial B\bigl(z_1, \frac{3}{4}\varepsilon R\bigr)$. We use A, α_1, α_2 and α' in Lemma 3.8. In the present situation, the value of d_1 of that lemma is $d_1 = 2\delta R + O(1)$. If we have a path connecting α' to $\partial D \cup \bar{\gamma}$, it must either exit $B\bigl(z_1, \frac{7}{8}\varepsilon R\bigr) \setminus B\bigl(z_1, \frac{5}{8}\varepsilon R\bigr)$, hit $\Upsilon \setminus B\bigl(z_1, \frac{5}{8}\varepsilon R\bigr)$ (whose distance from $\bar{\gamma}$ is at least $\frac{1}{5}\varepsilon R$) or connect α_0 to $\widetilde{\chi}_{z_1}$ inside A. Consequently, the minimum on the right-hand side in (3.24) is presently at least $\frac{1}{16}\varepsilon R$. The lemma now implies that if we choose our current $\delta = \delta(\varepsilon, m, \bar{\Lambda}) > 0$ sufficiently small

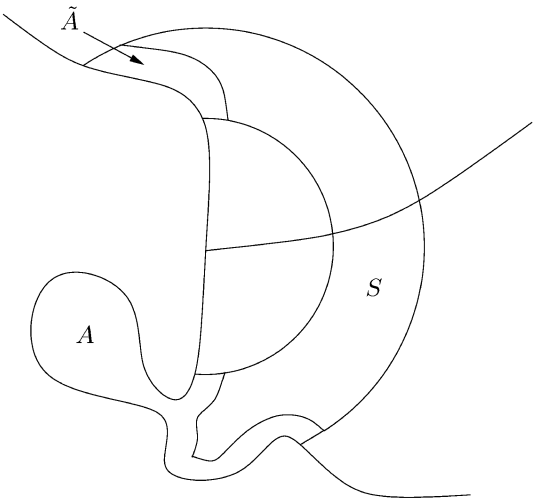


Figure 3.5. No way to penetrate.

and make sure that $R>1/\delta$, then

$$\mathbf{P}[\mathcal{C} \,|\, \mathcal{K}] \leqslant \frac{1}{10m},$$

where $\mathcal{C}$ is the event that there is a crossing of hexagons satisfying $h<0$ between α_1 and α_2 inside A. If there is no such crossing, then also $\widehat{\gamma}_g$ does not make such a crossing, because $g \geqslant 0$ in A.

Likewise, we may define $\widetilde{\alpha}_1$, $\widetilde{\alpha}_2$, $\widetilde{\alpha}'$, $\tilde{A}$ and $\widetilde{\mathcal{C}}$, when we replace S' by $\widetilde{S}$ in the above paragraph. The same argument shows that $\mathbf{P}[\widetilde{\mathcal{C}}|\mathcal{K}] \leqslant \frac{1}{10} m^{-1}$.

Condition (3) of the definition of a barrier implies that $\bar{S} \cup A \cup \tilde{A} \cup \chi_{z_1}$ separates $\Upsilon \cap A_{z_1}(\Upsilon)$ from all the endpoints of the strands $\gamma_1, ..., \gamma_k$. Consequently, in order for $\widehat{\gamma}_g$ to hit $\Upsilon \cap A_{z_1}(\Upsilon)$, we must have $\widehat{\gamma}_g \cap S \neq \varnothing$ or $\mathcal{C} \cup \widetilde{\mathcal{C}}$. See Figure 3.5 (also compare Figure 3.4). A similar argument applies for every other endpoint of a barrier. Since $S \subset Z$, we conclude that

$$\mathbf{P}[\widehat{\gamma}_g \cap (\cup Y) \neq \varnothing \,|\, \mathcal{K}] \leqslant \tfrac{3}{10}. \tag{3.32}$$

We are really more interested in $\widehat{\gamma}$ than in $\widehat{\gamma}_g$. To do the translation, we will appeal to Lemma 3.6. For this purpose, note that

$$\|\nabla g\|_\infty = O(1)\frac{c_0}{\delta R} = O_{\varepsilon,\bar{\Lambda},m}\left(\frac{1}{R}\right)$$

and that g is supported in a union of m sets of diameter $O(R)$. Consequently, we have $\|\nabla g\|_2 = O_{\varepsilon,\bar{\Lambda},m}(1)$. Let U be the set of vertices in $\partial D(\bar{\gamma})$, and let h_U denote the restriction

of h to U. By (3.32), even if we further condition on h_U, the left-hand side stays bounded away from one on an event whose probability is bounded away from zero, namely,

$$\mathbf{P}\big[\mathbf{P}[\widehat{\gamma}_g \cap (\bigcup Y) = \varnothing \,|\, \mathcal{K}, h_U] > \tfrac{1}{10} \,\big|\, \mathcal{K}\big] > \tfrac{1}{10}.$$

Because $g=0$ on U, we may apply Lemma 3.6 and conclude that for h_U such that the inner inequality above holds,

$$O_{\varepsilon,\bar{\Lambda},m}(1)\mathbf{P}[\widehat{\gamma} \cap (\bigcup Y) = \varnothing \,|\, \mathcal{K}, h_U] \geqslant 1.$$

Since this set of h_U has conditioned probability at least $\frac{1}{10}$, the theorem follows.

It remains to remove the assumption that R is larger than some fixed constant $R_0 = R_0(\varepsilon, m, \bar{\Lambda})$. Assume now that R is bounded. It is not too hard to see that the event $\mathcal{H}$ that $h(H) \in (0,1)$ for every hexagon meeting $\bigcup Y_+$ and $h \in (-1,0)$ for every hexagon meeting $\bigcup Y_-$ has probability bounded below by (a rather small) positive constant. This is proved by considering these $O(mR^2)$ hexagons one by one. On the event $\mathcal{H}$, we have $\widehat{\gamma} \cap (\bigcup Y) = \varnothing$. This completes the proof. □

Remark 3.12. There is a corresponding analog of Theorem 3.11 in the case where the endpoints of the barriers are permitted to land on ∂D. In that case, it is necessary to assume that $h_\partial \geqslant -\Lambda_0$ (respectively, $h_\partial \leqslant \Lambda_0$) on $\partial A_z(\Upsilon) \cap V_\partial$ if $\Upsilon \in Y_+$ (respectively, Y_-) and $z \in \partial D \cap \Upsilon$, where $\Lambda_0 = \Lambda_0(\bar{\Lambda}) > 0$ is the constant given by Lemma 3.9. We refrain from stating a complete formulation of this variant, though it will be useful. The proof is the same, except that Lemma 3.9 is used to deal with the narrows near ∂D, instead of the narrows lemma (Lemma 3.8).

3.5. Meeting of random walk and interface

We now need to further develop the basic setup and introduce some more notation. If α is a path in the hexagonal grid TG^*, we let $V(\alpha)$ denote the set of TG-vertices adjacent to it. If α is an arc of an oriented zero-height interface of h, let $V_+(\alpha)$ denote the vertices adjacent to it on its right-hand side, and let $V_-(\alpha)$ denote the vertices adjacent to it on its left-hand side.

In addition to our previous assumptions (h) and (D), we will use the following setup:

(S) Let γ denote the interface of h from x_∂ to y_∂. Let v_0 be some vertex of TG in D, and let S be a simple random walk on the vertices of TG started at v_0 that is independent of h. Let τ be the first time t such that $S_t \in \partial D(\gamma)$.

The point S_τ will play a special role. Essentially, we will be interested in the configuration "as viewed from S_τ"; that is, in the coordinate system where S_τ is translated to 0.

In order to eliminate too much additional notation, it will be convenient to consider the event $S_\tau=0$ instead. Let τ_0 be the first t such that $S_t=0$, and let $\breve{S}$ denote the reversed walk $\breve{S}_t:=S_{\tau_0-t}$, $t=0,1,...,\tau_0$.

For $\sigma=0,1,...,5$, let e_σ denote the edge $\bigl[0,\exp\bigl(\frac{1}{3}\pi i\sigma\bigr)\bigr]$ of the triangular grid TG, and let e_σ^* denote the dual edge in TG*. Let $\mathcal{Z}_0^\sigma$ denote the event $\mathcal{Z}_0^\sigma:=\{S_\tau=0\}\cap\{e_\sigma^*\subset\gamma\}$.

Fix some large R. Suppose that $\mathfrak{B}_R\subset D$ and $v_0\notin\mathfrak{B}_R$. Let $\mathrm{ext}_R\gamma$ denote the union of the components of $\gamma\setminus\mathfrak{B}_R$ containing x_∂ and y_∂. (If $\gamma\cap\mathfrak{B}_R=\varnothing$, then $\mathrm{ext}_R\gamma=\gamma$.) If there is an interface of h containing e_σ^*, denote it by $\hat{\beta}=\hat{\beta}^\sigma$, and let $\beta=\beta_R^\sigma$ be the connected component of $\hat{\beta}\cap\mathfrak{B}_R$ that contains e_σ^*. Otherwise, set $\beta=\hat{\beta}=\varnothing$. Let $\mathrm{int}_R\breve{S}$ denote the part of $\breve{S}$ up to the first exit of $\mathfrak{B}_R$, and let $\mathrm{ext}_R S$ denote the part of S up to the first entry to $\mathfrak{B}_R$.

Set $\Phi_R:=(D,\partial_+,h_\partial,v_0,\mathrm{ext}_R\gamma,\mathrm{ext}_R S)$ and $\Theta_R=\Theta_R(\sigma):=(\beta_R^\sigma,\mathrm{int}_R\breve{S})$. Our goal is to show that conditioned on $\mathcal{Z}_0^\sigma$, the distribution of β does not depend strongly on Φ_{4R}. (A precise version of this statement is given in Corollary 3.16 below.) To this end we will use something like

$$\mathbf{P}[\Theta_R=\vartheta\mid\Phi_{3R},\mathcal{Z}_0^\sigma]=\frac{\mathbf{P}[\Theta_R=\vartheta\mid\Phi_{3R}]\mathbf{P}[\mathcal{Z}_0^\sigma\mid\Theta_R=\vartheta,\Phi_{3R}]}{\mathbf{P}[\mathcal{Z}_0^\sigma\mid\Phi_{3R}]}.\tag{3.33}$$

This equality is obtained by applying Bayes' formula to the measure $\mathbf{P}[\,\cdot\,|\Phi_{3R}]$. The following lemma takes care of the first factor in the numerator on the right-hand side.

LEMMA 3.13. *Assume* (h), (D) *and* (S). *There exists a constant* $c=c(\bar{\Lambda})>0$ *and a function* $p_R(\cdot)$ *such that if* $R>50$, $R'\in\bigl[\frac{5}{4}R,3R\bigr]$, $D\supset\mathfrak{B}_{4R}$ *and* $v_0\notin\mathfrak{B}_{4R}$, *then for all* $\vartheta=(\tilde{\beta},\widetilde{S})$ *such that* $\tilde{\beta}\neq\varnothing$ *one has*

$$\frac{1}{c}p_R(\vartheta)\leqslant\mathbf{P}[\Theta_R=\vartheta\mid\gamma\cap\mathfrak{B}_{4R}\neq\varnothing,\Phi_{R'}]\leqslant cp_R(\vartheta).$$

The function p_R *may depend on* R *and* ϑ, *but not on anything else* (*in particular*, *not on* D, v_0, $\Phi_{R'}$ *or* h_∂).

Proof. The corresponding statement with Θ_R replaced by β, the first coordinate of Θ_R, is an immediate consequence of Proposition 3.7.

We assume that $v_0\notin\mathfrak{B}_{4R}$. The configuration $\Phi_{R'}$ determines the first vertex, say q, inside $\mathfrak{B}_{R'}$ visited by the random walk S. The continuation of the walk is just a simple random walk starting at q. Suppose that we had another such walk starting at a vertex $q'\in\mathfrak{B}_{R'}$. It is easy to see that with probability bounded away from zero the walk starting at q visits q' before 0. If that happens, we couple the continuation of the walk to be the same as the walk which starts at q' (otherwise, we let them be independent). On the event that the walk started at q hits q' before 0, the corresponding $\mathrm{int}_R\breve{S}$ for both walks will be the same. This proves the corresponding statement about the second coordinate of Θ_R. Since the two coordinates are independent given $\Phi_{R'}$, the lemma follows. □

Proving an analogous result for the second factor in the numerator of the right-hand side of (3.33) will be considerably more difficult. To this end, we now define a measure of the *quality* $Q{=}Q_R$ of the configurations Φ_R and Θ_R.

If $\gamma{\cap}\mathfrak{B}_R{\neq}\varnothing$, let x^R (respectively, y^R) denote the endpoint in $\partial\mathfrak{B}_R$ of the component of $\mathrm{ext}_R\gamma$ containing x_∂ (respectively, y_∂). When $v_0{\notin}\mathfrak{B}_R$, let q^R denote the vertex in $\mathfrak{B}_R$ first visited by S. If $\gamma{\cap}\mathfrak{B}_R{=}\varnothing$ or $\mathrm{ext}_R S$ visits $\partial D(\mathrm{ext}_R\gamma)$, then set $Q(\Phi_R){=}0$. Otherwise, define

$$Q(\Phi_R) := \frac{\mathrm{dist}(x,\mathrm{ext}_R S){\wedge}\mathrm{dist}(y,\mathrm{ext}_R S){\wedge}\mathrm{dist}(q,\mathrm{ext}_R\gamma){\wedge}|x{-}y|}{R}{\wedge}\frac{1}{100},$$

where $x{=}x^R$, $y{=}y^R$ and $q{=}q^R$. This is a measure of the separation between the strands comprising Φ_R. Similarly, define $Q(\Theta_R)$, as follows. Suppose that $v_0{\notin}\mathfrak{B}_R{\subset}D$ and let $\hat{q}^R$ be the first vertex outside of $\mathfrak{B}_R$ visited by $\breve{S}$. Fix an orientation of e^*_σ. If $\hat{\beta}{\not\subset}\mathfrak{B}_R$, let $\hat{x}^R$ and $\hat{y}^R$ be the two endpoints of the component of $\beta{=}\beta_R$ containing e^*_σ, chosen so that the orientation of the arc of β from $\hat{x}^R$ to $\hat{y}^R$ agrees with that of e^*_σ. If $\hat{\beta}{\subset}\mathfrak{B}_R$ or if $\breve{S}$ visits any vertex in $\partial D(\beta_R)\setminus\{0\}$, then set $Q(\Theta_R){=}0$. Otherwise, set

$$Q(\Theta_R) := \frac{\mathrm{dist}(\hat{x},\mathrm{int}_R\breve{S}){\wedge}\mathrm{dist}(\hat{y},\mathrm{int}_R\breve{S}){\wedge}\mathrm{dist}(\hat{q},\beta){\wedge}|\hat{x}{-}\hat{y}|}{R}{\wedge}\frac{1}{100},$$

where $\hat{x}{=}\hat{x}^R$, $\hat{y}{=}\hat{y}^R$ and $\hat{q}{=}\hat{q}^R$.

LEMMA 3.14. (Compatibility) *Assume* (h), (D) *and* (S). *For every* $\varepsilon{>}0$ *there is a constant* $c{=}c(\varepsilon,\bar{\Lambda}){>}0$ *such that*

$$\frac{1}{c}\,1_{\{Q(\Theta_R)>\varepsilon\}}1_{\{Q(\Phi_{R'})>\varepsilon\}} \leqslant \mathbf{P}[\mathcal{Z}_0^\sigma \mid \Theta_R,\Phi_{R'}]\log R \leqslant c$$

holds whenever $R{>}c$, $5R{>}R'{>}\frac{9}{8}R$ *and* $v_0{\notin}\mathfrak{B}_{6R}{\subset}D$.

Proof. We start by proving the lower bound on $\mathbf{P}[\mathcal{Z}_0^\sigma\,|\,\Theta_R,\Phi_{R'}]$, which is the harder estimate. Assume that $Q(\Theta_R){\wedge}Q(\Phi_{R'}){>}\varepsilon$, $R{>}c{>}10^{10}/\varepsilon$, $5R{>}R'{>}\frac{9}{8}R$ and $v_0{\notin}\mathfrak{B}_{6R}{\subset}D$. Let $\widehat{D}_S$ denote the connected component of $\mathfrak{B}_R\setminus\beta_R$ that intersects $\mathrm{int}_R\breve{S}$ and let D_S denote the connected component of $D\setminus(\mathrm{ext}_{R'}\gamma{\cup}\mathfrak{B}_{R'})$ that contains v_0 and therefore $\mathrm{ext}_{R'}S$. Note that the sign of h on vertices in $\widehat{D}_S$ adjacent to β is constant, as is the sign of h on vertices in D_S adjacent to $\mathrm{ext}_{R'}\gamma$. Let $\mathcal{D}$ denote the event that these signs are the same, namely, the sign of h on vertices in $V(\mathrm{ext}_{R'}\gamma){\cap}D_S$ is the same as on vertices in $V(\beta_R){\cap}\widehat{D}_S$. Using symmetry, Proposition 3.7 immediately implies that $\mathbf{P}[\mathcal{D}|\Phi_{R'},\Theta_R]$ is bounded away from zero. (Although β is determined by Θ_R, its orientation as a subarc of an oriented zero-height interface of h is not determined by Θ_R.)

We now construct some barriers, as illustrated in Figure 3.6. Let a be the initial point of the arc $\partial\mathfrak{B}_{R'}{\cap}\overline{D}_S$, when the arc is oriented counterclockwise around $\mathfrak{B}_{R'}$, and

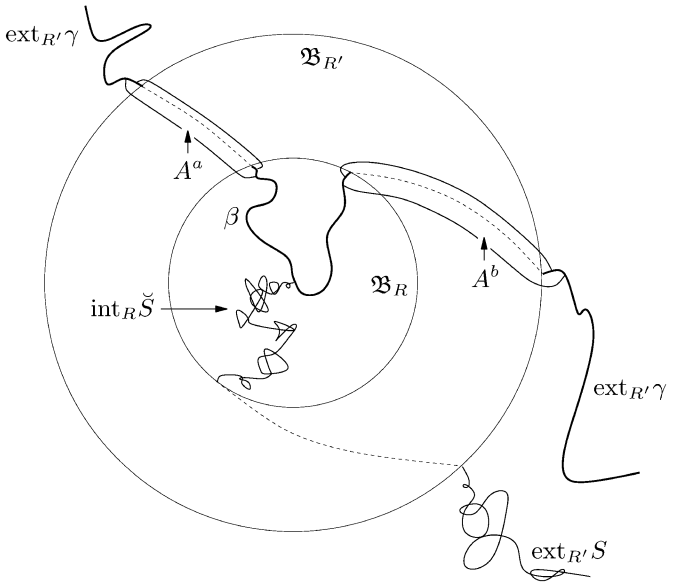


Figure 3.6. The construction of the barriers.

let b be the other endpoint of this arc. Likewise, let $\hat{a}$ be the initial point of the arc $\partial\mathfrak{B}_R\cap\overline{\widehat{D}}_S$, when the arc is oriented counterclockwise around $\mathfrak{B}_R$, and let $\hat{b}$ be the other endpoint of this arc. Note that $\{\hat{a},\hat{b}\}=\{\hat{x}^R,\hat{y}^R\}$ and $\{a,b\}=\{x^{R'},y^{R'}\}$.

We now describe a path α_a connecting a to $\hat{a}$ and a path α_b connecting b to $\hat{b}$ and a path α_q connecting $\hat{q}^R$ to $q^{R'}$ such that these paths do not come too close to each other. For example, if the arguments $\arg a$, $\arg b$, $\arg q^{R'}$, $\arg\hat{a}$, $\arg\hat{b}$ and $\arg\hat{q}^R$ are chosen so that $\arg\hat{a}<\arg\hat{q}^R<\arg\hat{b}<\arg\hat{a}+2\pi$, $\arg a<\arg q^R<\arg b<\arg a+2\pi$ and $|\arg a-\arg\hat{a}|\leqslant\pi$, then we may take α_a to be defined in polar coordinates by $\theta=s+tr$, with s and t chosen so that a and $\hat{a}$ are on the path, and similarly for α_b and α_q. It is easy to check that our assumptions guarantee that the distance between any two of these paths is at least $c_1\varepsilon R$ for some constant $c_1>0$. Set $\varepsilon'=\frac{1}{10}c_1\varepsilon$.

Let D_+ be the connected component of $D\setminus(\mathrm{ext}_{R'}\gamma\cup\beta_R\cup\alpha_a\cup\alpha_b)$ that contains $V_+(\mathrm{ext}_{R'}\gamma)$, and let D_- be the other connected component. Let α'_a be the connected component of $\{z\in D_+:\mathrm{dist}(z,\alpha_a)=\varepsilon' R\}$ that meets the circle $\partial B\bigl(0,\frac{1}{2}(R+R')\bigr)$. Note that α'_a is a simple path which intersects the circle $\partial B\bigl(0,\frac{1}{2}(R+R')\bigr)$ at one point. Let a_0 denote that point. We want to construct a pertubation of α'_a, which will be some $(\varepsilon'',12R)$-barrier, with ε'' not much smaller than ε'. Let a_1 be the closest point to a_0 along α'_a such that the distance from a_1 to $\partial D(\mathrm{ext}_{R'}\gamma)$ is $\frac{1}{10}\varepsilon' R$, and let $\hat{a}_1$ be the closest point to a_0 along α'_a such that the distance from $\hat{a}_1$ to $\partial D(\beta_R)$ is $\frac{1}{10}\varepsilon' R$. Let α^a_+ be the path which is the union of the arc of α'_a connecting a_1 and $\hat{a}_1$ together with a shortest

line segment connecting a_1 to $\partial D(\mathrm{ext}_{R'}\gamma)$ and a shortest line segment connecting $\hat{a}_1$ to $\partial D(\beta_R)$.

We claim that α^a_+ is an $(\varepsilon'', 12R)$-barrier for the configuration $(D, \mathrm{ext}_{R'}\gamma \cup \beta_R)$ with $\varepsilon'' = \frac{1}{1000}\varepsilon'$. Indeed, conditions (1), (2), (4) and (3) (b) in the definition of the barrier clearly hold. To verify condition (3) (a), let z_1 be the endpoint of α^a_+ on $\partial D(\mathrm{ext}_{R'}\gamma)$, and let z_1' and z_1'' be the two endpoints of χ_{z_1} on $\partial D(\mathrm{ext}_{R'}\gamma)$. Consider the simple arc $\widetilde{\chi}$ connecting z_1' to z_1'' in $\partial D(\mathrm{ext}_{R'}\gamma)$. By the Jordan curve theorem, $\widetilde{\chi} \cup \chi_{z_1}$ separates the plane into two connected components. Since α^a_+ crosses χ_{z_1}, it follows that the part of α^a_+ inside $\mathfrak{B}_{R'}$ is outside of A_{z_1}, and thus the endpoints $x^{R'}$, $y^{R'}$ and also β_R are all outside A_{z_1}. It follows that $\partial A_{z_1} = \chi_{z_1} \cup \widetilde{\chi}$, as required. A similar argument applies near the endpoint of α^a_+ on $\partial D(\beta_R)$. Thus, α^a_+ is indeed an (ε'', R')-barrier. Note also that the above easily implies that $\partial A_{z_1} \subset D_+$. This will be useful below when we apply the barriers theorem (Theorem 3.11).

We similarly construct a path α^a_- in D_- close to α_a. Likewise, we construct barriers α^b_+ and α^b_- near the path α_b. The construction is the same, except that we replace α_a by α_b.

On the event $\mathcal{D}$, we may apply Theorem 3.11 with $Y_+ = \{\alpha^a_+, \alpha^b_+\}$, $Y_- = \{\alpha^a_-, \alpha^b_-\}$, $Y = Y_+ \cup Y_-$ and $\bar{\gamma} = \mathrm{ext}_{R'}\gamma \cup \beta_R$. (Here we use the assumption that $R > c$.) Note that conditioning on Θ_R, $\Phi_{R'}$ and $\mathcal{D}$, amounts to conditioning on $\mathcal{K}$ in the theorem and on the behavior of $\mathrm{int}_R \breve{S} \cup \mathrm{ext}_{R'} S$, which is anyway independent of h. Therefore, there is a $p = p(\varepsilon, \bar{\Lambda}) > 0$ such that

$$\mathbf{P}[\mathcal{Y} \mid \Theta_R, \Phi_{R'}, \mathcal{D}] \geqslant p,$$

where $\mathcal{Y}$ denotes the event

$$\mathcal{Y} := \{(\gamma \setminus \bar{\gamma}) \cap (\bigcup Y) = \varnothing\}.$$

Let A^a be the connected component of $D(\mathrm{ext}_{R'}\gamma \cup \beta_R) \setminus (\alpha^a_+ \cup \alpha^a_-)$ that contains α_a, and let A^b be the connected component of $D(\mathrm{ext}_{R'}\gamma \cup \beta_R) \setminus (\alpha^b_+ \cup \alpha^b_-)$ that contains α_b. Again, using the Jordan curve theorem, it is easy to verify that $A^a \cap A^b = \varnothing$. On the event $\mathcal{Y}$, there is no other choice for the strand of γ extending $\mathrm{ext}_{R'}\gamma$ at a, but to be confined to A^a until it hooks up with β at $\hat{a}$, since every other exit from A^a is blocked. Consequently, on $\mathcal{Y}$, we have $\gamma \supset \beta$. A similar argument applies to A^b, and we get

$$\beta \subset \gamma \subset \mathrm{ext}_{R'}\gamma \cup \beta \cup A^a \cup A^b \quad \text{on } \mathcal{Y}.$$

We now turn to the random walk S. For $\mathcal{Z}_0^\sigma$ to hold, we must make sure that $\{S_t : t < \tau_0\}$ does not meet any vertex neighboring with γ. First consider $\mathrm{ext}_{R'}S$. Note that $\mathrm{ext}_{R'}S$ does not intersect ∂A^a, because ∂A^a is contained in $\mathfrak{B}_{R'} \cup B(a, 2\varepsilon' R)$, and we

are assuming that $Q(\Phi_{R'})>\varepsilon$. (Recall that $\{a,b\}=\{x^{R'},y^{R'}\}$.) Thus, $\mathrm{ext}_{R'}S\cap A^a=\varnothing$, and we may also conclude that $\mathrm{ext}_{R'}S$ does not visit any vertex adjacent to A^a when R is large. Similar arguments apply to A^b and to $\mathrm{int}_R\breve{S}$. Thus, $\mathrm{int}_R\breve{S}\cup\mathrm{ext}_{R'}S$ does not visit any vertex adjacent to γ on the event $\mathcal{Y}\cap\{Q(\Phi_{R'})>\varepsilon, Q(\Theta_R)>\varepsilon\}$.

Now let S^* be the walk S from the first time it visits $q^{R'}$ until the first time it visits $\hat{q}^R$. Then, conditioned on Θ_R and $\Phi_{R'}$, S^* is just a simple random walk started at $q^{R'}$ conditioned to hit $\hat{q}^R$ before hitting 0. Let A^q denote the $\varepsilon' R$-neighborhood of α_q. Clearly, $\mathrm{dist}(A^q, A^a\cup A^b)\geqslant\varepsilon' R$. The probability that S^* gets within distance $\frac{1}{4}\varepsilon' R$ of $\hat{q}^R$ before exiting A^q is at least some (perhaps small) positive constant depending only on ε' (and hence on ε). Conditional on this event, the probability that S^* visits $\hat{q}^R$ before exiting A^q is within a constant multiple of $1/\log(\varepsilon' R)$, by (2.6). Now let S^{**} be the walk S from the first visit of $\hat{q}^R$ to the last visit of $\hat{q}^R$ before time τ_0. Note that S^{**} and $\mathrm{int}_R\breve{S}$ are independent given $\hat{q}^R$. Thus, given $\Phi_{R'}$, Θ_R, $\mathcal{D}$ and S^*, we may sample S^{**} by starting a random walk from $\hat{q}^R$, stopping when it hits 0, and then removing the part of that walk after the last visit to $\hat{q}^R$. When the latter walk first gets to distance $\frac{1}{2}\varepsilon' R$ from $\hat{q}^R$, it has probability bounded away from zero (by a constant depending only on ε') to hit 0 before $\hat{q}^R$. (This follows, for example, from Lemma 2.2 applied to the function giving for every vertex the probability to hit 0 before $\hat{q}^R$ for a random walk started at that vertex.) Thus, conditioned on $(S^*,\Theta_R,\Phi_{R'})$, with probability bounded away from zero, $S^{**}\subset A^q$. Since

$$\mathcal{Z}_0^\sigma\supset\{S^*\cup S^{**}\subset A^q\}\cap\mathcal{Y},$$

we conclude that

$$O_{\varepsilon,\bar{\Lambda}}(1)\mathbf{P}[\mathcal{Z}_0^\sigma\mid\Phi_{R'},\Theta_R,\mathcal{D}]\geqslant\frac{1}{\log R}.$$

Above, we have argued that $\mathbf{P}[\mathcal{D}|\Phi_{R'},\Theta_R]$ is bounded away from zero, and so we conclude that the lower bound estimate in the proposition holds.

It remains to prove the upper bound. Conditional on γ, $\mathrm{int}_R\breve{S}$ and $\mathrm{ext}_{R'}S$, the probability that S^* (as defined in the proof of the lower bound) hits $\hat{q}^R$ before hitting γ is clearly $O(1)/\log R$, since the conditional law of S^* is that of a random walk started at $q^{R'}$ and conditioned to hit $\hat{q}^R$ before 0, and the probability that an ordinary random walk started at $q^{R'}$ hits $\hat{q}^R$ before 0 is bounded away from zero. The upper bound now follows, and the proof is complete. □

To make the previous lemma useful, we will need to argue that configurations with quality bigger than ε are not too rare, in an appropriate sense. This is achieved by the following lemma.

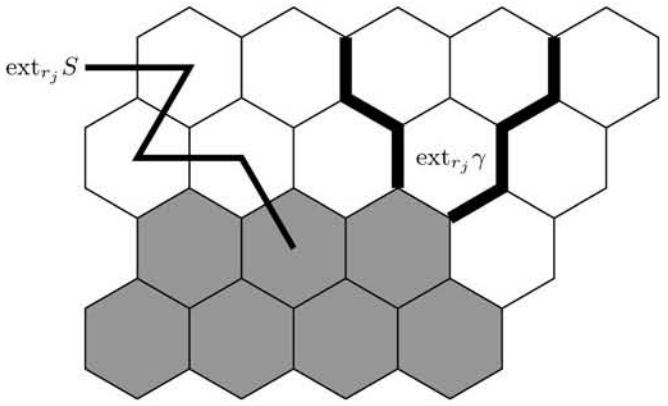


Figure 3.7. An example for $Q_j{=}0{\neq}Q(\Phi_{r_j})$. The set $\mathfrak{B}_{r_j}$ is shaded.

LEMMA 3.15. (Separation) *Assume* (h), (D) *and* (S). *Let* $p{<}1$. *There exists some constant* $c{=}c(p,\bar{\Lambda}){>}0$ *such that if* $R{>}1/c$ *and* $v_0{\notin}\mathfrak{B}_{6R}{\subset}D$, *then*

$$\mathbf{P}[Q(\Phi_{3R}){\wedge}Q(\Theta_{2R}){>}c\,|\,\Theta_R,\Phi_{4R},\mathcal{Z}_0^\sigma]{>}p, \tag{3.34}$$

provided that $\mathbf{P}[\mathcal{Z}_0^\sigma\,|\Theta_R,\Phi_{4R}]{>}0$.

The proof of this lemma is modeled after Lawler's separation lemma for Brownian motions from [L2, Lemma 4.2].

Proof. To keep the notation simple, we start by proving a simpler version of the lemma, where we also assume that $Q(\Theta_R){\geqslant}\frac{1}{100}$, say (and therefore $Q(\Theta_R){=}\frac{1}{100}$), and we prove that

$$\mathbf{P}[Q(\Phi_{3R}){>}c\,|\,\Theta_R,\Phi_{4R},\mathcal{Z}_0^\sigma]{>}p. \tag{3.35}$$

We define inductively a random sequence $r_0,r_1,...$ as follows. Set $r_0{:=}4R$. Suppose that r_j is defined. Set

$$Q_j := \begin{cases} Q(\Phi_{r_j}), & \text{if } \mathbf{P}[\mathcal{Z}_0^\sigma\,|\,\Phi_{r_j},\Theta_R]{>}0,\\ 0, & \text{otherwise,}\end{cases}$$

and $r_{j+1}{:=}(1{-}10Q_j)r_j$. Note that $Q_j{=}0$ implies that $\mathbf{P}[\mathcal{Z}_0^\sigma\,|\Phi_{r_j},\Theta_R]{=}0$. (An example showing that $Q_j{=}0{\neq}Q(\Phi_{r_j})$ is possible is given in Figure 3.7. Such a situation can only occur when $|x^{r_j}{-}y^{r_j}|{=}O(1)$.) Also note that $\mathbf{P}[\mathcal{Z}_0^\sigma\,|\Phi_{r_j},\Theta_R]{>}0$ if and only if there are paths γ_* and S_* satisfying the following: (1) γ_* is a simple TG*-path in $\overline{D}$ containing β^R and $\mathrm{ext}_{r_j}\gamma$, (2) S_* is a TG-path in D containing $\mathrm{ext}_{r_j}S$ and the reversal of $\mathrm{int}_R\breve{S}$ and (3) S_* does not visit any vertex in $\partial D(\gamma_*)$, except for 0.

We claim that for every $j{\in}\mathbb{N}$,

$$\mathbf{P}\big[Q_{j+1}{\geqslant}(2Q_j){\wedge}\tfrac{1}{100}\,\big|\,\Phi_{r_j},r_j{>}2R,\Theta_R\big]{>}c_0 \tag{3.36}$$

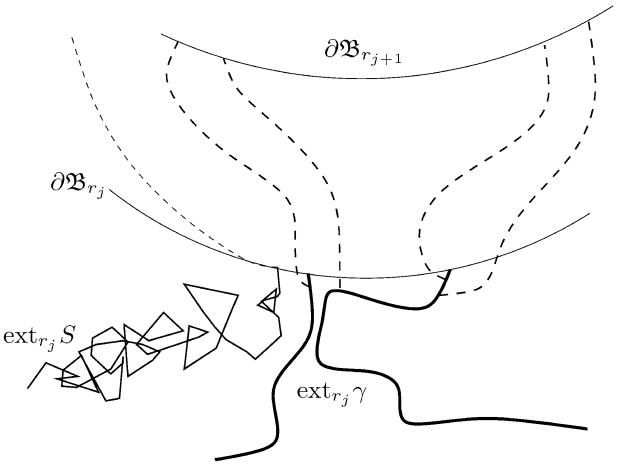


Figure 3.8. The construction of the barriers giving (3.36).

for some constant $c_0=c_0(\bar{\Lambda})>0$. Clearly it suffices to prove this in the case $Q_j>0$. The gap between $\partial\mathfrak{B}_{r_j}$ and $\partial\mathfrak{B}_{r_{j+1}}$ is larger than but comparable to $5r_jQ_j$. Note also that because $\mathfrak{B}_{r_j}$ is a union of TG*-hexagons, $r_jQ_j\geqslant 1/\sqrt{3}$. If $r_jQ_j>20$, say, then, we can easily use barriers as in the proof of Lemma 3.14 (see Figure 3.8) to direct and separate the two strands of $\mathrm{ext}_{r_{j+1}}\gamma\backslash\mathrm{ext}_{r_j}\gamma$ and the walk $\mathrm{ext}_{r_{j+1}}S\backslash\mathrm{ext}_{r_j}S$ so as to obtain (3.36).

Now assume that $r_jQ_j\leqslant 20$. In this case the discrete structure of the lattice is "visible". Let q' be the point on $\partial\mathfrak{B}_{r_j}$ crossed by $\mathrm{ext}_{r_j}S$ in its last step, and let α be a longest arc among the three connected components of $\partial\mathfrak{B}_{r_j}\setminus\{q',x^{r_j},y^{r_j}\}$. Suppose first that q' is not an endpoint of α. Let α be oriented counterclockwise around $\mathfrak{B}_{r_j}$, and let a and b be the initial and terminal points of α, respectively. Let η_a (respectively, η_b) denote the connected component of $\mathrm{ext}_{r_{j+1}}\gamma\backslash\mathrm{ext}_{r_j}\gamma$ that has a (respectively, b) as an endpoint. Assume that $|q'-a|<500$ and $|q'-b|<500$. Consider the event $\mathcal{X}$ that η_a goes as far to the right as possible subject to the conditions that it remains inside $B(q',550)$ and avoids $\mathrm{ext}_{r_j}\gamma$ and that η_b goes as far to the left as possible subject to the conditions that it remains inside $B(q',550)$ and avoids $\mathrm{ext}_{r_j}\gamma$. See Figure 3.9. It is easy to see that $\mathbf{P}[\mathcal{Z}_0^\sigma\,|\,\Phi_{r_j},\Theta_R]>0$ implies that on $\mathcal{X}$ there is a simple TG-path in $B(q',250)$ from q' to $\mathfrak{B}_{r_{j+1}}$ that avoids $\partial D(\mathrm{ext}_{r_{j+1}}\gamma)$. Since the number of edges traversed by these paths is bounded, it is easy to see that the probability that S follows the latter path and $\mathcal{X}$ holds given Φ_{r_j} and Θ_R satisfying the above assumptions is bounded away from zero (for $\mathcal{X}$, note that we can extend the interfaces one step at a time and the probability for every specific step given the previous ones is bounded away from zero). This gives (3.36) in this case. Similar, or simpler, arguments apply if one or more of the assumptions $|q'-a|<500$ and $|q'-b|<500$ do not hold.

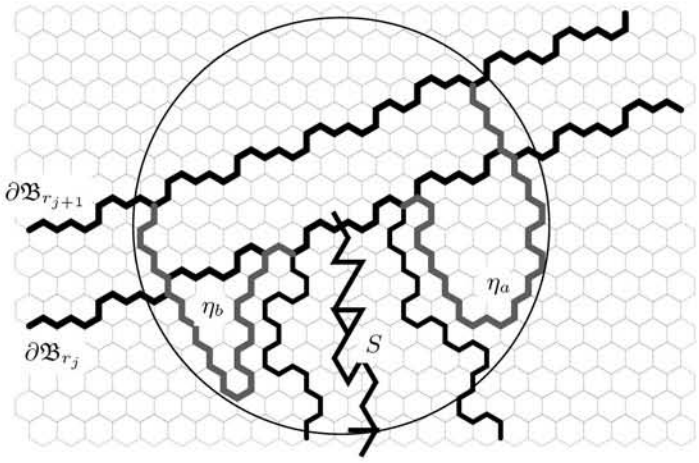


Figure 3.9. The interfaces spreading out.

If q' is an endpoint of α a similar argument may be used. Suppose, for example, that q' is the initial point of α, that y^{r_j} is the other endpoint of α and that $|q'-y^{r_j}|<300$. Then we may consider the possibility that the connected component of $\mathrm{ext}_{r_{j+1}}\gamma\backslash\mathrm{ext}_{r_j}\gamma$ that has y^{r_j} as an endpoint goes as far to the left as possible subject to the requirements that it stays inside $B(y^{r_j},5000)$ and avoids hexagons containing vertices visited by $\mathrm{ext}_{r_j}S$, and that the connected component of $\mathrm{ext}_{r_{j+1}}\gamma\backslash\mathrm{ext}_{r_j}\gamma$ that has x^{r_j} as an endpoint goes as far to the left as possible subject to the requirements that it stays inside $B(x^{r_j},4000)$ and avoids the previous strand extending $\mathrm{ext}_{r_j}\gamma$ at y^{r_j} and finally, the random walk avoids $\partial D(\mathrm{ext}_{r_{j+1}}\gamma)$ and stays in $B(q',300)$ until it hits $\mathfrak{B}_{r_{j+1}}$. A similar argument applies if $|q'-y^{r_j}|\geqslant 300$. This proves (3.36).

We now prove that for every $j\in\mathbb{N}$,

$$\mathbf{P}[Q_{j+2}=0\,|\,\Phi_{r_j},r_j>2R,\Theta_R,Q_{j+1}<2Q_j]\geqslant c_1, \tag{3.37}$$

for some $c_1=c_1(\bar{\Lambda})>0$. Let $x=x^{r_{j+1}}$, $y=y^{r_{j+1}}$ and $q=q^{r_{j+1}}$. Let $S^\dagger$ be the part of the walk S from the first visit to q^{r_j} up to the first visit to q. Note that

$$\mathrm{dist}(x,\mathrm{ext}_{r_j}S)\geqslant 2r_jQ_j>2r_{j+1}Q_j,$$

if $Q_j>0$, and similarly for y. Thus, the event $Q_{j+1}<2Q_j$ is the union of the following five events

$$\begin{aligned}
&\mathcal{M}_0:=\{Q_{j+1}=0\},\\
&\mathcal{M}_1:=\{\mathrm{dist}(S^\dagger,\{x,y\})<(2Q_jr_{j+1})\wedge|x-y|\},\\
&\mathcal{M}_2:=\{r_{j+1}Q_{j+1}=\mathrm{dist}(q,\mathrm{ext}_{r_{j+1}}\gamma),Q_{j+1}<2Q_j\},
\end{aligned}$$

$$\mathcal{M}_3 := \{0 < |x-y| = r_{j+1}Q_{j+1}, Q_{j+1} < 2Q_j\},$$
$$\mathcal{M}_4 := \{Q_{j+1} = \tfrac{1}{100}, Q_{j+1} < 2Q_j\}.$$

(In the definition of $\mathcal{M}_1$, $\operatorname{dist}(S^\dagger, \{x,y\})$ means the least distance from a vertex visited by $S^\dagger$ to x or y, of course.) Clearly,

$$O_{\bar{\Lambda}}(1)\mathbf{P}[Q_{j+2}=0 \,|\, \Phi_{r_j}, \Theta_R, \mathcal{M}_k] \geqslant 1 \tag{3.38}$$

holds for $k=0$. The same is also true for $k=4$, because $Q_{j+1}=\frac{1}{100}$ implies that the random walk started at q has conditional probability bounded away from zero to hit $\operatorname{ext}_{r_{j+1}}\gamma$ before $\partial\mathfrak{B}_{r_{j+2}}$. A similar argument gives (3.38) when $k=2$.

Now condition on $\mathcal{M}_1$, and let v be the vertex first visited by $S^\dagger$ that is at distance less than $(2Q_jr_{j+1})\wedge|x-y|$ from $\{x,y\}$. Conditioned additionally on Θ_R, $\operatorname{ext}_{r_{j+1}}\gamma$ and the walk $S^\dagger$ until it hits v, there is clearly probability bounded away from zero that $S^\dagger$ hits a vertex adjacent to $\operatorname{ext}_{r_{j+1}}\gamma$ before $\mathfrak{B}_{r_{j+2}}$, and in this case we have $Q_{j+2}=0$. Consequently, (3.38) also holds for $k=1$.

Now condition on $\mathcal{M}_3$, $\Phi_{r_{j+1}}$ and Θ_R. Let z be the midpoint of the segment $[x,y]$, and consider the circle $\partial B(z, 2|x-y|)$. We may build a barrier by using the connected component of $\partial B(z,2|x-y|)\setminus\partial D(\operatorname{ext}_{r_{j+1}}\gamma)$ that intersects $\mathfrak{B}_{r_{j+1}}$ (and possibly perturbing it slightly near its endpoints). If γ does not cross this barrier, then $Q_{j+2}=0$ holds. Thus, we get from Theorem 3.11 that (3.38) also holds for $k=3$. Since $\{Q_{j+1}<2Q_j\}=\bigcup_{k=0}^4\mathcal{M}_k$, and (3.38) holds for $k=0,1,...,4$, it follows that (3.37) holds as well.

Set $s_n := 2R\prod_{k=0}^{n-1}(1-2^{-k}/10)^{-1}$. It follows from Lemma 3.14 and our assumption that $Q(\Theta_R)\geqslant\frac{1}{100}$ that for any $j\in\mathbb{N}$,

$$\mathbf{P}[\mathcal{Z}_0^\sigma \,|\, \Phi_{r_j}, \Theta_R] \geqslant \frac{c_2}{\log R}1_{\{Q_j\geqslant 1/100\}}1_{\{r_j\geqslant s_0\}}$$

for some $c_2=c_2(\bar{\Lambda})>0$. An appeal to (3.36) therefore implies that

$$\mathbf{P}[\mathcal{Z}_0^\sigma \,|\, \Phi_{r_j}, \Theta_R] \geqslant \frac{c_0c_2}{\log R}1_{\{Q_j\geqslant 2^{-1}/100\}}1_{\{r_j\geqslant s_1\}}.$$

Continuing inductively, we get for every $n\in\mathbb{N}$,

$$\mathbf{P}[\mathcal{Z}_0^\sigma \,|\, \Phi_{r_j}, \Theta_R] \geqslant \frac{c_0^nc_2}{\log R}1_{\{Q_j\geqslant 2^{-n}/100\}}1_{\{r_j\geqslant s_n\}}.$$

Since

$$\sup_n s_n < 2R\left(1-\sum_{k=0}^\infty\frac{2^{-k}}{10}\right)^{-1} = \frac{5}{2}R,$$

we get for every $j\in\mathbb{N}$,

$$\mathbf{P}[\mathcal{Z}_0^\sigma \mid \Phi_{r_j}, \Theta_R] \geqslant \frac{c_0^{-\log_2 Q_j} c_2}{\log R} 1_{\{r_j \geqslant 5R/2\}} = c_2 \frac{Q_j^{-\log_2 c_0}}{\log R} 1_{\{r_j \geqslant 5R/2\}}. \tag{3.39}$$

From (3.37) we get that for every j the conditional probability that there is some $k>j$ such that $r_k>2R$, $Q_k<2Q_j$ and $Q_{k+1}>0$, given Φ_{r_j} and Θ_R, is at most $1-c_1$. Let m_n denote the number of $k\in\mathbb{N}$ such that $r_k>3R$ and $Q_k\in(2^{-n}, 2^{1-n}]$. Fix some $n\in\mathbb{N}$, and suppose that $\mathbf{P}[m_n>0|\Phi_{4R}, \Theta_R]>0$. On the event $m_n>0$, let k_n be the first k such that $Q_k\in(2^{-n}, 2^{1-n}]$. By induction and (3.37), for every $m\in\mathbb{N}$,

$$\mathbf{P}[m_n > 2m \,|\, m_n > 0, \Phi_{r_{k_n}}, \Theta_R] \leqslant (1-c_1)^m.$$

An appeal to the upper bound in Lemma 3.14 gives

$$\mathbf{P}[\mathcal{Z}_0^\sigma, m_n > 2m \,|\, m_n > 0, \Phi_{r_{k_n}}, \Theta_R] \leqslant \frac{c_3(1-c_1)^m}{\log R}$$

for some $c_3=c_3(\bar{\Lambda})$. On the other hand, (3.39) gives

$$\mathbf{P}[\mathcal{Z}_0^\sigma \,|\, m_n > 0, \Phi_{r_{k_n}}, \Theta_R] \geqslant \frac{c_2 c_0^n}{\log R}.$$

Comparing the last two inequalities, we get

$$\mathbf{P}[m_n > 2m \,|\, \mathcal{Z}_0^\sigma, \Phi_{r_0}, \Theta_R] \leqslant \frac{c_3 c_0^{-n}(1-c_1)^m}{c_2}.$$

In particular, there is an $n_0=n_0(\bar{\Lambda}, p)\in\mathbb{N}$ and a $c_4=c_4(\bar{\Lambda})$ such that

$$\mathbf{P}[\text{there exists } n \geqslant n_0 : m_n > c_4 n \,|\, \mathcal{Z}_0^\sigma, \Phi_{r_0}, \Theta_R] \leqslant \tfrac{1}{3}(1-p).$$

Let n_1 be the least integer larger than 3 such that $\prod_{n=n_1}^\infty (1-10\cdot 2^{1-n})^{c_4 n} > \frac{7}{8}$, and let $n_2=n_1\vee n_0$. On the event $\mathcal{Z}_0^\sigma \cap \bigcap_{n>n_2}\{m_n \leqslant c_4 n\}$ we must have some $j\in\mathbb{N}$ with $r_j>\frac{7}{2}R$ and $Q_j\geqslant 2^{-n_2}$ (because $r_{j+1}/r_j=1-10Q_j$ and $n_2\geqslant n_1$). Consequently,

$$\mathbf{P}\big[\text{there exists } j : Q_j > c_5, r_j > \tfrac{7}{2}R \,\big|\, \mathcal{Z}_0^\sigma, \Phi_{4R}, \Theta_R\big] \geqslant 1-\tfrac{1}{3}(1-p) \tag{3.40}$$

holds for some $c_5=c_5(p, \bar{\Lambda})>0$. Note that this almost achieves our goal of proving (3.35). The difference between (3.40) and (3.35) is that in the latter the radius r at which $Q(\Phi_r)$ is bounded from below is variable.

Let $\mathcal{A}$ be the event that there is a $j\in\mathbb{N}$ with $Q_j>c_5$ and $r_j>\frac{7}{2}R$, and on $\mathcal{A}$, let j_0 denote the first such j. We have from (3.39) that

$$\mathbf{P}[\mathcal{Z}_0^\sigma \,|\, \mathcal{A}, \Phi_{r_{j_0}}, \Theta_R] \geqslant \frac{c_2 c_5^{-\log_2 c_0}}{\log R}. \tag{3.41}$$

Fix some $s>0$ small. We now argue that

$$\mathbf{P}[\mathcal{Z}_0^\sigma \mid \mathcal{A}, \Phi_{r_{j_0}}, \Theta_R, |x^{3R}-y^{3R}|<sR] \leqslant \frac{c_7 s^{c_6}}{\log R} \tag{3.42}$$

for some positive constants c_7 and c_6 depending only on $\bar{\Lambda}$. The argument is similar to the one given in the proof of the case $k=3$ in (3.38). Let z be the midpoint of the segment $[x^{3R}, y^{3R}]$. We construct a barrier as a perturbation of the connected component of $\partial B(z, 2sR)\backslash\partial D(\mathrm{ext}_{3R}\gamma)$ that intersects $\mathfrak{B}_{3R}$. If that barrier is hit by the extension of $\mathrm{ext}_{3R}\gamma$ (which happens with probability bounded away from 1), then we condition on the extension up to that barrier, and construct another barrier at radius $4sR$, instead. We continue in this manner, constructing barriers at radii $2^n sR$ up to the least n such that $2^n s > \frac{1}{1000}$, say. Because the probability of avoiding the nth barrier given that the $(n-1)$th barrier has been reached is bounded away from 1, we find that $\mathbf{P}[\gamma\supset\beta_R|\mathcal{A}, \Phi_{r_{j_0}}, \Theta_R, |x^{3R}-y^{3R}|<sR]$ is bounded by a constant times some positive power of s. The estimate (3.42) follows by considering the behavior of S.

Suppose now that the random walk S after its first hit to $\mathfrak{B}_{r_{j_0}}$ but before its first hit to $\mathfrak{B}_{3R}$ gets within distance sR of $\mathrm{ext}_{3R}\gamma$. Then, by Lemma 2.1, conditional on S up to the first time this has happened and on $\mathcal{A}$, $\Phi_{r_{j_0}}$, Θ_R and $\mathrm{ext}_{3R}\gamma$, the conditional probability for S hitting $\mathfrak{B}_{5R/2}$ before hitting $\partial D(\mathrm{ext}_{3R}\gamma)$ is at most $c_8 s^{\zeta_1}$, for some universal constant c_8. Thus, the conditional probability for $\mathcal{Z}_0^\sigma$ is at most $c_8 s^{\zeta_1}/\log R$. Combining this with (3.42), one gets

$$\mathbf{P}[\mathcal{Z}_0^\sigma \mid \mathcal{A}, \Phi_{r_{j_0}}, \Theta_R, Q(\Phi_{3R})<s] \leqslant \frac{c_7 s^{c_6}+c_8 s^{\zeta_1}}{\log R}.$$

Comparison with (3.41) now gives

$$\begin{aligned}\mathbf{P}[Q(\Phi_{3R})<s \mid \mathcal{Z}_0^\sigma, \mathcal{A}, \Phi_{r_{j_0}}, \Theta_R] &= \frac{\mathbf{P}[Q(\Phi_{3R})<s, \mathcal{Z}_0^\sigma \mid \mathcal{A}, \Phi_{r_{j_0}}, \Theta_R]}{\mathbf{P}[\mathcal{Z}_0^\sigma \mid \mathcal{A}, \Phi_{r_{j_0}}, \Theta_R]} \\ &\leqslant \frac{\mathbf{P}[\mathcal{Z}_0^\sigma \mid \mathcal{A}, \Phi_{r_{j_0}}, \Theta_R, Q(\Phi_{3R})<s]}{\mathbf{P}[\mathcal{Z}_0^\sigma \mid \mathcal{A}, \Phi_{r_{j_0}}, \Theta_R]} \leqslant \frac{c_7 s^{c_6}+c_8 s^{\zeta_1}}{c_2 c_5^{-\log_2 c_0}}.\end{aligned}$$

Thus, we obtain, for all s sufficiently small,

$$\mathbf{P}[Q(\Phi_{3R})<s \mid \mathcal{Z}_0^\sigma, \mathcal{A}, \Phi_{4R}, \Theta_R] \leqslant \tfrac{1}{3}(1-p).$$

Taking (3.40) into account, this gives (3.35), and completes the proof of the simplified case.

The argument in the general case proceeds as follows. We define inductively two sequences r_j and $\hat{r}_j$, starting with $r_0=4R$ and $\hat{r}_0=R$. At each step j, we set

$$Q_j := Q(\Phi_{r_j}) \wedge Q(\Theta_{\hat{r}_j}) \wedge 1_{\{\mathbf{P}[\mathcal{Z}_0^\sigma|\Phi_{r_j},\Theta_{\hat{r}_j}]>0\}}$$

and take $r_{j+1}=(1-10Q_j)r_j$ and $\hat{r}_{j+1}=(1+10Q_j)\hat{r}_j$. The proof proceeds essentially as above. The straightforward details are left to the reader. □

COROLLARY 3.16. *There exists a constant $c=c(\bar{\Lambda})>0$ such that the following estimate holds. Let $(D', \partial'_+, \partial'_-, h'_\partial, h', \gamma', v'_0)$ satisfy the same assumptions as we have for $(D, \partial_+, \partial_-, h_\partial, h, \gamma, v_0)$. Let $R>c$, and assume that $\mathfrak{B}_{6R}\subset D'\cap D$ and $v_0, v'_0\notin\mathfrak{B}_{6R}$. Let Θ'_R, Φ'_{4R} and $\mathcal{Z}_0^{\sigma\prime}$ be the objects corresponding to Θ_R, Φ_{4R} and $\mathcal{Z}_0^\sigma$ for the system in D' (with the same σ, that is, $\sigma'=\sigma$). Then*

$$\mathbf{P}[\Theta_R=\vartheta\,|\,\mathcal{Z}_0^\sigma, \Phi_{4R}]\leqslant c\mathbf{P}[\Theta'_R=\vartheta\,|\,\mathcal{Z}_0^{\sigma\prime}, \Phi'_{4R}] \tag{3.43}$$

holds for all ϑ and for all Φ_{4R} and Φ'_{4R} satisfying $\mathbf{P}[\mathcal{Z}_0^\sigma\,|\,\Phi_{4R}]>0$ and $\mathbf{P}[\mathcal{Z}_0^{\sigma\prime}\,|\,\Phi'_{4R}]>0$, respectively. Consequently, under the same assumptions, there exists a coupling of the conditional laws of Θ_R and Θ'_R such that

$$\mathbf{P}[\Theta_R=\Theta'_R\,|\,\mathcal{Z}_0^\sigma, \mathcal{Z}_0^{\sigma\prime}, \Phi_{4R}, \Phi'_{4R}]\geqslant\frac{1}{c}.$$

Proof. It is enough to prove the first claim, since the latter claim immediately follows. Let $c'>0$ be the constant denoted as c in the separation lemma (Lemma 3.15) with $p=\frac{1}{2}$. Let $\mathcal{Q}$ denote the event $Q(\Phi_{3R})\wedge Q(\Theta_{2R})\geqslant c'$. Let X be the collection of all θ such that $\Theta_{2R}=\theta$ is possible and $Q(\theta)\geqslant c'$, and let X_ϑ be the collection of all $\theta\in X$ that are compatible with $\Theta_R=\vartheta$; that is, such that $\{\Theta_{2R}=\theta$ and $\Theta_R=\vartheta\}$ is possible. In the following, $f\approx g$ will mean that f/g is contained in $[1/c, c]$ for some constant $c=c(\bar{\Lambda})>0$. By Lemma 3.15 and the choice of p,

$$\mathbf{P}[\Theta_R=\vartheta\,|\,\mathcal{Z}_0^\sigma, \Phi_{4R}]\approx\mathbf{P}[\Theta_R=\vartheta\,|\,\mathcal{Q}, \mathcal{Z}_0^\sigma, \Phi_{4R}]=\sum_{\theta\in X_\vartheta}\mathbf{P}[\Theta_{2R}=\theta\,|\,\mathcal{Q}, \mathcal{Z}_0^\sigma, \Phi_{4R}]. \tag{3.44}$$

Now, if $Q(\Phi_{3R})>c'$ and $\theta\in X$, then

$$\begin{aligned}\mathbf{P}[\Theta_{2R}=\theta\,|\,\mathcal{Q}, \mathcal{Z}_0^\sigma, \Phi_{3R}]&=\frac{\mathbf{P}[\Theta_{2R}=\theta, \mathcal{Q}, \mathcal{Z}_0^\sigma\,|\,\Phi_{3R}]}{\mathbf{P}[\mathcal{Z}_0^\sigma, \mathcal{Q}\,|\,\Phi_{3R}]}\\&=\frac{\mathbf{P}[\Theta_{2R}=\theta, \mathcal{Z}_0^\sigma\,|\,\Phi_{3R}]}{\mathbf{P}[\mathcal{Z}_0^\sigma, \mathcal{Q}\,|\,\Phi_{3R}]}\\&=\frac{\mathbf{P}[\Theta_{2R}=\theta\,|\,\Phi_{3R}]\mathbf{P}[\mathcal{Z}_0^\sigma\,|\,\Theta_{2R}=\theta, \Phi_{3R}]}{\mathbf{P}[\mathcal{Z}_0^\sigma, \mathcal{Q}\,|\,\Phi_{3R}]}.\end{aligned}$$

We apply Lemma 3.13 to the first factor in the numerator and Lemma 3.14 to the second factor, and get

$$\mathbf{P}[\Theta_{2R}=\theta\,|\,\mathcal{Q}, \mathcal{Z}_0^\sigma, \Phi_{3R}]\approx\frac{p_{2R}(\theta)/\log R}{\mathbf{P}[\mathcal{Z}_0^\sigma, \mathcal{Q}\,|\,\Phi_{3R}]}.$$

The sum of the left-hand side over all $\theta\in X$ is 1. Consequently,

$$\mathbf{P}[\Theta_{2R}=\theta_0\,|\,\mathcal{Q}, \mathcal{Z}_0^\sigma, \Phi_{3R}]\approx\frac{p_{2R}(\theta_0)}{\sum_{\theta\in X}p_{2R}(\theta)}.$$

By taking expectation conditioned on $\mathcal{Q}$, $\mathcal{Z}_0^\sigma$ and Φ_{4R}, it follows that the same relation holds when we replace Φ_{3R} by Φ_{4R}. We now sum over $\theta_0 \in X_\vartheta$ and invoke (3.44), to obtain

$$\mathbf{P}[\Theta_R = \vartheta \mid \mathcal{Z}_0^\sigma, \Phi_{4R}] \approx \frac{\sum_{\theta \in X_\vartheta} p_{2R}(\theta)}{\sum_{\theta \in X} p_{2R}(\theta)}.$$

This implies (3.43), and completes the proof. □

Our intermediate goal to show that the dependence between the local behavior near S_τ and the global behavior far away is now accomplished. Roughly, the next objective will be to show that it is unlikely that γ contains an arc with a very large diameter whose endpoints are both relatively close to S_τ.

For $R > r > 0$, let $\mathcal{J} = \mathcal{J}(r, R)$ denote the event that there are more than two disjoint arcs of γ connecting $\mathfrak{B}_r$ and $\partial\mathfrak{B}_R$ or that S exits $\mathfrak{B}_R$ between the time it first hits $\mathfrak{B}_r$ and τ_0. Set $\mathcal{Z}_0 := \bigcup_{\sigma=0}^5 \mathcal{Z}_0^\sigma$. Our next objective is to show that conditioned on $\mathcal{Z}_0^\sigma$ or $\mathcal{Z}_0$, $\mathcal{J}(r, R)$ is unlikely if $R \gg r > 0$. More precisely, the claim is as follows.

LEMMA 3.17. *Assume* (h), (D) *and* (S). *For every* $p > 0$ *there is some* $a = a(p, \bar{\Lambda}) > 10$ *such that if* $r > 1$, $R > ar$, $v_0 \notin \mathfrak{B}_{4R} \subset D$ *and* $\mathbf{P}[\mathcal{Z}_0^\sigma \mid \Phi_R, \Theta_r] > 0$, *then*

$$\mathbf{P}[\mathcal{J}(r, R) \mid \Phi_R, \Theta_r, \mathcal{Z}_0^\sigma] < p$$

and also

$$\mathbf{P}[\mathcal{J}(r, R) \mid \Phi_R, \mathcal{Z}_0] < p.$$

One may first think that this can be proved by repeating the argument in the proof of Lemma 3.14. The difficulty in carrying out this idea is that the sets A^a and A^b described in the proof of Lemma 3.14 may extend beyond $\mathfrak{B}_{sR'}$ for large s (where R' is as in that lemma) if there are more than two disjoint arcs of $\mathrm{ext}_{R'}\gamma$ connecting $\partial\mathfrak{B}_{R'+\varepsilon' R}$ and $\partial\mathfrak{B}_{sR'}$.

Since the proof of the lemma is a bit involved and somewhat indirect, we take a few moments to give an overview of the strategy. First, it is established that under the conditioning the simple random walk S is unlikely to backtrack to $\partial\mathfrak{B}_R$ after hitting $\mathfrak{B}_r$ and before τ_0. Next, we identify a pair of arcs α_1 and α_2 that are defined from Φ_{3r} each of which has one endpoint on $\mathrm{ext}_{3r}S$ and the other on $\mathrm{ext}_{3r}\gamma$. A barriers argument is then used to show that with high conditional probability $\gamma \backslash \mathrm{ext}_{3r}\gamma$ does not hit $\alpha_1 \cup \alpha_2$. In this case, we see that $\alpha_1 \cup \alpha_2$ has an alternative definition in terms of γ and S, which is in some sense more symmetric. Next we define another pair of arcs $\widetilde{\alpha}_1 \cup \widetilde{\alpha}_2$, which have a similar definition as $\alpha_1 \cup \alpha_2$, except that they are defined from $\Theta_{R/3}$. Again, the same barriers argument can be used to show that with high conditional probability these arcs are not visited by $\gamma \backslash \beta_{R/3}$. In this case, these arcs have a more symmetric definition,

which leads us to conclude that with high conditional probability $\alpha_1 \cup \alpha_2 = \widetilde{\alpha}_1 \cup \widetilde{\alpha}_2$. This is then used to establish that the endpoints of these arcs on γ belong to $\mathrm{ext}_{3r}\gamma$ as well as $\beta_{R/3}$. Next, we prove that these two endpoints belong to different connected components of $\gamma \backslash e^*_\sigma$. This then implies that each of the two strands of $\mathrm{ext}_{3r}\gamma$ merges with $\beta_{R/3}$, which implies that there are no more than two disjoint crossings between $\partial\mathfrak{B}_R$ and $\mathfrak{B}_{3r}$ in γ.

Proof. The second claimed inequality with p replaced by $6p$ follows from the first inequality and taking conditional expectation, since $\mathcal{Z}_0 = \bigcup_{\sigma=0}^5 \mathcal{Z}_0^\sigma$. Thus, we only need to prove the first inequality. By Lemma 3.15, there is a constant $c_0 = c_0(\bar{\Lambda}, p) > 0$ such that

$$\mathbf{P}[Q(\Phi_{3R/4}) \wedge Q(\Theta_{R/2}) \wedge Q(\Phi_{3r}) \wedge Q(\Theta_{2r}) \leqslant c_0 \,|\, \Phi_R, \Theta_r, \mathcal{Z}_0^\sigma] < \tfrac{1}{10}p. \tag{3.45}$$

Let r' be in the range $\big[\sqrt{rR}, \sqrt{rR}+1\big]$, chosen so that the circle $\partial B(0, r')$ does not contain any TG-vertices nor any TG*-vertices. Let S_* denote the part of the walk S from its first visit to q^{3r} until its last visit to $\hat{q}^r$ prior to τ_0. Conditional on Φ_{3r} and Θ_r, the probability that S_* exits $\mathfrak{B}_{r'/3}$ without hitting $\mathrm{ext}_{3r}\gamma$ decays to zero as $a \to \infty$ (by Lemma 2.1). On the event that this happens, let S_{**} be the initial segment of the walk S_* until it exits $\mathfrak{B}_{r'/3}$. By the proof of the upper bound in Lemma 3.14,

$$\mathbf{P}[\mathcal{Z}_0^\sigma \,|\, \Theta_{2r}, \Phi_{3r}, S_* \not\subset \mathfrak{B}_{r'/3}, S_{**}] \leqslant \frac{O(1)}{\log r}.$$

By the lower bound in that lemma, on the event $Q(\Phi_{3r}) \wedge Q(\Theta_{2r}) \geqslant c_0$ we have

$$O_{p,\bar{\Lambda}}(1)\mathbf{P}[\mathcal{Z}_0^\sigma \,|\, \Theta_{2r}, \Phi_{3r}] \geqslant \frac{1}{\log r}.$$

Consequently, a may be chosen sufficiently large so that

$$1_{\{Q(\Phi_{3r}) \wedge Q(\Theta_{2r}) \geqslant c_0\}} \mathbf{P}[S_* \not\subset \mathfrak{B}_{r'/3} \,|\, \Theta_{2r}, \Phi_{3r}, \mathcal{Z}_0^\sigma] < \tfrac{1}{10}p.$$

Hence, (3.45) implies that

$$\mathbf{P}[S_* \not\subset \mathfrak{B}_{r'/3} \,|\, \Theta_r, \Phi_R, \mathcal{Z}_0^\sigma] < \tfrac{1}{5}p. \tag{3.46}$$

Let S' be the path traced by $\mathrm{ext}_{3r}S$ from the last time in which $\mathrm{ext}_{3r}S$ was outside $\mathfrak{B}_{R/3}$ until its terminal point $q^{3r} \in \mathfrak{B}_{3r}$. Let $\widetilde{S}$ be the path traced by $\mathrm{int}_{R/3}\breve{S}$ from the last time in which $\mathrm{int}_{R/3}\breve{S}$ was inside $\mathfrak{B}_{3r}$ until its terminal point $\hat{q}^{R/3}$. Observe that S' is Φ_{3r}-measurable, $\widetilde{S}$ is $\Theta_{R/3}$-measurable, and when $S_* \subset \mathfrak{B}_{r'/3}$, we have $S' = \widetilde{S}$ as unoriented paths.

Observe that there is a connected component α of $\partial B(0,r')\setminus S'$ such that $S'\cup\alpha$ separates ∂D from $\mathfrak{B}_{3r}$. We fix such an α, and if there is more than one possible choice, we choose one in a way which depends only on S'.

On the event $Q(\Phi_{3r})>0$ each strand of $\mathrm{ext}_{3r}\gamma$ connects ∂D with $\mathfrak{B}_{3r}$, and hence $\mathrm{ext}_{3r}\gamma$ intersects α. Thus, there are precisely two connected components of $\alpha\cap D(\mathrm{ext}_{3r}\gamma)$ which have one endpoint in S' and the other in $\partial D(\mathrm{ext}_{3r}\gamma)$. Let α_1 and α_2 be these two arcs.

We now argue that

$$1_{\{Q(3r)\wedge Q(2r)\geqslant c_0\}}\mathbf{P}[\gamma\cap\alpha_1\neq\varnothing\,|\,\mathcal{Z}_0^\sigma,\Phi_{3r},\Theta_{2r}]<\tfrac{1}{10}p \tag{3.47}$$

if a is sufficiently large.

The basic idea of the proof of (3.47) is to construct a sequence of barriers separating $\mathfrak{B}_{3r}$ from α_1 in $D(\mathrm{ext}_{3r}\gamma)$ such that if γ hits a barrier in the sequence, the conditional probability that it will hit the next barrier is bounded away from 1.

Let D' be the connected component of $D(\mathrm{ext}_{3r}\gamma)\setminus\mathfrak{B}_{3r}$ that contains α_1. Note that D' is a simply connected domain. See Figure 3.10. Let z_{α_1} denote the endpoint of α_1 on $\partial D'$ and let ξ_1 and ξ_2 be the two connected components of $\partial D'\setminus(\{z_{\alpha_1}\}\cup\overline{\mathfrak{B}}_{3r})$. (Both have z_{α_1} as an endpoint and the other endpoint in $\partial\mathfrak{B}_{3r}$.) Note that any path in $D'\setminus\alpha_1$ connecting ξ_1 and ξ_2 separates α_1 from $\partial\mathfrak{B}_{3r}$ in $D(\mathrm{ext}_{3r}\gamma)$. For each $\varrho\in(3r+3,r'-3)$ let $A(\varrho)$ denote the connected component of $D'\setminus\overline{B(0,\varrho)}$ that contains α_1, and let $\alpha(\varrho)$ denote the connected component of $\partial A(\varrho)\cap\partial B(0,\varrho)$ that separates α_1 from $\mathfrak{B}_{3r}$ in D'. Observe that $\alpha(\varrho)$ has one endpoint on ξ_1 and the other on ξ_2. If $3r+3<\varrho<\varrho'<r'-3$, then $A(\varrho)\supset A(\varrho')$ and therefore $\alpha(\varrho')$ separates $\alpha(\varrho)$ from α_1 in $D(\mathrm{ext}_{3r}\gamma)$ and $\alpha(\varrho)$ separates $\alpha(\varrho')$ from $\mathfrak{B}_{3r}$. When $3r+3<\varrho<\varrho'<r'-3$, let $A(\varrho,\varrho')$ denote the connected component of $D'\setminus(\alpha(\varrho)\cup\alpha(\varrho'))$ whose boundary contains $\alpha(\varrho)\cup\alpha(\varrho')$.

For $n\in\mathbb{N}$, let $\varrho_n:=2^n(3r+3)$, and let N be the largest n such that $8\varrho_n<r'$. Fix some $n\in\{1,\dots,N\}$ and some small $\delta>0$ ($\delta=\frac{1}{100}$ should do). Set $A_n^\delta:=A((1-\delta)\varrho_n,(1+\delta)\varrho_n)$. By continuity, A_n^δ contains points w such that $\mathrm{dist}(w,\xi_1)=\mathrm{dist}(w,\xi_2)$. Set

$$\eta(n):=\min\bigl\{\mathrm{dist}(w,\xi_1):w\in\overline{A_n^\delta},\mathrm{dist}(w,\xi_1)=\mathrm{dist}(w,\xi_2)\bigr\}.$$

First, assume that $\eta(n)>\delta\varrho_n$. In that case, a barrier Υ_n is defined as follows. By continuity, there is a subarc Υ' of $\alpha(\varrho_n)\subset A_n^\delta$ with endpoints z_1 and z_2 such that $\mathrm{dist}(z_j,\xi_j)=\frac{1}{10}\delta\varrho_n$, $j=1,2$, and $\mathrm{dist}(\Upsilon',\partial D')=\frac{1}{10}\delta\varrho_n$. Let z_j' be a point in ξ_j at distance $\frac{1}{10}\delta\varrho_n$ from z_j, $j=1,2$. Then we take Υ_n as the union of Υ' with the two line segments $[z_1,z_1']$ and $[z_2,z_2']$. Recall the definition of $\mathrm{dist}(\cdot,\cdot;\cdot)$ from (3.30), and note that $\mathrm{dist}(z_j',\xi_{3-j};D')>\frac{1}{10}\delta\varrho_n$, $j=1,2$, for otherwise, by continuity again, there would be a

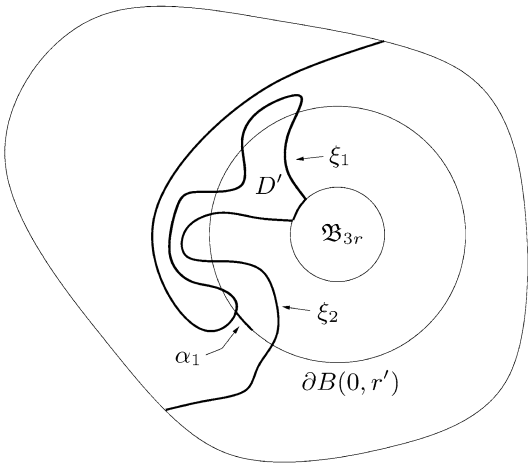


Figure 3.10. The domain D'.

point $w\in\overline{D}'$ satisfying $\operatorname{dist}(w,z'_j;D')\leqslant\frac{1}{10}\delta\varrho_n$ (and therefore $w\in\overline{A_n^\delta}$) that is at equal distance from ξ_1 and from ξ_2, which would contradict our assumption $\eta(n)>\delta\varrho_n$. It easily follows that in this case Υ_n is a $\big(\frac{1}{20}\delta,2\varrho_n\big)$-barrier.

We now assume that $\eta(n)\leqslant\delta\varrho_n$. Let $w_1\in\overline{A_n^\delta}$ be a point satisfying

$$\operatorname{dist}(w_1,\xi_1)=\operatorname{dist}(w_1,\xi_2)\leqslant\delta\varrho_n.$$

For $j=1,2$, let $p_j\in\xi_j$ be a point satisfying $|w_1-p_j|=\operatorname{dist}(w_1,\partial D')$. If

$$\operatorname{dist}(p_j,\xi_{3-j};D')\geqslant\delta^2\varrho_n\quad\text{for } j=1,2,$$

then we may take as our barrier the union $[p_1,w_1]\cup[w_1,p_2]$. This will be a $\big(\frac{1}{4}\delta,2\delta\varrho_n\big)$-barrier. Otherwise, fix a point w_2 satisfying

$$\operatorname{dist}(w_2,p_j;D')\leqslant\delta^2\varrho_n\quad\text{and}\quad\operatorname{dist}(w_2,\xi_1)=\operatorname{dist}(w_2,\xi_2)\leqslant\delta^2\varrho_n$$

and consider the above construction with w_2 in place of w_1. It may happen that the construction succeeds now, and we construct a $\big(\frac{1}{4}\delta,2\delta^2\varrho_n\big)$-barrier. Otherwise, we find a point $w_3\in D'$ satisfying $\operatorname{dist}(w_3,w_1;D')\leqslant(\delta+\delta^2+\delta^3)\varrho_n$ such that

$$\operatorname{dist}(w_3,\xi_1)=\operatorname{dist}(w_3,\xi_2)\leqslant\delta^3\varrho_n.$$

We continue this procedure until some $\big(\frac{1}{4}\delta,2\delta^m\varrho_n\big)$-barrier is obtained. The procedure must terminate successfully at some finite m, for otherwise the points w_m would converge to some point in $\xi_1\cap\xi_2$ within distance $2\delta\varrho_n$ from $\overline{A_n^\delta}$, which is clearly impossible. Note

that the barrier Υ_n thus constructed is contained in $A_n^{2\delta}$. Thus, when $1\leqslant n'<n\leqslant N$, $n,n'\in\mathbb{N}$, we have $\operatorname{dist}(\Upsilon_n,\Upsilon_{n'};D')>\frac{1}{4}\varrho_n$ and Υ_n separates α_1 from $\Upsilon_{n'}$ in D'.

Suppose $n\in\{1,...,N-1\}$. Note that (contrary to what appears in Figure 3.10, which does not show the scale of the lattice) the endpoints of Υ_n are not on γ, since $\xi_1\cup\xi_2$ are disjoint from γ, by construction. On the event $\gamma\cap\Upsilon_n\neq\varnothing$, let γ_1 and γ_2 be the two arcs of γ extending from the endpoints of ∂_+ to the first encounter with Υ_n. Now we apply Theorem 3.11 with $Y=\{\Upsilon_{n+1}\}$. Our careful construction above ensures that Υ_{n+1} is an $(\varepsilon,\operatorname{diam}\Upsilon_{n+1})$-barrier for some universal constant $\varepsilon>0$. Note that ξ_1 and ξ_2 contain vertices on which h takes the same sign. We conclude from the theorem that

$$\mathbf{P}[\gamma\cap\Upsilon_{n+1}\neq\varnothing\,|\,\gamma_1,\gamma_2,\Phi_{3r},\Theta_{2r}]<1-c_1$$

for some $c_1=c_1(\bar{\Lambda})>0$. The above implies that

$$\mathbf{P}[\gamma\cap\Upsilon_{n+1}\neq\varnothing\,|\,\gamma\cap\Upsilon_n\neq\varnothing,\Phi_{3r},\Theta_{2r}]<1-c_1,$$

which gives

$$\mathbf{P}[\gamma\cap\alpha_1\neq\varnothing\,|\,\Phi_{3r},\Theta_{2r}]\leqslant(1-c_1)^{N-1}. \tag{3.48}$$

Conditioned on γ,Φ_{3r} and Θ_{2r}, the probability of $\mathcal{Z}_0^\sigma$ is at most $O(1)/\log r$, by the proof of the upper bound in Lemma 3.14. Thus,

$$\mathbf{P}[\gamma\cap\alpha_1\neq\varnothing,\mathcal{Z}_0^\sigma\,|\,\Phi_{3r},\Theta_{2r}]\leqslant\frac{O(1)(1-c_1)^{N-1}}{\log r}.$$

On the other hand, the lower bound tells us that on the event $\{Q(3r)\wedge Q(2r)\geqslant c_0\}$, we have $O_{\bar{\Lambda},p}(1)\mathbf{P}[\mathcal{Z}_0^\sigma\,|\,\Phi_{3r},\Theta_{2r}]\geqslant 1/\log r$. Thus, (3.47) follows.

Clearly, (3.47) also holds for α_2. On $\mathcal{Z}_0^\sigma$ let α_1^* and α_2^* be the two connected components of $\alpha\cap D(\gamma)$ that have one endpoint in S' and the other in $\partial D(\gamma)$. Note that when $\mathcal{Z}_0^\sigma$ holds and $\gamma\cap(\alpha_1\cup\alpha_2)=\varnothing$, we have $\alpha_1^*\cup\alpha_2^*=\alpha_1\cup\alpha_2$. Thus, (3.47) for α_1 and for α_2 together with (3.45) now gives

$$\mathbf{P}[\alpha_1^*\cup\alpha_2^*\neq\alpha_1\cup\alpha_2\,|\,\Phi_R,\Theta_r,\mathcal{Z}_0^\sigma]<\tfrac{3}{10}p. \tag{3.49}$$

We now follow an analog of the above argument with the roles of inside and outside switched. Observe that there is a connected component $\widetilde{\alpha}$ of $\partial B(0,r')\setminus\widetilde{S}$ such that $\widetilde{S}\cup\widetilde{\alpha}$ separates ∂D from $\mathfrak{B}_{3r}$. We fix such an $\widetilde{\alpha}$, and if there is more than one possible choice, we choose it in the same way in which α was chosen from S'; that is, we make sure that $\alpha=\widetilde{\alpha}$ if $S'=\widetilde{S}$ (as unoriented paths). The point is that although α is Φ_{3r}-measurable and $\widetilde{\alpha}$ is $\Theta_{R/3}$-measurable, we have $\alpha=\widetilde{\alpha}$ if $S'=\widetilde{S}$.

On the event $Q(\Theta_{R/3})>0$, let $\widetilde{\alpha}_1$ and $\widetilde{\alpha}_2$ be the two connected components of $\widetilde{\alpha}\cap D(\beta_{R/3})$ that have one endpoint on $\widetilde{S}$ and the other on $\partial D(\beta_{R/3})$. On the event $\mathcal{Z}_0^\sigma$, let $\widetilde{\alpha}_1^*$ and $\widetilde{\alpha}_2^*$ be the two connected components of $\widetilde{\alpha}\cap D(\gamma)$ that have one endpoint on $\widetilde{S}$ and the other on $\partial D(\gamma)$. Essentially the same proof which gave (3.49) now gives

$$\mathbf{P}[\widetilde{\alpha}_1^*\cup\widetilde{\alpha}_2^*\neq\widetilde{\alpha}_1\cup\widetilde{\alpha}_2\,|\,\Phi_R,\Theta_r,\mathcal{Z}_0^\sigma]<\tfrac{3}{10}p. \tag{3.50}$$

But observe that when $\mathcal{Z}_0^\sigma$ and $S'=\widetilde{S}$ hold, we clearly have $\widetilde{\alpha}_1^*\cup\widetilde{\alpha}_2^*=\alpha_1^*\cup\alpha_2^*$. Also recall that $S'=\widetilde{S}$ when $S_*\subset\mathfrak{B}_{r'/3}$. Thus, from (3.46), (3.49) and (3.50), we get

$$\mathbf{P}[\widetilde{\alpha}_1\cup\widetilde{\alpha}_2\neq\alpha_1\cup\alpha_2\text{ or }S_*\not\subset\mathfrak{B}_{r'/3}\,|\,\Phi_R,\Theta_r,\mathcal{Z}_0^\sigma]<\tfrac{4}{5}p.$$

Assume that $\mathcal{Z}_0^\sigma$, $\widetilde{\alpha}_1\cup\widetilde{\alpha}_2=\alpha_1^*\cup\alpha_2^*=\alpha_1\cup\alpha_2$ and $S_*\subset\mathfrak{B}_{r'/3}$ hold. It remains to show that in this case the path γ has no more than two disjoint arcs connecting $\mathfrak{B}_r$ and $\partial\mathfrak{B}_R$. Recall that $\widetilde{\alpha}_1$ has an endpoint on $\partial D(\beta_{R/3})$. This endpoint is on a TG-triangle containing a TG*-vertex $v_1\in\beta_{R/3}$. Similarly, there is a TG*-vertex $v_2\in\beta_{R/3}$ for which the TG-triangle containing it has an endpoint of $\widetilde{\alpha}_2$. From $\widetilde{\alpha}_1\cup\widetilde{\alpha}_2=\alpha_1\cup\alpha_2$, we conclude that $v_1,v_2\in\mathrm{ext}_{3r}\gamma$ as well.

Shortly, we will prove that v_1 and v_2 are in separate connected components of $\beta_{R/3}\backslash e_\sigma^*$. This implies that each connected component of $\beta_{R/3}\backslash e_\sigma^*$ intersects $\mathrm{ext}_{3r}\gamma$. Since $\beta_{R/3}\cup\mathrm{ext}_{3r}\gamma\subset\gamma$, and γ is a simple path, it easily follows that $\gamma=\beta_{R/3}\cup\mathrm{ext}_{3r}\gamma$, which implies that there are at most two disjoint crossings in γ between $\mathfrak{B}_r$ and $\partial\mathfrak{B}_R$.

It remains to prove that v_1 and v_2 are in different connected components of $\beta_{R/3}\backslash e_\sigma^*$. This will be established using planar topology arguments. Let $\widehat{\alpha}$ consist of $\widetilde{\alpha}_1\cup\widetilde{\alpha}_2$, a simple path $\widetilde{S}_0\subset\widetilde{S}$ connecting them, and short line segments (contained in the TG-triangles containing v_1 and v_2) from the endpoints of $\widetilde{\alpha}_1$ and $\widetilde{\alpha}_2$ on $\partial D(\beta_{R/3})$ to v_1 and v_2. Then $\widehat{\alpha}$ is a simple path and only the endpoints of $\widehat{\alpha}$ are on γ. Let $\tilde{\beta}$ be the connected component of $\beta_{R/3}\backslash\{v_1,v_2\}$ with endpoints v_1 and v_2. Then $\tilde{\beta}\cup\widehat{\alpha}$ is a simple closed path and it suffices to show that $e_\sigma^*\subset\tilde{\beta}$. If $\tilde{\beta}\cup\widehat{\alpha}$ separates 0 from ∂D, then $\tilde{\beta}$ must contain e_σ^*, because each connected component of $\gamma\backslash\{e_\sigma^*\}$ connects ∂D to TG*-vertices adjacent to 0 and is disjoint from $\widehat{\alpha}\backslash\{v_1,v_2\}$.

Suppose that $\tilde{\beta}\cup\widehat{\alpha}$ does not separate 0 from ∂D. Recall that $\widetilde{\alpha}\cup\widetilde{S}$ separates ∂D from $\mathfrak{B}_{3r}$ and therefore from 0. Since $\widetilde{S}$ itself does not separate ∂D from 0, it follows that the winding number of $\widetilde{\alpha}\cup\widetilde{S}_0$ around 0 is ±1 (depending on orientation). Since $\tilde{\beta}\cup\widehat{\alpha}$ does not separate 0 from ∂D, its winding number around 0 is zero. If we remove from the union of the two paths $\widetilde{\alpha}\cup\widetilde{S}_0$ and $\tilde{\beta}\cup\widehat{\alpha}$ all the non-trivial arcs where they agree, we get a closed curve χ, which consists of $\tilde{\beta}$, a segment of $\widetilde{\alpha}$ and the two short connecting segments near v_1 and v_2, and χ has odd winding number around 0. Consequently, it

separates 0 from ∂D. But observe that $\widetilde{S}$ is disjoint from χ. Moreover, since we are assuming $S_* \subset \mathfrak{B}_{r'/3}$, it follows that $\mathrm{int}_{R/3}\breve{S}$ is also disjoint from it. But this contradicts the fact that χ separates ∂D from 0, since $\mathrm{int}_{R/3}\breve{S}$ can be extended to a path disjoint from χ and connecting 0 and ∂D. Thus, the proof is now complete. □

3.6. Coupling and limit

In this subsection, we retain our previous assumptions (h), (D) and (S) about the system D, ∂_+, ∂_-, h_∂, v_0, h, γ. Moreover, we also consider another such system D', ∂'_+, ∂'_-, h'_∂, v'_0, h', γ', which is supposed to satisfy the same assumptions. In particular, $\|h'_\partial\|_\infty \leqslant \bar{\Lambda}$. Generally, we will use $'$ to denote objects related to the system in D'. For example, $\mathcal{J}'(r,R)$ will denote the event corresponding to $\mathcal{J}(r,R)$.

Definition 3.18. Fix $R>r>0$, and suppose that $\mathfrak{B}_R \subset D \cap D'$. Consider the intersection $\mathrm{ext}_r\gamma \cap \mathfrak{B}_R$ as a collection of oriented paths, oriented so as to have vertices in $V_+(\gamma)$ on the right. We say that Φ_r and Φ'_r *match* in $\mathfrak{B}_R$ if the set of vertices in $\mathfrak{B}_R$ visited by $\mathrm{ext}_r S$ is the same as the corresponding set for S' and $\mathrm{ext}_r\gamma \cap \mathfrak{B}_R = \mathrm{ext}_r\gamma' \cap \mathfrak{B}_R$ with all the orientations agreeing or with all the orientations reversed.

We now show that if the configurations match in a big annulus, then it is likely that the interfaces agree in the inner disk; more precisely, we have the following result.

LEMMA 3.19. *For every $\delta>0$, $r \geqslant 10$ and $R^*>r+3$ there is an $R=R(\delta,r,R^*,\bar{\Lambda})>R^*$ such that the following holds. Suppose that $v'_0, v_0 \notin \mathfrak{B}_R \subset D\cap D'$, $\mathbf{P}[\neg\mathcal{J}(r,R^*), \mathcal{Z}_0 | \Phi_r]>0$ and $\mathbf{P}[\neg\mathcal{J}'(r,R^*), \mathcal{Z}'_0 | \Phi'_r]>0$. Assume that Φ_r and Φ'_r match in $\mathfrak{B}_R$. In particular, the endpoint q^r of $\mathrm{ext}_r S$ in $\mathfrak{B}_r$ is the same as that of $\mathrm{ext}_r S'$. Let ν be the law of $\gamma_* := \gamma \setminus \mathrm{ext}_r\gamma$ (as an unoriented path) conditioned on $\mathcal{Z}_0$, Φ_r and $\neg\mathcal{J}(r,R^*)$, and let ν' be the law of $\gamma'_* := \gamma' \setminus \mathrm{ext}_r\gamma'$ conditioned on $\mathcal{Z}'_0$, Φ'_r and $\neg\mathcal{J}'(r,R^*)$. Then $\|\nu-\nu'\|<\delta$.*

Here, $\|\nu-\nu'\|$ denotes the total variation norm $\sum_\vartheta |\nu[\gamma_*=\vartheta]-\nu'[\gamma'_*=\vartheta]|$.

Proof. Assume that the orientation of $\mathrm{ext}_r\gamma \cap \mathfrak{B}_R$ agrees with that of $\mathrm{ext}_r\gamma' \cap \mathfrak{B}_R$. This involves no loss of generality, since we may replace ∂'_+ with ∂'_-, replace h' by $-h'$, etc.

Since we are assuming that $\mathbf{P}[\neg\mathcal{J}(r,R^*), \mathcal{Z}_0 | \Phi_r]>0$, there is a path $\vartheta \subset \mathfrak{B}_{R^*}$ such that $\mathbf{P}[\gamma_*=\vartheta, \neg\mathcal{J}(r,R^*), \mathcal{Z}_0 | \Phi_r]>0$. Let $\widehat{\Gamma}$ be the collection of all such ϑ, and fix some $\vartheta \in \widehat{\Gamma}$. Obviously, the length of ϑ is $O(R^*)^2$. We start extending $\mathrm{ext}_r\gamma$ starting at one of the endpoints, say x^r, and consider the conditional probability that each successive step follows ϑ, given that the previous steps follow θ and given Φ_r. Each step is decided by the sign of h on a specific vertex v. When we condition on the values of h on the

neighbors of v, the conditional law of $h(v)$ is a Gaussian with some constant positive variance. It follows from (3.2) that with high probability (conditioned on the success of the previous steps) the mean of this Gaussian random variable is unlikely to be large. Thus, the probability for either sign is bounded away from zero, which means that each step is successful with probability bounded away from zero. By (3.3) it is unlikely that $h(v)$ will be very close to zero. Proposition 3.3 therefore implies that if $R>R^*$ is very large, the probability for a successful one-step extension for γ' is almost the same as for γ. Thus, we conclude that for sufficiently large $R>R^*$,

$$(1-\delta)\mathbf{P}[\gamma'_*=\vartheta\,|\,\Phi'_r]\leqslant\mathbf{P}[\gamma_*=\vartheta\,|\,\Phi_r]\leqslant(1+\delta)\mathbf{P}[\gamma'_*=\vartheta\,|\,\Phi'_r]$$

holds for all $\vartheta\in\widehat{\Gamma}$. It is moreover clear that

$$\mathbf{P}[\mathcal{Z}_0,\neg\mathcal{J}(r,R^*)\,|\,\gamma_*=\vartheta,\Phi_r]=\mathbf{P}[\mathcal{Z}'_0,\neg\mathcal{J}'(r,R^*)\,|\,\gamma'_*=\vartheta,\Phi'_r],$$

because under $\neg\mathcal{J}(r,R^*)$ the random walk S cannot get close to any place where $\mathrm{ext}_r\gamma$ differs from $\mathrm{ext}_r\gamma'$ between the first visit to q^r and time τ_0. Thus,

$$1-\delta\leqslant\frac{\mathbf{P}[\gamma_*=\vartheta,\mathcal{Z}_0,\neg\mathcal{J}(r,R^*)\,|\,\Phi_r]}{\mathbf{P}[\gamma'_*=\vartheta,\mathcal{Z}'_0,\neg\mathcal{J}'(r,R^*)\,|\,\Phi'_r]}\leqslant1+\delta.$$

The lemma follows (though perhaps δ needs to be readjusted). □

The next lemma shows that given Φ_R the events $\mathcal{Z}_0^\sigma$ have comparable probabilities for different σ.

LEMMA 3.20. *As usual, assume* (h), (D) *and* (S). *There is a constant* $c=c(\bar{\Lambda})\geqslant1$ *such that for all* R *sufficiently large and every* $\sigma,\sigma'\in\{0,1,...,5\}$ *we have*

$$\mathbf{P}[\mathcal{Z}_0^\sigma\,|\,\Phi_R]\leqslant c\mathbf{P}[\mathcal{Z}_0^{\sigma'}\,|\,\Phi_R].$$

Proof. The statement is clear when $R=100$, because in that case if it is at all possible to extend Φ_R in such a way that $\mathcal{Z}_0^\sigma$ holds, then the probability that $\mathcal{Z}_0^{\sigma'}$ holds is bounded away from zero (by a function of $\bar{\Lambda}$). (We may choose the continuations of γ and S as we please, and as long as the continuations involve a bounded number of steps, the probability for these continuations are bounded away from zero, as in the proof of Lemma 3.19.) When $R>100$, we may just condition on the corresponding extension of Φ_R up to radius 100. □

We now come to one of the main results in this section—the existence of a limiting interface.

THEOREM 3.21. (Limit existence) *There is a (unique) probability measure μ_∞ on the space of two-sided infinite simple* TG**-paths $\dot\gamma$ which is the limit of the law of γ (unoriented) conditioned on $\mathcal{Z}_0$ and Φ_R, in the following sense. Assume* (h), (D) *and* (S). *For every finite set of* TG**-edges E_0 and every $\delta>0$ there is an $R_0=R_0(\delta,E_0,\bar\Lambda)$ such that if $R>R_0$, $v_0\notin\mathfrak{B}_R\subset D$ and $\mathbf{P}[\mathcal{Z}_0|\Phi_R]>0$, then*

$$|\mathbf{P}[E_0\subset\gamma\,|\,\mathcal{Z}_0,\Phi_R]-\mu_\infty[E_0\subset\dot\gamma]|<\delta.$$

Proof. Clearly, it suffices to show that, for every $r>0$, if R is sufficiently large, $v_0,v_0'\notin\mathfrak{B}_R\subset D\cap D'$, $\mathbf{P}[\mathcal{Z}_0|\Phi_R]>0$ and $\mathbf{P}[\mathcal{Z}_0'|\Phi_R']>0$, then we may couple the conditioned laws of γ given Φ_R and $\mathcal{Z}_0$, and γ' given Φ_R' and $\mathcal{Z}_0'$ such that

$$\mathbf{P}[\gamma\backslash\mathrm{ext}_r\gamma=\gamma'\backslash\mathrm{ext}_r\gamma'\,|\,\Phi_R,\mathcal{Z}_0,\Phi_R',\mathcal{Z}_0']>1-\delta. \tag{3.51}$$

(Here, the equivalence is an equivalence of unoriented paths.) Let a_n be the constant a given by Lemma 3.17 when one takes $p=\delta_n:=\frac18 2^{-n}\delta$. We define a sequence of radii $r_0,r_1,\ldots$ inductively, as follows. Let c_1 be the constant c given by Corollary 3.16, and set $r_0:=r\vee10\vee c_1$. Given r_n, let $\hat r_n$ be the R promised by Lemma 3.19 when we take r_n for r, δ_n for δ and a_nr_n for R^*. Finally, set $r_{n+1}=4a_n\hat r_n$. Let c_2 be the constant promised by Lemma 3.20. We assume, with no loss of generality, that $\delta<1/4c_1$. Let $N\in\mathbb{N}$ be sufficiently large so that $(1-1/72c_1c_2^2)^{N-1}<\frac12\delta$. We will prove (3.51) on the assumption that $R>6r_N$.

The construction of the coupling is as follows. First, we choose $\Phi_{r_{N-1}}$ and $\Phi'_{r_{N-1}}$ independently according to their conditional distribution given Φ_R, $\mathcal{Z}_0$, Φ_R' and $\mathcal{Z}_0'$. We proceed by reverse induction. Suppose that $n\in[1,N-1]\cap\mathbb{N}$ and that Φ_{r_n} and Φ'_{r_n} have been determined. If Φ_{r_n} and Φ'_{r_n} match inside $\mathfrak{B}_{\hat r_n}$, then we couple γ and γ' in such a way as to maximize the probability that $\gamma\backslash\mathrm{ext}_{r_n}\gamma=\gamma'\backslash\mathrm{ext}_{r_n}\gamma'$, subject to maintaining their correct conditional distributions given the choices previously made. If they do not match, then we couple $\Phi_{r_{n-1}}$ and $\Phi'_{r_{n-1}}$ in such a way as to maximize the probability that they match in $\mathfrak{B}_{\hat r_{n-1}}$, subject to their correct conditional distributions. If Φ_{r_0} and Φ'_{r_0} have been determined, but γ and γ' have not, then we couple γ and γ' arbitrarily, subject to their correct conditional distributions.

We claim that the coupling just described achieves the bound (3.51). Let $\mathcal{M}_n$ denote the event that Φ_{r_n} and Φ'_{r_n} match inside $\mathfrak{B}_{\hat r_n}$, where $n\in\{1,\ldots,N-1\}$, and let $\mathcal{M}$ denote the union of these events $\mathcal{M}_n$. It follows from the choice of $\hat r_n$ that if $\mathcal{M}_n$ holds and n is the largest $n'\in\{1,\ldots,N-1\}$ with that property, then there is a coupling of the appropriately conditioned laws of γ and γ' such that

$$\begin{aligned}\mathbf{P}[\gamma\backslash\mathrm{ext}_{r_n}\gamma\neq\gamma'\backslash\mathrm{ext}_{r_n}\gamma'\,|\,\Phi_{r_n},\Phi'_{r_n},\mathcal{Z}_0,\mathcal{Z}_0']&\\ \leqslant\delta_n+\mathbf{P}[\mathcal{J}(r_n,a_nr_n)\,|\,\Phi_{r_n},\mathcal{Z}_0']+\mathbf{P}[\mathcal{J}'(r_n,a_nr_n)\,|\,\Phi'_{r_n},\mathcal{Z}_0'],&\end{aligned}$$

and hence this also holds for our coupling. By taking conditional expectation and summing over n, we get

$$\begin{aligned}\mathbf{P}[\mathcal{M}, \gamma\backslash\mathrm{ext}_r\gamma \neq \gamma'\backslash\mathrm{ext}_r\gamma' \mid \Phi_R, \mathcal{Z}_0, \Phi'_R, \mathcal{Z}'_0] \\ \leqslant \frac{1}{8}\delta + \sum_{n=1}^{N-1} (\mathbf{P}[\mathcal{J}(r_n, a_n r_n) \mid \Phi_R, \mathcal{Z}_0] + \mathbf{P}[\mathcal{J}'(r_n, a_n r_n) \mid \Phi'_R, \mathcal{Z}'_0]) \leqslant \frac{3}{8}\delta.\end{aligned}$$

(where the last inequality follows by the choice of a_n).

Now fix some $n \in \{1, ..., N-2\}$, and suppose that none of the events $\mathcal{M}_{n'}$, $n' > n$, occurs. Fix some arbitrary $\sigma \in \{0, 1, ..., 5\}$. Conditional on $\Phi_{r_{n+1}}$ and $\mathcal{Z}_0$, by the choice of c_2, there is probability at least $1/6c_2$ that $\mathcal{Z}_0^\sigma$ holds, and the same is true for the system in D'. If we additionally condition on $\mathcal{Z}_0^\sigma$ and $\mathcal{Z}_0^{\sigma\prime}$, then, by the choice of c_1, there is a coupling of the appropriate conditioned laws of $\Theta_{r_{n+1}/4}$ and $\Theta'_{r_{n+1}/4}$ such that

$$\mathbf{P}[\Theta_{r_{n+1}/4} = \Theta'_{r_{n+1}/4} \mid \Phi_{r_{n+1}}, \Phi'_{r_{n+1}}, \mathcal{Z}_0^\sigma, \mathcal{Z}_0^{\sigma\prime}] \geqslant \frac{1}{c_1}.$$

But note that if $\Theta_{r_{n+1}/4} = \Theta'_{r_{n+1}/4}$ and $\neg\big(\mathcal{J}\big(\hat{r}_n, \frac{1}{4}r_{n+1}\big) \cup \mathcal{J}'\big(\hat{r}_n, \frac{1}{4}r_{n+1}\big)\big)$ both hold (as well as $\mathcal{Z}_0^\sigma \cap \mathcal{Z}_0^{\sigma\prime}$), then $\mathcal{M}_n$ holds as well. We may then consider a coupling of the two systems which first decides the two events $\mathcal{Z}_0^\sigma$ and $\mathcal{Z}_0^{\sigma\prime}$ independently, and if both hold (which happens with probability at least $1/(6c_2)^2$), then with conditional probability at least $1/c_1$ we also have $\Theta_{r_{n+1}/4} = \Theta'_{r_{n+1}/4}$. Thus, under this coupling,

$$\begin{aligned}(6c_2)^2 \mathbf{P}[\mathcal{M}_n \mid \Phi_{r_{n+1}}, \Phi'_{r_{n+1}}, \mathcal{Z}_0, \mathcal{Z}'_0] \\ &\geqslant c_1^{-1} - \mathbf{P}\big[\mathcal{J}\big(\hat{r}_n, \tfrac{1}{4}r_{n+1}\big) \cup \mathcal{J}'\big(\hat{r}_n, \tfrac{1}{4}r_{n+1}\big) \mid \Phi_{r_{n+1}}, \Phi'_{r_{n+1}}, \mathcal{Z}_0^\sigma, \mathcal{Z}_0^{\sigma\prime}\big] \\ &\geqslant c_1^{-1} - 2\delta_n \quad \text{(by the choices of } r_{n+1} \text{ and } a_n) \\ &\geqslant \tfrac{1}{2}c_1^{-1} \quad \text{(by our assumption } 4c_1\delta < 1).\end{aligned}$$

This must hold for our coupling as well. Consequently, induction gives

$$\mathbf{P}[\neg\mathcal{M}] \leqslant \left(1 - \frac{1}{72c_2^2c_1}\right)^{N-1},$$

which is less than $\frac{1}{2}\delta$ by the choice of N. Therefore (3.51) follows, and the proof is complete. □

3.7. Boundary values of the interface

Consider the random path $\dot{\gamma}$ whose law is the measure μ_∞ provided by Theorem 3.21. We orient $\dot{\gamma}$ so that the edges e_σ^*, $\sigma = 0, 1, ..., 5$, that are in $\dot{\gamma}$ are oriented clockwise around the

hexagon $\bigcup_{\sigma=0}^{5} e_\sigma^*$ centered at 0. Let U_+ denote the set of TG-vertices adjacent to $\dot\gamma$ on its right-hand side, and let U_- denote the set of vertices adjacent to $\dot\gamma$ on its left-hand side. Using the heights interface continuity (Proposition 3.3), it is clear that given $\dot\gamma$ we may define the DGFF $\dot h$ on all of TG conditioned to be positive on U_+ and negative on U_-, as a limit of an appropriately conditioned DGFF on bounded domains. Moreover, many properties of the DGFF on bounded domains easily transfer to $\dot h$. In particular, (3.2) applies, to give $\mathbf{E}[|\dot h(0)||\dot\gamma]=O(1)$. Set

$$\lambda := \mathbf{E}[\dot h(0)]. \tag{3.52}$$

Clearly, $0<\lambda<\infty$.

Recall that τ is the first time t such that $S_t\in\partial D(\gamma)$ and recall the notation $\operatorname{dist}(\cdot,\cdot;\cdot)$ from (3.30). In this subsection we will show that in the limit as $\operatorname{dist}(v_0,\partial D)\to\infty$ we have

$$\mathbf{E}[(\pm h(S_\tau)-\lambda)1_{\{S_\tau\in V_\pm(\gamma)\}}\,|\,\gamma]\to 0$$

in probability, under the assumption that

$$h_\partial(\partial_+\cap V_\partial)\subset[-\Lambda_0,\bar\Lambda]\quad\text{and}\quad h_\partial(\partial_-\cap V_\partial)\subset[-\bar\Lambda,\Lambda_0], \tag{∂}$$

where $\Lambda_0=\Lambda_0(\bar\Lambda)>0$ is the constant given by Lemma 3.9. The importance of the assumption (∂) is that, by Remark 3.12, it enables the application of Theorem 3.11 to barriers with endpoints on ∂D, provided that the barriers in Y_+ do not have endpoints in ∂_- and those in Y_- do not have endpoints in ∂_+. This allows us to prove the following result.

THEOREM 3.22. *Assume* (h), (D), (S) *and* (∂). *There are positive constants* $\zeta_2=\zeta_2(\bar\Lambda)$ *and* $c=c(\bar\Lambda)$ *such that the following holds true. Let* z *be any point on* ∂_+. *Then for every* $r>1$,

$$\mathbf{P}[\operatorname{dist}(z,\gamma;D)<r]\leqslant c\biggl(\frac{r}{\operatorname{dist}(z,\partial_-;D)}\biggr)^{\zeta_2}.$$

Proof. Let β_1 and β_2 be the two components of $\partial_+\backslash\{z\}$. Let $R=\operatorname{dist}(z,\partial_-;D)$. For each $\varrho\in(0,R)$, let $A(\varrho)$ denote the connected component of $B(z,\varrho)\cap D$ that has z in its boundary, and let $\alpha(\varrho)$ denote the connected component of $\partial A(\varrho)\backslash\partial D$ that separates z from ∂_- in D. Using this construction, the proof proceeds as in the proof (3.48), except that the barriers start from the outside and get closer to z, and we appeal to Remark 3.12 instead of the barriers theorem. We leave it to the reader to verify that the proof carries over with no other significant modifications. □

Our next lemma shows that it is unlikely that S_τ is adjacent to γ and is near ∂D.

LEMMA 3.23. *Assume* (h), (D), (S) *and* (∂). *For all* $p>0$ *there is some* $\delta=\delta(p,\bar{\Lambda})>0$ *such that*

$$\mathbf{P}[0<\operatorname{dist}(S_\tau,\partial D)<\delta\operatorname{dist}(v_0,\partial D)]<p.$$

Proof. Let τ_D be the first t such that $S_t\in\partial D$, and let

$$M=M(S,\gamma):=\{S_t:S_t\in\partial D(\gamma)\text{ and }0\leqslant t<\tau_D\}.$$

We will prove the stronger statement

$$\mathbf{P}[\operatorname{dist}(M,\partial D)<\delta\operatorname{dist}(v_0,\partial D)]<p. \tag{3.53}$$

(By convention $\operatorname{dist}(\varnothing,\partial D)=\infty$.) Fix $r>0$ and set $R=\operatorname{dist}(v_0,\partial D)$. Conditioned on $\operatorname{dist}(M,\partial D)<r$, we have $\operatorname{dist}(S_{\tau_D},\gamma;D)<4r$ with probability bounded away from zero, since the random walk started at any $v\in M$ such that $\operatorname{dist}(v,\partial D)<r$ has probability bounded away from 0 to surround the closest point to v on ∂D (and therefore hit ∂D) before exiting the ball of radius $2r$ about that point. It therefore suffices to prove that

$$\mathbf{P}[\operatorname{dist}(S_{\tau_D},\gamma;D)<\delta R]<p \tag{3.54}$$

for $\delta=\delta(p,\bar{\Lambda})>0$. Let $\mathcal{A}_+=\mathcal{A}_+(\delta)$ denote the event $\operatorname{dist}(S_{\tau_D},\partial_-;D)>\delta^{1/2}R$, and similarly define $\mathcal{A}_-$ with ∂_- replaced by ∂_+. By conditioning on S_{τ_D}, Theorem 3.22 shows that

$$\mathbf{P}[\mathcal{A}_+,\operatorname{dist}(S_{\tau_D},\gamma;D)<\delta R]<\tfrac{1}{3}p$$

for an appropriate choice of δ. A symmetric argument applies on $\mathcal{A}_-$. Consequently, it is enough to prove that $\mathbf{P}[\neg(\mathcal{A}_+\cup\mathcal{A}_-)]<\frac{1}{3}p$ for an appropriate choice of δ.

Fix $r_0:=\delta^{1/2}R$, let L_{r_0} denote the set of points that lie on some path in $\overline{D}$ of diameter at most r_0 connecting ∂_+ and ∂_-, and let D_* be the connected component of v_0 in $D\setminus L_{r_0}$. (See Figure 3.11.) We now prove that

$$\partial D_*\setminus\partial D\text{ is contained in the union of two balls of radius }2r_0. \tag{3.55}$$

Every path connecting ∂_+ and ∂_- in $\overline{D}$ must separate v_0 from x_∂ or from y_∂ in $\overline{D}$ (because, by Jordan's theorem, it separates x_∂ from y_∂ in $\overline{D}$). Let Γ_1 (respectively, Γ_2) denote the collection of paths in $\overline{D}$ of diameter at most r_0 that connect ∂_+ and ∂_- and separate x_∂ (respectively, y_∂) from v_0. Then $\partial D_*\setminus\partial D$ is contained in the union of the set of points belonging to a path in Γ_1 and the set of points belonging to a path in Γ_2. Suppose that α and α' are two paths in Γ_1, both of which intersect ∂D_*. Let β be a path connecting x_∂ with $\alpha\cup\alpha'$ in $\overline{D}$, which is disjoint from $\alpha\cup\alpha'$, except for its endpoint. If $\beta\cap\alpha\neq\varnothing$,

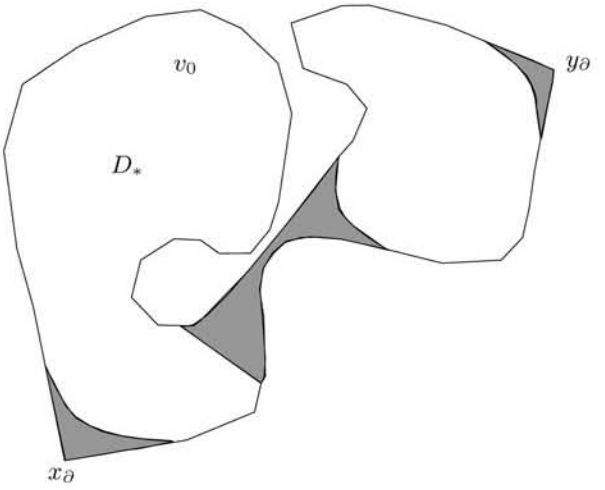


Figure 3.11. The set L_{r_0} (shaded) and D_*.

then one can connect x_∂ to v_0 in $\beta \cup \alpha \cup D_*$, and therefore $\alpha \cap \alpha' \neq \varnothing$ (since α' separates v_0 from x_∂ in $\overline{D}$). Similar reasoning applies if $\beta \cap \alpha' \neq \varnothing$. It follows that any two paths in Γ_1 that intersect ∂D_* must intersect each other, and hence the collection of all such paths is covered by the ball of radius $2r_0$ centered at any point on any such path. Since a similar argument applies to Γ_2, (3.55) follows.

By (3.55) and Lemma 2.1,

$$\mathbf{P}[\text{there exists } t \leqslant \tau_D : S_t \in L_{r_0}] < \tfrac{1}{3}p$$

for all sufficiently small $\delta > 0$. Thus $\mathbf{P}[\neg(\mathcal{A}_+ \cup \mathcal{A}_-)] = \mathbf{P}[S_{\tau_D} \in L_{r_0}] < \frac{1}{3}p$, and the proof is complete. □

Next, we show that S_τ is unlikely to be close to v_0 by proving the same for γ.

LEMMA 3.24. *Assume* (h) *and* (D). *There are constants* $c > 0$ *and* $\zeta_3 > 0$, *both depending only on* $\bar{\Lambda}$, *such that for every* $\delta > 0$,

$$\mathbf{P}[\operatorname{dist}(v_0, \gamma) < \delta \operatorname{dist}(v_0, \partial D)] < c\delta^{\zeta_3}.$$

We expect that the left-hand side is bounded by $\delta^{1/2+o(1)}$, using the corresponding result [RS] for SLE(4).

Proof. Let Υ_n denote the circle of radius $2^{-n-1} \operatorname{dist}(v_0, \partial D)$ about v_0. As in the proof of (3.48), Theorem 3.11 implies that, given that γ intersects Υ_n (where $n \in \mathbb{N}$), the conditional probability that γ does not intersect Υ_{n+1} is bounded away from zero by a function of $\bar{\Lambda}$, provided that $2^{-n-1} \operatorname{dist}(v_0, \partial D) > 10$, say. The lemma follows by induction. □

PROPOSITION 3.25. *Assume* (h), (D), (S) *and* (∂). *For every* $\varepsilon>0$ *there exists an* $R=R(\varepsilon,\bar{\Lambda})$ *such that*

$$\mathbf{E}[\mathbf{E}[(|h(S_\tau)|-\lambda)1_{\{S_\tau\notin\partial D\}}\,|\,\gamma]^2]<\varepsilon \tag{3.56}$$

holds, provided that $\operatorname{dist}(v_0,\partial D)>R$, *where* λ *is the constant given by* (3.52).

The proof is based on the simple idea that given a single instance of γ we consider two independent copies of (h,S).

Proof. Set

$$X:=\mathbf{E}[(|h(S_\tau)|-\lambda)1_{\{S_\tau\notin\partial D\}}\,|\,\gamma].$$

To get a handle on $\mathbf{E}[X^2]$, let (h',S') be independent of (h,S) given γ and have the same conditional law as that of (h,S) given γ. Thus, (h,S,γ) has the same law as (h',S',γ). Let $\tau':=\min\{t:S'_t\in\partial D(\gamma)\}$, $y:=\mathbf{E}[|h(S_\tau)|-\lambda|S_\tau,\gamma]$ and $y':=\mathbf{E}[|h'(S'_{\tau'})|-\lambda|S'_{\tau'},\gamma]$. Then $X^2=\mathbf{E}[yy'1_{\{S_\tau,S'_{\tau'}\notin\partial D\}}|\gamma]$ and hence

$$\mathbf{E}[X^2]=\mathbf{E}[yy'1_{\{S_\tau,S'_{\tau'}\notin\partial D\}}]. \tag{3.57}$$

Fix some $r_3\gg r_2\gg r_1\gg 0$, and assume that $\operatorname{dist}(0,\{v_0,\partial D\})>r_3$. Suppose that we condition on $\mathcal{Z}_0$; that is, on $S_\tau=0$. Then, with high conditional probability $|S'_{\tau'}|>2r_2$, and moreover $\operatorname{dist}(0,\{S'_0,S'_1,...,S'_{\tau'}\})>2r_2$. By the heights interface continuity (Proposition 3.3), given $\operatorname{ext}_{r_1}\gamma$ and $S'_{\tau'}$ and $\gamma\setminus\operatorname{ext}_{r_1}\gamma\subset\mathfrak{B}_{r_2}$, the actual choice of $\gamma\setminus\operatorname{ext}_{r_1}\gamma$ can change the value of y' by very little if $|S'_{\tau'}|>2r_2$. Thus, we conclude that $y'1_{S'_{\tau'}\notin\partial D}$ is nearly independent of y given $\mathcal{Z}_0$, $\neg\mathcal{J}(r_1,r_2)$ and $\operatorname{ext}_{r_1}\gamma$. Since y and y' are bounded, in the limit as $r_3\to\infty$, we have

$$\begin{aligned}&\mathbf{E}[yy'1_{\{S'_{\tau'}\notin\partial D\}}\,|\,\mathcal{Z}_0,\operatorname{ext}_{r_1}\gamma,\neg\mathcal{J}]\\&\quad=o(1)+\mathbf{E}[y'1_{\{S'_{\tau'}\notin\partial D\}}\,|\,\mathcal{Z}_0,\operatorname{ext}_{r_1}\gamma,\neg\mathcal{J}]\mathbf{E}[y\,|\,\mathcal{Z}_0,\operatorname{ext}_{r_1}\gamma,\neg\mathcal{J}],\end{aligned} \tag{3.58}$$

where $\mathcal{J}=\mathcal{J}(r_1,r_2)$. By Lemma 3.17, $\mathbf{P}[\mathcal{J}|\mathcal{Z}_0]\to 0$ as $r_2/r_1\to\infty$. Thus

$$\mathbf{E}[y\,|\,\mathcal{Z}_0,\operatorname{ext}_{r_1}\gamma,\neg\mathcal{J}]-\mathbf{E}[y\,|\,\mathcal{Z}_0,\operatorname{ext}_{r_1}\gamma]\to 0$$

in probability as $r_2/r_1\to\infty$. A similar remark applies to the other terms in (3.58). Since y and y' are bounded, taking conditional expectation given $\mathcal{Z}_0$ in (3.58) gives

$$\mathbf{E}[yy'1_{\{S'_{\tau'}\notin\partial D\}}\,|\,\mathcal{Z}_0]=o(1)+\mathbf{E}[\mathbf{E}[y'1_{\{S'_{\tau'}\notin\partial D\}}\,|\,\mathcal{Z}_0,\operatorname{ext}_{r_1}\gamma]\mathbf{E}[y\,|\,\mathcal{Z}_0,\operatorname{ext}_{r_1}\gamma]\,|\,\mathcal{Z}_0]. \tag{3.59}$$

By the limit existence theorem (Theorem 3.21), given $\mathcal{Z}_0$ and $\operatorname{ext}_{r_1}\gamma$, near 0 the path γ_* is close in distribution to $\dot{\gamma}$ when r_1 is large. Consequently, Proposition 3.3 implies that $\mathbf{E}[|h(0)||\mathcal{Z}_0,\operatorname{ext}_{r_1}\gamma]-\lambda=o(1)$ as $r_1\to\infty$, which gives that

$$\mathbf{E}[y\,|\,\mathcal{Z}_0,\operatorname{ext}_{r_1}\gamma]=o(1).$$

Now (3.59) implies that

$$\mathbf{E}[yy'1_{\{S'_\tau \notin \partial D\}} \mid \mathcal{Z}_0] \to 0 \tag{3.60}$$

as $\operatorname{dist}(0, \partial D \cup \{v_0\}) \to \infty$. Lemmas 3.23 and 3.24 tell us that, for $r_0 < \infty$ fixed,

$$\mathbf{P}[S_\tau \notin \partial D, \operatorname{dist}(S_\tau, \partial D \cup \{v_0\}) < r_0] \to 0$$

as $\operatorname{dist}(v_0, \partial D) \to \infty$. Since there is nothing special about the vertex at 0, except for our assumption that $\operatorname{dist}(0, \partial D \cup \{v_0\})$ is large, we conclude from (3.60) that

$$\lim \mathbf{E}[yy'1_{\{S_\tau, S'_\tau \notin \partial D\}} \mid S_\tau] = 0$$

in probability, as $\operatorname{dist}(v_0, \partial D) \to \infty$. Now, equation (3.57) implies that $\mathbf{E}[X^2] \to 0$, since $yy'1_{\{S_\tau, S'_\tau \notin \partial D\}}$ is bounded. This gives (3.56) and completes the proof. □

Let F be the function that is equal to h_∂ on ∂D, λ on $V_+(\gamma)$, $-\lambda$ on $V_-(\gamma)$ and is discrete-harmonic on all other TG-vertices in D. Since $\mathbf{E}[h(v_0) \mid \gamma] = \mathbf{E}[h(S_\tau) \mid \gamma]$, Proposition 3.25 gives

$$\mathbf{E}[h(v_0) \mid \gamma] - F(v_0) \to 0 \tag{3.61}$$

in probability as $\operatorname{dist}(v_0, \partial D) \to \infty$.

We now need to generalize the proposition and (3.61) to apply when γ is replaced by an appropriate initial segment of γ.

Let T be some stopping time for γ started at x_∂ and let γ^T denote γ stopped at T. (Note that the relevant filtration here, the one generated by intitial segments of γ, only reveals the signs of h on vertices adjacent to these initial segments, but not the actual values of h.) Let z_T denote the vertex in $\partial D(\gamma^T)$ first visited by S, and let S^T denote the initial segment of S up to its first visit to z_T.

LEMMA 3.26. *Assume* (h), (D), (S) *and* (∂). *For every* $p_0 > 0$ *there is some* $s = s(p_0, \bar{\Lambda}) > 0$ *such that*

$$\mathbf{P}[\operatorname{dist}(z_T, \gamma \setminus \gamma^T) < s \operatorname{dist}(v_0, \partial D), z_T \notin \partial D] < p_0.$$

Note that we could rather easily prove the estimate with $\operatorname{dist}(z_T, \gamma \setminus \gamma^T; D(\gamma^T))$ instead of $\operatorname{dist}(z_T, \gamma \setminus \gamma^T)$, by the argument giving (3.48), but this is not sufficient for our purposes. The idea of the proof of the lemma is to first show that $S_\tau = z_T$ is usually not too unlikely given γ^T and S. Then Lemma 3.17 may be used in conjunction with the argument giving (3.48) to deduce the required result. Note that the event $S_\tau = z_T$ is the event that $\gamma \setminus \gamma^T$ is not adjacent to any vertex visited by S prior to z_T.

Proof. We first show that for every $\varepsilon>0$ there is a $p>0$ and an $R>0$, both depending only on ε and $\bar{\Lambda}$, such that

$$\mathbf{P}[\mathbf{P}[z_T = S_\tau \mid S, \gamma^T] < p] < \varepsilon \tag{3.62}$$

holds provided that $\operatorname{dist}(v_0, \partial D) > R$.

We choose $\delta = \delta(\varepsilon, \bar{\Lambda}) > 0$ very small. Set $r = \operatorname{dist}(v_0, \partial D)$, and assume that $r > 100\delta^{-2}$. Set further $D_T := D(\gamma^T)$. Let b_T be the point on ∂D_T near the tip of γ^T that is at equal distance from $V_+(\gamma^T)$ and $V_-(\gamma^T)$ along ∂D_T. Let ∂_+^T and ∂_-^T denote the two connected components of $\partial D_T \setminus \{y_\partial, b_T\}$ that have y_∂ and b_T as their endpoints, with ∂_+^T being the one containing vertices in ∂_+.

Let $\mathcal{A}_1$ be the event $\operatorname{dist}(z_T, \partial_+^T; D_T) \vee \operatorname{dist}(z_T, \partial_-^T; D_T) \geqslant 2\delta r$, let $\mathcal{A}_2$ be the event $\operatorname{diam}(S^T) < \delta^{-1} r$, let $\mathcal{A}_3$ be the event that the diameter of the segment of S^T after the first time at which it is distance at most $\delta^2 r$ from ∂D_T is less than $\frac{1}{2}\delta r$, and let $\mathcal{A}_4$ be the event $\operatorname{dist}(v_0, \gamma^T) \geqslant \delta^{1/2} r$.

Lemma 3.24 shows that if $\delta = \delta(\bar{\Lambda}, \varepsilon)$ is sufficiently small, then $\mathbf{P}[\neg\mathcal{A}_4] < \frac{1}{4}\varepsilon$. On the other hand, Lemma 2.1 implies that, by choosing δ sufficiently small, one can ensure that $\mathbf{P}[\neg\mathcal{A}_j \mid \gamma^T] < \frac{1}{4}\varepsilon$ for $j = 2, 3$. We now prove the same for $j = 1$. Assume that $\mathcal{A}_4$ holds. Let L be the set of points in D_T that lie on a path of diameter at most $4\delta r$ in $\overline{D}_T$ connecting ∂_+^T and ∂_-^T, and let D_* be the connected component of $D_T \setminus L$ that contains v_0 (we know that $v_0 \notin L$, since δ is small and $\mathcal{A}_4$ holds). By (3.55) applied to D_T in place of D, $\partial D_* \cap D_T$ may be covered by two balls of radius $8\delta r$. Thus, Lemma 2.1 shows that if $\delta = \delta(\varepsilon)$ is chosen sufficiently small, then $\mathbf{P}[S \text{ hits } L \text{ before } \partial D_T, \mathcal{A}_4] < \frac{1}{4}\varepsilon$. This implies that $\mathbf{P}[\neg\mathcal{A}_1, \mathcal{A}_4] < \frac{1}{4}\varepsilon$. Thus $\mathbf{P}[\neg\mathcal{A}] < \varepsilon$, where $\mathcal{A} := \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3 \cap A_4$.

We now complete the proof of inequality (3.62) by showing that the event

$$\mathbf{P}[S_\tau = z_T \mid S, \gamma^T] < p$$

is contained in $\neg\mathcal{A}$ if $p = p(\bar{\Lambda}, \delta) > 0$ is chosen sufficiently small. The latter is equivalent to showing that the random variable $\mathbf{P}[S_\tau = z_T \mid S, \gamma^T]$ is bounded away from zero on $\mathcal{A}$ by a function of δ and $\bar{\Lambda}$.

Suppose that $\mathcal{A}$ holds, and that $z_T \in \partial_+^T$. Let ξ_1 and ξ_2 be the two connected components of $\partial_+^T \setminus \{z_T\}$. The construction of Υ_n in the proof (3.48) shows that there is a path Υ_0 connecting ξ_1 and ξ_2 in $B(z_T, \delta r) \setminus B\big(z_T, \frac{1}{2}\delta r\big)$ that separates z_T from ∂_-^T in D_T such that Υ_0 is an $(s, \operatorname{diam}\Upsilon_0)$-barrier for (D, γ_T) for every $s \in (0, \delta']$, where $\delta' \in (0,1)$ is a universal constant. (The assumption that $\mathcal{A}_1$ holds is used here.) If S^T never visits a vertex adjacent to Υ_0, then Υ_0 separates S^T from ∂_-^T. In this situation, if $\gamma \setminus \gamma^T$ does not hit Υ_0, then S^T does not visit a vertex adjacent to it, and therefore $z_T = S_\tau$. Thus, Theorem 3.11 (or Remark 3.12) applies to give the needed lower bound on $\mathbf{P}[S_\tau = z_T \mid S, \gamma^T]$ when S^T does not visit a vertex adjacent to Υ_0.

Suppose now that S^T does visit vertices adjacent to Υ_0. We can then construct a path Υ whose image is Υ_0 as well as all the boundaries of hexagons visited by S^T that are not separated from ∂_-^T by Υ_0. Since we are assuming that $\mathcal{A}_2$ holds, $\operatorname{diam}\Upsilon\leqslant 2\delta^{-1}r$. As $\mathcal{A}_3$ holds, $\operatorname{dist}(\Upsilon\setminus\Upsilon_0,\partial D_T)>\frac{1}{2}\delta^2 r$ and therefore also $\operatorname{diam}\Upsilon_0>\frac{1}{2}\delta^2 r$. Consequently, Υ is a $\bigl(\frac{1}{2}\delta'\delta^3,2\delta^{-1}r\bigr)$-barrier. Now Theorem 3.11 and Remark 3.12 may be used again to give a similar lower bound on $\mathbf{P}[S_\tau=z_T\,|\,S,\gamma^T]$. As a similar argument applies when $S_\tau\in\partial_-^T$, the proof of (3.62) is now complete.

We now choose $\varepsilon=\frac{1}{9}p_0$ and take a $p>0$ and $R>0$ depending only on ε and $\bar\Lambda$ and satisfying (3.62). Let a be such that the estimate given in Lemma 3.17 holds with the p there replaced by $\frac{1}{9}p_0p$. Let $\delta=\delta(p_0,\bar\Lambda)>0$ be sufficiently small so that $\delta^{-1}>R\vee a$. We assume that $r>10\delta^{-5}$. For $z\in D$, let $\mathcal{J}_z$ denote the event that there are more than two disjoint arcs in γ joining the two circles $\partial B(z,\delta^4 r)$ and $\partial B(z,\delta^5 r)$, and let $\mathcal{J}_z^T$ denote the event that there are more than two such arcs in γ^T. By the choice of a and δ, the probability that $\operatorname{dist}(S_\tau,\partial D\cup\{v_0\})>4\delta^4 r$ and $\mathcal{J}_{S_\tau}$ holds is at most $\frac{1}{9}p_0p$. Consequently, the same bound applies for the probability that $\operatorname{dist}(z_T,\partial D\cup\{v_0\})>4\delta^4 r$, $z_T=S_\tau$ and $\mathcal{J}_{z_T}^T$ holds. Thus, as the events $\operatorname{dist}(z_T,\partial D\cup\{v_0\})>4\delta^4 r$ and $\mathcal{J}_{z_T}^T$ are (γ^T,S)-measurable,

$$\begin{aligned}\tfrac{1}{9}p_0p&\geqslant\mathbf{P}[z_T=S_\tau,\operatorname{dist}(z_T,\partial D\cup\{v_0\})>4\delta^4 r,\mathcal{J}_{z_T}^T]\\&=\mathbf{E}[\mathbf{P}[z_T=S_\tau,\operatorname{dist}(z_T,\partial D\cup\{v_0\})>4\delta^4 r,\mathcal{J}_{z_T}^T\,|\,\gamma^T,S]]\\&=\mathbf{E}[\mathbf{P}[z_T=S_\tau\,|\,\gamma^T,S]1_{\{\operatorname{dist}(z_T,\partial D\cup\{v_0\})>4\delta^4 r\}}1_{\mathcal{J}_{z_T}^T}].\end{aligned}$$

By (3.62) and our choice of ε, we therefore have

$$\mathbf{P}[\operatorname{dist}(z_T,\partial D\cup\{v_0\})>4\delta^4 r,\mathcal{J}_{z_T}^T]\leqslant\tfrac{2}{9}p_0.\tag{3.63}$$

Let $\mathcal{H}$ denote the event $(\gamma\setminus\gamma^T)\cap\partial B(z_T,\delta^5 r)\neq\varnothing$. Now condition on γ^T and S such that $\operatorname{dist}(z_T,\partial D)>\delta r$ and $\neg\mathcal{J}_{z_T}^T$ holds. Suppose also that $\operatorname{dist}(z_T,b_T)>\delta^3 r$. Then there are precisely two connected components of $B(z_T,\delta^4 r)\cap D_T$ that intersect $\partial B(z_T,\delta^5 r)$. Let $w_1,w_2\in\partial B(z_T,\delta^5 r)\cap D_T$ be points in each of these two connected components. By constructing barriers as in the proof of (3.48), it is easy to see that if $\delta=\delta(p_0,\bar\Lambda)$ is sufficiently small, then

$$\mathbf{P}[\operatorname{dist}(w_j,\gamma\setminus\gamma^T;D(\gamma^T))\leqslant\delta^4 r\,|\,\gamma^T,S]\leqslant\tfrac{1}{9}p_0$$

for $j=1,2$. Now note that if $\operatorname{dist}(w_j,\gamma\setminus\gamma^T;D(\gamma^T))>\delta^4 r$ for $j=1,2$, then $\neg\mathcal{H}$ holds. Consequently,

$$\mathbf{P}[\mathcal{H},\neg\mathcal{J}_{z_T}^T,\operatorname{dist}(z_T,b_T)>\delta^3 r,\operatorname{dist}(z_T,\partial D)>\delta r]\leqslant\tfrac{2}{9}p_0.$$

We combine this with (3.63), and get

$$\mathbf{P}[\mathcal{H}, \operatorname{dist}(v_0, z_T) > 4\delta^4 r, \operatorname{dist}(z_T, b_T) > \delta^3 r, \operatorname{dist}(z_T, \partial D) > \delta r] \leqslant \tfrac{4}{9} p_0. \tag{3.64}$$

Since $r = \operatorname{dist}(v_0, \partial D)$, provided that we take $\delta = \delta(p_0, \bar{\Lambda}) > 0$ sufficiently small, Lemma 3.24 gives $\mathbf{P}[\operatorname{dist}(v_0, z_T) \leqslant \delta r] < \frac{1}{9} p_0$, Lemma 2.1 gives $\mathbf{P}[\operatorname{dist}(v_0, z_T) > \delta r, \operatorname{dist}(z_T, b_T) \leqslant \delta^2 r] < \frac{1}{9} p_0$ and (3.53) gives $\mathbf{P}[\operatorname{dist}(z_T, \partial D) \leqslant \delta r, z_T \notin \partial D] < \frac{1}{9} p_0$. These last three estimates may be combined with (3.64), to yield $\mathbf{P}[\mathcal{H}, z_T \notin \partial D] \leqslant \frac{7}{9} p_0$, which completes the proof. □

We now prove the analog of (3.61) with γ^T replacing γ. Let F_T denote the function that is $+\lambda$ on $V_+(\gamma^T)$, $-\lambda$ on $V_-(\gamma^T)$, equal to h_∂ on TG-vertices in ∂D, and is discrete-harmonic at all other vertices in $\overline{D}$.

PROPOSITION 3.27. *Assume* (h), (D), (S) *and* (∂). *Then*

$$\mathbf{E}[h(v_0) \,|\, \gamma^T] - F_T(v_0) \to 0$$

in probability as $\operatorname{dist}(v_0, \partial D) \to \infty$ *while* $\bar{\Lambda}$ *is held fixed.*

Proof. Fix $\varepsilon > 0$ and set $r := \operatorname{dist}(v_0, \partial D)$. We have, by the heights interface continuity (Proposition 3.3),

$$\big|\mathbf{E}[h(z_T) \,|\, \gamma, z_T] - \mathbf{E}[h(z_T) \,|\, \gamma^T, z_T]\big| < \varepsilon$$

if $\operatorname{dist}(z_T, \gamma \setminus \gamma_T) > R_0$, where $R_0 = R_0(\varepsilon, \bar{\Lambda})$. Therefore, Lemma 3.26 with $p_0 = \varepsilon / O_{\bar{\Lambda}}(1) > 0$ gives

$$\mathbf{E}\big[\big|\mathbf{E}[h(z_T) \,|\, \gamma, z_T] - \mathbf{E}[h(z_T) \,|\, \gamma^T, z_T]\big|\big] < 2\varepsilon \tag{3.65}$$

when $r > s^{-1} R_0$, and s is as given by the lemma.

(Note that $\mathbf{E}[h(z_T) | \gamma, z_T] = \mathbf{E}[h(z_T) | \gamma^T, z_T]$ when $z_T \in \partial D$.)

In the following, we will use a parameter $\delta > 0$. The notation $o(1)$ will be shorthand for any quantity g satisfying $\lim_{\delta \to 0} \lim_{r \to \infty} |g| = 0$ while $\bar{\Lambda}$ is fixed. Let Z be a maximal set of TG-vertices in $D \cap B(v_0, 2\delta^{-1} r)$ such that the distance between any two such vertices is at least $\delta^2 r$ and the distance between any such vertex to ∂D is at least $\delta^3 r$. Then $|Z| = O(\delta^{-6})$. By (3.61) (with each $u \in Z$ in place of v_0), we therefore have

$$\mathbf{P}[\text{there exists } u \in Z : \mathbf{E}[h(u) - F(u) \,|\, \gamma]^2 > \varepsilon] = o(1). \tag{3.66}$$

Let t_0 be the first $t \in \mathbb{N}$ such that $\operatorname{dist}(S_t, \partial D(\gamma^T)) \leqslant \delta r$. By Lemma 2.1,

$$\mathbf{P}[\operatorname{dist}(z_T, S_{t_0}; D) \geqslant \delta^{1/2} r] = o(1). \tag{3.67}$$

By (3.54),

$$\mathbf{P}[\operatorname{dist}(z_T, \gamma; D) < \delta^{1/3} r, z_T \in \partial D] = o(1),$$

while Lemma 3.26 gives

$$\mathbf{P}[\mathrm{dist}(z_T, \gamma\backslash\gamma_T) < \delta^{1/3} r, z_T \notin \partial D] = o(1).$$

Thus,

$$\mathbf{P}[\mathrm{dist}(z_T, \gamma\backslash\gamma_T; D) < \delta^{1/3} r] = o(1).$$

This and (3.67) imply that

$$\mathbf{P}\big[\mathrm{dist}(S_{t_0}, \gamma\backslash\gamma^T; D) < \tfrac{1}{2}\delta^{1/3} r\big] = o(1). \tag{3.68}$$

Since $\mathrm{dist}(S_{t_0}, \partial D(\gamma^T)) < \delta r$, this and Lemma 2.1 imply that with probability $1-o(1)$ the L^1 norm of the difference between the discrete-harmonic measure from S_{t_0} on $\partial D(\gamma^T)$ and the discrete-harmonic measure from S_{t_0} on $\partial D(\gamma)$ is $o(1)$. Because $\mathbf{E}[h(S_{t_0})|\gamma, S_{t_0}]$ is the average of $\mathbf{E}[h(z)|\gamma]$, where z is selected according to the discrete-harmonic measure on $\partial D(\gamma)$ from S_{t_0} and similarly for γ^T, we conclude from the above and (3.65) that

$$\mathbf{E}[h(S_{t_0}) \,|\, \gamma^T, S_{t_0}] - \mathbf{E}[h(S_{t_0}) \,|\, \gamma, S_{t_0}] = o(1)$$

in probability. Now, Lemma 2.1 implies that $\mathbf{P}[\mathrm{dist}(S_{t_0}, v_0) > \delta^{-1} r] = o(1)$. On the event $\mathrm{dist}(S_{t_0}, v_0) \leqslant \delta^{-1} r$, fix some $z_0 \in Z$ within distance $\delta^2 r$ from S_{t_0} (if there is more than one such z_0, let z_0 be chosen uniformly at random among these given (h, γ, S)). By the discrete Harnack principle (Lemma 2.2) and (3.68), we have

$$\mathbf{E}[h(S_{t_0}) \,|\, \gamma, S_{t_0}] - \mathbf{E}[h(z_0) \,|\, \gamma, z_0] = o(1)$$

in probability, which in conjunction with (3.66) yields $\mathbf{E}[h(S_{t_0})|\gamma, S_{t_0}] - F(z_0) = o(1)$. The discrete Harnack principle now implies that $F(z_0) - F(S_{t_0}) = o(1)$ in probability, and (3.68) gives $F(S_{t_0}) - F_T(S_{t_0}) = o(1)$ in probability. Consequently,

$$\mathbf{E}[h(S_{t_0}) \,|\, \gamma^T, S_{t_0}] - F_T(S_{t_0}) = o(1)$$

in probability. Since $\mathbf{E}[h(\cdot)|\gamma^T]$ and F_T are discrete-harmonic in $D(\gamma^T)$, the proposition follows. □

We can now prove the height gap theorem.

THEOREM 3.28. *Assume* (h), (D), (S) *and* (∂). *As above, let* T *denote a stopping time for* γ, *let* $\gamma^T := \gamma[0, T]$, *let* D_T *be the complement in* D *of the closed triangles meeting* $\gamma[0, T)$. *Let* h_T *denote the restriction of* h *to* $V \cap \partial D_T$ *and let* v_0 *be some vertex in* D. *Then*

$$\mathbf{E}[h(v_0) \,|\, \gamma^T, h_T] - F_T(v_0) \to 0$$

in probability as $\mathrm{dist}(v_0, \partial D) \to \infty$ *while* $\bar{\Lambda}$ *is held fixed, where* F_T *is as in Proposition* 3.27.

Proof. Set

$$X := \mathbf{E}[h(v_0) \,|\, \gamma^T, h_T] - \mathbf{E}[h(v_0) \,|\, \gamma^T].$$

By Proposition 3.27, it suffices to show that $X \to 0$ in probability as $\operatorname{dist}(v_0, \partial D) \to \infty$. For $v \in V \cap \partial D_T$, let a_v denote the conditional probability that a simple random walk started at v_0 first hits ∂D_T at v, given γ^T. Then

$$X = \sum_{v \in V \cap \partial D_T} a_v (h(v) - \mathbf{E}[h(v) \,|\, \gamma^T]).$$

Consequently,

$$X^2 = \sum_{v,u} a_v a_u (h(v) - \mathbf{E}[h(v) \,|\, \gamma^T])(h(u) - \mathbf{E}[h(u) \,|\, \gamma^T]).$$

Now Corollary 3.5 implies that it suffices to show that for every $r > 0$,

$$\sum_{v,u} a_v a_u 1_{\{|v-u|<r\}} 1_{\{v,u \notin V_\partial\}} \to 0$$

in probability. This follows by Lemmas 2.1 and 3.24. □

4. Recognizing the driving term

In this section we use a technique introduced in [LSW4] and used again in [SS] in order to show that the driving term for the Loewner evolution given by the DGFF interface with boundary values $-a$ and b converges to the driving term of SLE$(4; a/\lambda - 1, b/\lambda - 1)$ if $a, b \in [-\Lambda_0, \bar{\Lambda}]$. The reader unfamiliar with this method is advised to first learn the technique from [SS, §4] or [LSW4, §3.3]. The account in [SS] is closer to the present setup and somewhat simpler, but some parts of the argument there are referred back to [LSW4].

The present argument is more involved than those of the above mentioned papers, because we prove convergence to an instance of SLE$(\varkappa; \varrho_1, \varrho_2)$ rather than just plain SLE(4). The main added difficulty comes from the fact that the drift term in SLE$(\varkappa; \varrho_1, \varrho_2)$ becomes unbounded as W_t comes close to the force points. These difficulties disappear if $a = b = \lambda$, in which case the convergence is to ordinary SLE and the argument giving the convergence of the driving term to scaled Brownian motion is easily established with minor adaptations of the established method. We therefore forego dwelling on this simpler case, and move on to the more general setting, assuming that the reader is already familiar with the fundamentals of the method.

4.1. About the definition of SLE(κ; ϱ_1, ϱ_2)

Throughout this section, given a Loewner evolution defined by a continuous W_t, we let x_t and y_t be defined as in §1.4 by

$$x_t := \sup\{g_t(x) : x < 0 \text{ and } x \notin K_t\} \quad \text{and} \quad y_t := \inf\{g_t(x) : x > 0 \text{ and } x \notin K_t\},$$

and we make use of the definition of SLE(κ; ϱ_1, ϱ_2) by means of the SDE (1.5). As we mentioned in §1.4, some subtlety is involved in extending the definition of SLE(κ; ϱ_1, ϱ_2) beyond times when W_t hits the force points, and in starting the process from the natural initial values $x_0 = W_0 = y_0 = 0$. This is closely related to the issues involved in defining the Bessel process, which we presently recall.

The Bessel process Z_t of dimension $\delta > 0$ and initial value $x \neq 0$ satisfies the SDE

$$dZ_t = \frac{\delta - 1}{2Z_t}\, dt + dB_t, \quad Z_0 = x, \tag{4.1}$$

which we also write in integral form as

$$Z_t = x + \int_0^t \frac{\delta - 1}{2Z_s}\, ds + B_t - B_0, \tag{4.2}$$

up until the first time t for which $Z_t = 0$. When defining Z_t for all times, this SDE is awkward to work with directly since the drift blows up whenever Z_t gets close to zero (and some of the standard existence and uniqueness theorems for SDE solutions, as given, e.g., in [RY], do not apply in this situation). However, for every $\delta > 0$, the square of the Bessel process Z_t^2 turns out to satisfy an SDE whose drift remains bounded and for which existence and uniqueness of solutions follow easily from standard theorems. For this reason, many authors construct the Bessel process by first defining the square of the Bessel process via an SDE that it satisfies and then taking its square root [RY]. (Recall also that when $\delta \leqslant 1$ the Bessel process itself does not satisfy (4.2) at all without a principal value correction. Even when $1 < \delta < 2$, which, as we will see below, is the case that corresponds to SLE(κ; ϱ) that hit the boundary and can be continued after hitting the boundary, the solution to (4.2) is not unique unless we restrict attention to non-negative solutions.)

The formal definition for SLE(κ; ϱ) with one force point (i.e., $\varrho_1 = \varrho$ and $\varrho_2 = 0$) was given in [LSW3]. It was observed there that in this case, (1.5) implies that the process $W_t - x_t$ satisfies the same SDE as the Bessel process of dimension $\delta = 1 + 2(\varrho + 2)/\kappa$ up until the first time t for which $W_t = x_t$. Thus, to define SLE(κ; ϱ), the paper [LSW3] starts with a constant multiple of a Bessel process Z_t of the appropriate dimension and defines the evolution of the force point x_t by $x_t = x_0 + \int_0^t (2/Z_s)\, ds$ and the driving term by $W_t = x_t + Z_t$.

Defining SLE($\varkappa;\varrho_1,\varrho_2$) is a slightly more delicate matter since neither W_t-x_t nor y_t-W_t is exactly a Bessel process (although each one is quite close to a Bessel process when the other force point is relatively far away). Although this is not a very difficult issue, it seems that there does not yet exist, in the literature, an adequate definition of SLE($\varkappa;\varrho_1,\varrho_2$) that is valid beyond the time that the driving term hits a force point. Since we prove the convergence to SLE($\varkappa;\varrho_1,\varrho_2$), we have to define it.

The approach we adopt is basically similar to the way in which the Bessel process (and hence SLE($\varkappa;\varrho$)) is usually defined: we pass to a coordinate system in which the corresponding SDE becomes tractable. We will describe the coordinate change we use in §4.2. Within this new coordinate system, we then prove the convergence of the Loewner driving parameters of our discrete processes to those of the corresponding SLE($\varkappa;\varrho_1,\varrho_2$) in §4.3 and §4.4. Then §4.5 describes the reverse coordinate transformation and use it to give a formal definition of SLE($\varkappa;\varrho_1,\varrho_2$), Definition 4.14.

We remark that there are many equivalent ways to define SLE($\varkappa;\varrho_1,\varrho_2$) (for example, one can probably show directly that (1.5) has a unique strong solution for which $x_t\leqslant W_t\leqslant y_t$ for all t), but ours seems most efficient given that the coordinate change also simplifies the proofs in §4.3 and §4.4.

4.2. A coordinate change

In this subsection, we recall a different coordinate system for Loewner evolutions, which is virtually identical to the setup used in [LSW2, §3]. Suppose that $\gamma\colon[0,\infty)\to\overline{\mathbb{H}}$ is a continuous simple path that starts at $\gamma(0)=0$, does not hit $\mathbb{R}\setminus\{0\}$, satisfies $\lim_{t\to\infty}|\gamma(t)|=\infty$ and is parameterized by half-plane capacity from ∞. Let $g_t\colon\mathbb{H}\setminus\gamma[0,t]\to\mathbb{H}$ be the conformal map satisfying the hydrodynamic normalization at ∞, let $W_t=g_t(\gamma(t))$ be the corresponding Loewner driving term. Loewner's theorem says that g_t satisfies Loewner's chordal equation (1.3). Now we introduce a 1-parameter family of maps $G_t^*\colon\mathbb{H}\setminus\gamma[0,t]\to\mathbb{H}$ satisfying the normalization for $t>0$,

$$G_t^*(\infty)=\infty,\quad G_t^*((0,\infty))=(1,\infty)\quad\text{and}\quad G_t^*((-\infty,0))=(-\infty,-1).$$

That is,

$$G_t^*(z)=\frac{2g_t(z)-x_t-y_t}{y_t-x_t},\tag{4.3}$$

where x_t and y_t (as defined earlier) are the two images under g_t of 0 and $x_t<y_t$. Set

$$W_t^*:=G_t^*(\gamma(t))=\frac{2W_t-x_t-y_t}{y_t-x_t}.$$

By differentiating (4.3) and using (1.3) and (1.4) it is immediate to verify that G_t^* satisfies

$$\frac{dG_t^*(z)}{dt} = \frac{8}{(y_t - x_t)^2} \frac{1 - G_t^*(z)^2}{(G_t^*(z) - W_t^*)(1 - (W_t^*)^2)}.$$

We now define a new time parameter

$$s(t) = \log(y_t - x_t) = \log 2 - \log(G_t^*)'(\infty).$$

It is easy to verify that s is continuous and monotone increasing and $s((0,\infty)) = (-\infty, \infty)$. Set $G_s = G_t^*$ and $\widetilde{W}_s = W_t^*$ when $s = s(t)$. Differentiation gives

$$ds = \frac{8(y_t - x_t)^{-2} dt}{1 - (W_t^*)^2}. \tag{4.4}$$

Consequently, this change of time variable allows us to write the ODE satisfied by G as

$$\frac{dG_s(z)}{ds} = \frac{1 - G_s(z)^2}{G_s(z) - \widetilde{W}_s}, \tag{4.5}$$

where all the terms come from the new coordinate system. Later, in §4.5, we explain how to go back to the standard chordal coordinate system.

4.3. The Loewner evolution of the DGFF interface

In addition to our previous assumptions (h) and (D) about the domain D and the boundary conditions, we now add the assumption that

(ab) there are constants a and b such that $h_\partial = b$ on ∂_+, $h_\partial = -a$ on ∂_- and

$$\min\{a, b\} > -\Lambda_0,$$

where $\Lambda_0 > 0$ is given by Lemma 3.9 with $\bar{\Lambda} := \max\{|a|, |b|\}$.

In this case, clearly (∂) holds. In the following, a and b will be considered as constants, and the dependence of various constants on a and b will sometimes be suppressed (for example, when using the $O(\cdot)$ notation).

Let $\phi: D \to \mathbb{H}$ be a conformal map that corresponds ∂_+ with the positive real ray. Let γ be the zero-height interface of h joining the endpoints of ∂_+, and let γ^ϕ denote the image of γ under ϕ. Now, γ^ϕ satisfies the assumptions in the previous subsection. Consequently, we may parameterize γ^ϕ according to the time parameter $s = s(t)$ and consider the conformal maps $G_s: \mathbb{H} \setminus \gamma^\phi(-\infty, s] \to \mathbb{H}$ as defined in §4.2. As above, we set

$\widetilde{W}_s = G_s(\gamma^\phi(s))$ and have the differential equation (4.5). Our goal now is to determine the limit of the law of $\widetilde{W}$ as $\mathrm{rad}_{\phi^{-1}(i)}(D)\to\infty$. Set for $x\in[-1,1]$,

$$q_1(x) := 2(1-x^2) \quad \text{and} \quad q_2(x) := -\frac{\lambda+a}{2\lambda}(x-1) - \frac{\lambda+b}{2\lambda}(x+1). \tag{4.6}$$

We extend the definitions of q_1 and q_2 to all of $\mathbb{R}$ by taking each q_j to be constant in each of the two intervals $(-\infty,-1]$ and $[1,\infty)$. Consider the SDE

$$dY_s = q_2(Y_s)\,ds + q_1(Y_s)^{1/2}\,dB_s, \tag{4.7}$$

where B is a standard 1-dimensional Brownian motion. A weak solution is known to exist (see [KS, §5.4.D]). We also recall that the weak solution is strong and pathwise unique (see [RY, §IX, Theorems 1.7 and 3.5]).

THEOREM 4.1. *There is a time-stationary solution $Y\colon(-\infty,\infty)\to[-1,1]$ of (4.7). Moreover, for every finite $S>1$ and $\varepsilon>0$, there is an $R_0=R_0(S,\varepsilon)$ such that if $R:=\mathrm{rad}_{\phi^{-1}(i)}(D)>R_0$ and the assumptions* (h), (D) *and* (ab) *hold, then there is a coupling of Y_s with h such that*

$$\mathbf{P}[\sup\{|Y_s-\widetilde{W}_s| : s\in[-S,S]\}>\varepsilon] < \varepsilon.$$

The following proposition is key in the proof of the theorem. In essence, it states that $\widetilde{W}_s$ satisfies a discrete version of (4.7). Let $\mathcal{F}_s$ be the σ-field generated by $(\widetilde{W}_r : r\leqslant s)$. (Note that, although the filtration defining $\widetilde{W}_s$ is discrete, there is no problem in considering $\mathcal{F}_s$ for arbitrary s, though the behavior of $\widetilde{W}_r$ for r in some neighborhood of s might be determined by $\mathcal{F}_s$.)

PROPOSITION 4.2. *Assume* (h), (D) *and* (ab). *Fix some $S>1$ large and some $\delta,\eta>0$ small. There is a constant $C>0$, depending only on a, b and S, and there is a function $R_0=R_0(S,\delta,\eta)$, depending only on a, b, S, δ and η, such that the following holds. If $R:=\mathrm{rad}_{\phi^{-1}(i)}(D)>R_0$ and s_0 and s_1 are two stopping times for $\widetilde{W}_s$ such that almost surely $-S\leqslant s_0\leqslant s_1\leqslant S$, $\Delta s:=s_1-s_0\leqslant\delta^2$ and $\sup_{s\in[s_0,s_1]}|\widetilde{W}_s-\widetilde{W}_{s_0}|\leqslant\delta$, then the following two estimates hold with probability at least $1-\eta$:*

$$\bigl|\mathbf{E}[\Delta\widetilde{W} - q_2(\widetilde{W}_{s_0})\Delta s \,|\, \mathcal{F}_{s_0}]\bigr| \leqslant C\delta^3, \tag{4.8}$$

$$\bigl|\mathbf{E}[(\Delta\widetilde{W})^2 - q_1(\widetilde{W}_{s_0})\Delta s \,|\, \mathcal{F}_{s_0}]\bigr| \leqslant C\delta^3, \tag{4.9}$$

where $\Delta\widetilde{W}:=\widetilde{W}_{s_1}-\widetilde{W}_{s_0}$.

To prepare for the proof of the proposition, we need the following easy lemma. The first two statements in this lemma should be rather obvious to anyone with a solid background on conformal mappings.

LEMMA 4.3. *Set* $c_1=c_1(S)=100e^S$. *There are finite constants* $c_2=c_2(S)>0$ *and* $R_0=R_0(S)>0$, *depending only on* S, *such that if* $R:=\mathrm{rad}_{\phi^{-1}(i)}(D)>R_0$ *and if* $z\in\mathbb{H}$ *satisfies* $5c_1\geqslant \mathrm{Im}\, z\geqslant c_1\geqslant|\mathrm{Re}\, z|$, *then the following holds true*:

(1) $\mathrm{rad}_{\phi^{-1}(z)}(D)>c_2R$;

(2) *there is a* TG*-vertex* $v\in D$ *satisfying* $|\phi(v)-z|<\frac{1}{100}|z|$;

(3) $\mathrm{Im}\, G_s(z)\geqslant\frac{1}{2}e^{-s}c_1$ *for* $s\in[-S,S]$ (*and in particular*, $G_s(z)$ *is well defined in that range*);

(4) $|G_s(z)-2e^{-s}z|\leqslant 2$ *for* $s\in[-S,S]$.

Proof. Consider the conformal map $\psi(z)=(z-i)/(z+i)$ from $\mathbb{H}$ onto the unit disk $\mathbb{U}$ taking i to 0 and set $f=(\psi\circ\phi)^{-1}$. The Schwarz lemma applied to the map $z\mapsto f^{-1}(f(0)+Rz)$ restricted to $\mathbb{U}$ gives $1/|f'(0)|=|(f^{-1})'(f(0))|\leqslant 1/R$. Thus $|f'(0)|\geqslant R$. For a fixed c_1 the set of possible z is a compact subset of $\mathbb{H}$, and its image under ψ is a compact subset of $\mathbb{U}$. Consequently, the Koebe distortion theorem (see, e.g., [P, Theorem 1.3]) implies that $|f'(\psi(z))|\geqslant c_2'R$ for some c_2' depending only on c_1. Now the Koebe $\frac{1}{4}$-theorem (see, e.g., [P, Corollary 1.4]) gives that $\mathrm{rad}_{f(\psi(z))}(D)>c_2R$ for some c_2 depending on c_1. This takes care of statement (1).

Let B be the open disk of radius $\frac{1}{200}|z|$ about z. Clearly, $B\subset\mathbb{H}$. We conclude from $|f'(\psi(z))|\geqslant c_2'R$ that $|(\phi^{-1})'(z)|>c_2''R$ for some c_2'' depending only on c_1. Thus, the Koebe $\frac{1}{4}$-theorem implies that

$$\operatorname*{rad}_{\phi^{-1}(z)}(\phi^{-1}(B))\geqslant\tfrac{1}{4}c_2''\,\mathrm{rad}(B)R.$$

Consequently, statement (2) holds once $R_0>4/c_2''\,\mathrm{rad}(B)$. This takes care of (2), because $1/\mathrm{rad}(B)$ is bounded by a function of c_1.

It is easy to check that (3) follows from (4). It remains to prove the latter. Let x_t and y_t be as in §4.2. Note that $x_t<W_t<y_t$ for all $t>0$ and $\lim_{t\searrow 0}x_t=W_0=0=\lim_{t\searrow 0}y_t$. Therefore, (1.4) implies that $x_t<0<y_t$ for all $t>0$. By (1.4),

$$\partial_t(y_t-x_t)=\frac{2(y_t-x_t)}{(y_t-W_t)(W_t-x_t)}\geqslant\frac{8}{y_t-x_t}.$$

Therefore, $\partial_t((y_t-x_t)^2)\geqslant 16$, which gives

$$e^{2s(t)}\geqslant 16t. \tag{4.10}$$

Observe, by (1.3), that

$$\partial_t((\mathrm{Im}\, g_t(z))^2)\geqslant -4.$$

Thus, $(\mathrm{Im}\, g_t(z))^2\geqslant(\mathrm{Im}\, z)^2-4t\geqslant c_1^2-4t$. Another appeal to (1.3) now gives

$$|g_t(z)-z|\leqslant\frac{2t}{\sqrt{c_1^2-4t}}.$$

By (4.3) and the definition of $s(t)$, the above gives

$$|G_s(z)-2e^{-s}z| \leqslant \left|\frac{x_t+y_t}{y_t-x_t}\right| + \frac{2te^{-s}}{\sqrt{c_1^2-4t}}.$$

Now, the first summand on the right-hand side is at most 1, because $y_t>0>x_t$. The second summand is also at most 1 in the range $s\in[-S,S]$, by (4.10) and the choice of c_1. This completes the proof of the lemma. □

Proof of Proposition 4.2. With the notation of Lemma 4.3, let $z_j:=2c_1i+\frac{1}{2}c_1j$ for $j=0,1$, and let v_j be a TG-vertex satisfying condition (2) of the lemma with z_j in place of z. Then $z_j':=\phi(v_j)$ satisfies in turn the assumptions required for z in the lemma. For $k=0,1$, note that there is a stopping time T_k for γ such that $\phi\circ\gamma(T_k)=\gamma^\phi(s_k)$. Fix $j\in\{0,1\}$ and set

$$X_k = X_k(j) := \mathbf{E}[h(v_j)\,|\,\mathcal{F}_{s_k}], \quad k=0,1.$$

Clearly,

$$\mathbf{E}[X_1\,|\,\mathcal{F}_{s_0}] = X_0. \tag{4.11}$$

Recall the definition of the function F_T from Proposition 3.27. Let $\mathcal{A}_k$ be the event $|X_k-F_{T_k}(v_j)|\geqslant\delta^5$. By that proposition and the fact that z_j' satisfies condition (1) of Lemma 4.3, if R is chosen sufficiently large then $\mathbf{P}[\mathcal{A}_k]<\frac{1}{4}\eta\delta^5$. Since $F_{T_k}(v_j)=O_{a,b}(1)$, and likewise $X_k=O_{a,b}(1)$ by (3.2), we get

$$\left|\mathbf{E}[X_1-F_{T_1}(v_j)\,|\,\mathcal{F}_{s_0}]\right| \leqslant \delta^5 + O_{a,b}(1)\mathbf{P}[\mathcal{A}_1\,|\,\mathcal{F}_{s_0}].$$

Let $\tilde{\mathcal{A}}$ be the event that $\mathbf{P}[\mathcal{A}_1\,|\,\mathcal{F}_{s_0}]>\delta^5$. Then $\mathbf{P}[\tilde{\mathcal{A}}]<\frac{1}{4}\eta$ (since we are assuming $\mathbf{P}[\mathcal{A}_1]<\frac{1}{4}\eta\delta^5$) and we have

$$\left|\mathbf{E}[X_1-F_{T_1}(v_j)\,|\,\mathcal{F}_{s_0}]\right| \leqslant O_{a,b}(\delta^5) \quad \text{on } \neg\tilde{\mathcal{A}}.$$

Thus we have, from (4.11),

$$\mathbf{E}[F_{T_1}(v_j)\,|\,\mathcal{F}_{s_0}] - F_{T_0}(v_j) = O_{a,b}(\delta^5) \quad \text{on } \neg(\tilde{\mathcal{A}}\cup\mathcal{A}_0). \tag{4.12}$$

Let H_k be the bounded function that is harmonic (not discrete-harmonic) in $D\setminus\gamma[0,T_k]$, has boundary values b on ∂_+, $-a$ on ∂_-, $+\lambda$ on the right-hand side of $\gamma[0,T_k]$, and $-\lambda$ on the left-hand side of $\gamma[0,T_k]$. We claim that the difference $H_k(v_j)-F_{T_k}(v_j)$ is small if $\operatorname{dist}(v_j,\partial D\cup\gamma[0,T_k])$ is large. Indeed, this easily follows by coupling the simple random walk on TG to stay with high probability relatively close to a Brownian motion and using (3.54) and Lemmas 3.23 and 2.1 to show that with high probability the boundary value sampled by the hitting point of the Brownian motion is the same as that sampled

by the hitting vertex of the simple random walk. Now, Lemma 3.24 guarantees that if R is sufficiently large, then with high probability $\operatorname{dist}(v_j, \partial D \cup \gamma[0,T_k])$ is large as well. Consequently, if R is chosen sufficiently large, we have $\mathbf{P}[|H_k(v_j)-F_{T_k}(v_j)|\geqslant\delta^5]<\frac{1}{4}\eta\delta^5$. Let $\mathcal{B}_k$ be the event $|H_k(v_j)-F_{T_k}(v_j)|\geqslant\delta^5$, let $\widetilde{\mathcal{B}}$ be the event $\mathbf{P}[\mathcal{B}_1\,|\,\mathcal{F}_{s_0}]\geqslant\delta^5$ and let $\bar{\mathcal{A}}:=\mathcal{A}_0\cup\tilde{\mathcal{A}}\cup\mathcal{B}_0\cup\widetilde{\mathcal{B}}$. Note that $\mathbf{P}[\bar{\mathcal{A}}]<\eta$. The above proof of (4.12) from (4.11) now gives

$$\mathbf{E}[H_1(v_j)\,|\,\mathcal{F}_{s_0}]-H_0(v_j)=O_{a,b}(\delta^5)\quad\text{on }\neg\bar{\mathcal{A}}. \tag{4.13}$$

Now, the point is that $H_k(v_j)$ can easily be expressed analytically in terms of $Z_k=Z_{k,j}:=G_{s_k}(z_j')=G_{s_k}(\phi(v_j))$ and $\widetilde{W}_{s_k}$. Indeed, conformal invariance implies that the harmonic measure of ∂_+ in $D\setminus\gamma[0,T_k]$ from v_j is the same as the harmonic measure of $[1,\infty)$ from Z_k, which is $1-\arg(Z_k-1)/\pi$, because $G_s\circ\phi$ corresponds ∂_+ with $[1,\infty)$. Likewise, the harmonic measure of the right-hand side of $\gamma[0,T_k]$ is

$$\frac{\arg(Z_k-1)-\arg(Z_k-\widetilde{W}_{s_k})}{\pi}.$$

Similar expressions hold for the harmonic measure of the left-hand side of $\gamma[0,T_k]$ and of ∂_-. These give

$$\pi(H_k(v_j)-b)=(\lambda-b)\arg(Z_k-1)-2\lambda\arg(Z_k-\widetilde{W}_{s_k})+(\lambda-a)\arg(Z_k+1). \tag{4.14}$$

Recall that $Z_k=G_{s_k}(z_j')$. By (4.5), we have in the interval $s\in[s_0,s_1]$,

$$G_s(z_j')=Z_0+\int_{s_0}^{s}\frac{1-G_r(z_j')^2}{G_r(z_j')-\widetilde{W}_r}\,dr. \tag{4.15}$$

Note that conditions (3) and (4) of Lemma 4.3 imply that the integrand is $O_S(1)$. Therefore $G_s(z_j')-Z_0=O_S(\Delta s)=O_S(\delta^2)$ for $s\in[s_0,s_1]$. Moreover, we have $|\widetilde{W}_s-\widetilde{W}_{s_0}|\leqslant\delta$ in that range. Thus, it follows from (4.15) and condition (3) of the lemma that

$$Z_1-Z_0=\Delta s\left(\frac{1-Z_0^2}{Z_0-\widetilde{W}_{s_0}}+O_S(\delta)\right)=\Delta s\frac{1-Z_0^2}{Z_0-\widetilde{W}_{s_0}}+O_S(\delta^3). \tag{4.16}$$

We will now write an expression for $H_1(v_j)-H_0(v_j)$ and then use (4.13) to complete the proof. Let us first look at the term $\arg(Z_k-\widetilde{W}_{s_k})$ on the right-hand side of (4.14), and see how it changes from $k=0$ to $k=1$. For this purpose, we expand $\log(Z-\widetilde{W})$ in Taylor series up to first order in $Z-Z_0$ and up to second order in $\widetilde{W}-\widetilde{W}_{s_0}$ (since $Z_1-Z_0=O_S(\delta^2)$ while $\widetilde{W}_{s_1}-\widetilde{W}_{s_0}=O(\delta)$), as follows:

$$\begin{aligned}\arg(Z_1-\widetilde{W}_{s_1})-\arg(Z_0-\widetilde{W}_{s_0})&=\operatorname{Im}\bigl(\log(Z_1-\widetilde{W}_{s_1})-\log(Z_0-\widetilde{W}_{s_0})\bigr)\\&=\operatorname{Im}\left(\frac{Z_1-Z_0}{Z_0-\widetilde{W}_{s_0}}-\frac{\widetilde{W}_{s_1}-\widetilde{W}_{s_0}}{Z_0-\widetilde{W}_{s_0}}-\frac{(\widetilde{W}_{s_1}-\widetilde{W}_{s_0})^2}{2(Z_0-\widetilde{W}_{s_0})^2}\right)+O_S(\delta^3).\end{aligned}$$

Similar (but simpler) expansions apply to the other arguments in (4.14). We use these expansions as well as (4.14) and (4.16), to write

$$\begin{aligned}\pi(H_1(v_j)-H_0(v_j)) &= (\lambda-b)\operatorname{Im}\frac{\Delta s(1-Z_0^2)}{(Z_0-\widetilde{W}_{s_0})(Z_0-1)}\\ &\quad -2\lambda\operatorname{Im}\frac{\Delta s(1-Z_0^2)-\Delta\widetilde{W}(Z_0-\widetilde{W}_{s_0})-\frac{1}{2}(\Delta\widetilde{W})^2}{(Z_0-\widetilde{W}_{s_0})^2}\\ &\quad +(\lambda-a)\operatorname{Im}\frac{\Delta s(1-Z_0^2)}{(Z_0-\widetilde{W}_{s_0})(Z_0+1)}+O_S(\delta^3).\end{aligned}$$

With the abbreviations $\tilde{x}:=\operatorname{Re}(Z_0-\widetilde{W}_{s_0})$ and $y:=\operatorname{Im} Z_0$, the above simplifies to

$$\begin{aligned}&\frac{\pi}{\lambda}(H_1(v_j)-H_0(v_j))\\ &\qquad=\frac{2y}{\tilde{x}^2+y^2}\left(\frac{\tilde{x}}{\tilde{x}^2+y^2}(\Delta sq_1(\widetilde{W}_{s_0})-(\Delta\widetilde{W})^2)+\Delta sq_2(\widetilde{W}_{s_0})-\Delta\widetilde{W}\right)+O_S(\delta^3).\end{aligned}$$

We know from (4.13) that on $\neg\bar{\mathcal{A}}$ the conditioned expectation given $\mathcal{F}_{s_0}$ of the left-hand side is $O_{a,b}(\delta^5)$. Since $(\tilde{x}^2+y^2)/y=O_S(1)$ (by statement (4) of Lemma 4.3), we have on $\neg\bar{\mathcal{A}}$,

$$\mathbf{E}\left[\frac{\tilde{x}}{\tilde{x}^2+y^2}\big(\Delta sq_1(\widetilde{W}_{s_0})-(\Delta\widetilde{W})^2\big)+\Delta sq_2(\widetilde{W}_{s_0})-\Delta\widetilde{W}\,\middle|\,\mathcal{F}_{s_0}\right]=O_{a,b,S}(\delta^3). \tag{4.17}$$

Now, this is valid for z_j', with $j=0,1$. The choice of j only affects the left-hand side in the term $\tilde{x}/(\tilde{x}^2+y^2)$. By the choice of the points z_j' and by statement (4) of Lemma 4.3, the factor $\tilde{x}/(\tilde{x}^2+y^2)=\operatorname{Re}((Z_0-\widetilde{W}_{s_0})^{-1})$ differs between the two z_j' by an amount that is bounded away from zero by a constant depending on S. Subtracting the above relation (4.17) for z_0' from that of z_1', we therefore get (4.9) on $\neg\bar{\mathcal{A}}$. When this is used in conjunction with (4.17) again, one obtains (4.8) on $\neg\bar{\mathcal{A}}$. This concludes the proof of the proposition. □

4.4. Approximate diffusions

In this subsection we embark on the general study of random processes satisfying the conclusions of Proposition 4.2 and show that the proposition essentially characterizes the macroscopic behavior of the process. As one of the referees of this paper pointed out, one can try to do this more "traditionally" by proving tightness of the driving term and characterizing the subsequential fine mesh limit using the appropriate martingale problem. However, our approach is somewhat different (though not necessarily better).

Motivated by the proposition, we say that a continuous random $W\colon[0,S]\to[-1,1]$ is a (C,δ,η)-*approximate* (q_1,q_2)-*diffusion* if it satisfies the conclusion of the proposition; namely, for every pair of stopping times s_0 and s_1 such that almost surely $0\leqslant s_0\leqslant s_1\leqslant S$, $\Delta s:=s_1-s_0\leqslant\delta^2$ and $\sup_{s\in[s_0,s_1]}|W_s-W_{s_0}|\leqslant\delta$ we have with probability at least $1-\eta$ that (4.8) and (4.9) hold with W in place of $\widetilde{W}$.

LEMMA 4.4. *Suppose that $a,b\geqslant-\lambda$ and that Y_s satisfies* (4.7)*, where q_1 and q_2 are given by* (4.6)*. Suppose that $Y_0\in[-1,1]$ almost surely. Then there is a $C>0$ such that Y_s is a $(C,\delta,0)$-approximate (q_1,q_2)-diffusion for every $\delta\in(0,1)$ and in every time interval $[0,S]$.*

Proof. First, note that $q_1(x)=0$ and $xq_2(x)\leqslant0$ for $|x|\geqslant1$. This clearly implies that $\{Y_s\colon s\geqslant0\}\subset[-1,1]$ almost surely. Now fix some $\delta>0$ and two stopping times $s_0\leqslant s_1$ satisfying the assumptions in the definition of approximate diffusions. Let $\mathcal{F}_{s_0}$ denote the σ-field generated by $(Y_s\colon s\leqslant s_0)$. Then

$$Y_s-Y_{s_0}=\int_{s_0}^{s}q_2(Y_r)\,dr+\int_{s_0}^{s}q_1(Y_r)^{1/2}\,dB_r.$$

The second summand is a martingale, and therefore

$$\mathbf{E}[Y_{s_1}-Y_{s_0}\mid\mathcal{F}_{s_0}]=\mathbf{E}\biggl[\int_{s_0}^{s_1}q_2(Y_s)\,ds\,\biggm|\,\mathcal{F}_{s_0}\biggr].$$

Since $|Y_s-Y_{s_0}|\leqslant\delta$ for $s\in[s_0,s_1]$ and q_2 is a Lipschitz function, we conclude that

$$\mathbf{E}[Y_{s_1}-Y_{s_0}\mid\mathcal{F}_{s_0}]=\mathbf{E}[\Delta sq_2(Y_{s_0})\mid\mathcal{F}_{s_0}]+O(\delta)\mathbf{E}[\Delta s\mid\mathcal{F}_{s_0}].$$

Thus, Y_s satisfies (4.8).

We now use Itô's formula to calculate $(Y_{s_1}-Y_{s_0})^2$:

$$\begin{aligned}(Y_{s_1}-Y_{s_0})^2&=\int_{s_0}^{s_1}2(Y_s-Y_{s_0})\,dY_s+\langle Y\rangle_{s_1}-\langle Y\rangle_{s_0}\\&=\int_{s_0}^{s_1}2(Y_s-Y_{s_0})q_2(Y_s)\,ds+\int_{s_0}^{s_1}2(Y_s-Y_{s_0})q_1(Y_s)^{1/2}\,dB_s+\int_{s_0}^{s_1}q_1(Y_s)\,ds.\end{aligned}$$

The left summand is $O(\delta^3)$ and the middle summand is a martingale and therefore its expectation given $\mathcal{F}_{s_0}$ is zero. Thus

$$\mathbf{E}[(Y_{s_1}-Y_{s_0})^2\mid\mathcal{F}_{s_0}]=\mathbf{E}\biggl[\int_{s_0}^{s_1}q_1(Y_s)\,ds\biggr]+O(\delta^3)=\mathbf{E}[\Delta sq_1(Y_{s_0})]+O(\delta^3),$$

because q_1 is Lipschitz. This shows that Y_s satisfies (4.9), and completes the proof. □

PROPOSITION 4.5. *Fix $S>2$. Let $q_1, q_2\colon[-1,1]\to\mathbb{R}$ be defined as in (4.6), where we assume that $a,b>-\lambda$. Suppose that $W^1\colon[0,S]\to[-1,1]$ is a (C,δ,η)-approximate (q_1,q_2)-diffusion and that $W^2\colon[0,S]\to[-1,1]$ is a solution of (4.7) with the same q_1 and q_2 and $W^1(0)=W^2(0)$ almost surely. Also assume that $\eta<\delta^5/S^2$. Then there is a coupling of W^1 and W^2 such that $\sup_{s\in[0,S-1]}|W^1_s-W^2_s|\to0$ in probability as $\delta\to0$, while C is fixed. Namely, for every $\varepsilon>0$ there is a $\delta_0>0$, depending only on a, b, S, C and ε such that $\sup_{s\in[0,S-1]}|W^1_s-W^2_s|\leqslant\varepsilon$ with probability at least $1-\varepsilon$ if $\delta<\delta_0$.*

A useful tool in the proof of the proposition is the following lemma.

LEMMA 4.6. *Let $W\colon[0,S]\to[-1,1]$ be a (C,δ,η)-approximate (q_1,q_2)-diffusion, and let τ_0 and τ_1 be two stopping times for W satisfying $0\leqslant\tau_0\leqslant\tau_1\leqslant S$. Assume that $C\delta<\frac12$. Let $f\colon[-1,1]\to\mathbb{R}$ be a function whose second derivative is Lipschitz with Lipschitz constant 1 and which satisfies $\|f'\|_\infty,\|f''\|_\infty\leqslant1$. Set*

$$Lf(x):=\tfrac12 q_1(x)f''(x)+q_2(x)f'(x).$$

Then there is a stopping time τ_1' satisfying $\tau_0\leqslant\tau_1'\leqslant\tau_1$ almost surely and $\mathbf{P}[\tau_1'\neq\tau_1]\leqslant\eta$ such that

$$\mathbf{E}\biggl[f(W_{\tau_1'})-f(W_{\tau_0})-\int_{\tau_0}^{\tau_1'}Lf(W_s)\,ds\,\biggm|\,\mathcal{F}_{\tau_0}\biggr]=O(C+1)\delta\mathbf{E}[\delta^2+\tau_1'-\tau_0\,|\,\mathcal{F}_{\tau_0}].$$

Moreover, in the above the function f may be random, provided that it is $\mathcal{F}_{\tau_0}$-measurable.

Proof. We inductively define the stopping times s_j as follows. Set $s_0:=\tau_0$, and $s_{j+1}:=\min\{s\geqslant s_j\colon s=s_j+\delta^2$ or $|W_s-W_{s_j}|=\delta$ or $s_j=\tau_1\}$. If there is a $j\in\mathbb{N}$ such that W does not satisfy (4.8) or (4.9) for the stopping times (s_j,s_{j+1}) in place of (s_0,s_1), then let n be the minimal such j. (Note that the event that W does not satisfy (4.8) or (4.9) for (s_j,s_{j+1}) is $\mathcal{F}_{s_j}$-measurable.) Otherwise, let n be the minimal j such that $s_j=\tau_1$. Note that (s_n,s_{n+1}) is a pair of stopping times and they do not satisfy both (4.8) and (4.9) unless $s_n=\tau_1$. Consequently,

$$\mathbf{P}[s_n\neq\tau_1]\leqslant\eta. \tag{4.18}$$

Since $|W_{s_{j+1}}-W_{s_j}|\leqslant\delta$ using a Taylor series for f around W_{s_j} we have

$$f(W_{s_{j+1}})-f(W_{s_j})=f'(W_{s_j})\Delta_jW+\tfrac12f''(W_{s_j})(\Delta_jW)^2+O(\delta^3),$$

where $\Delta_jW:=W_{s_{j+1}}-W_{s_j}$. We may use (4.8) and (4.9) to estimate the conditioned expectation of Δ_jW and $(\Delta_jW)^2$ given $\mathcal{F}_{s_j}$ and get for $j<n$

$$\mathbf{E}[f(W_{s_{j+1}})-f(W_{s_j})\,|\,\mathcal{F}_{s_j}]=Lf(W_{s_j})\mathbf{E}[s_{j+1}-s_j\,|\,\mathcal{F}_{s_j}]+O(1+C)\delta^3.$$

By our assumptions about f, this may also be written as

$$\mathbf{E}\left[f(W_{s_{j+1}})-f(W_{s_j})-\int_{s_j}^{s_{j+1}} Lf(W_s)\,ds \,\middle|\, \mathcal{F}_{s_j}\right]=O(1+C)\delta^3.$$

We sum this over j from 0 to $n-1$, then take expectations conditioned on F_{τ_0}, to obtain

$$\mathbf{E}\left[f(W_{s_n})-f(W_{\tau_0})-\int_{\tau_0}^{s_n} Lf(W_s)\,ds \,\middle|\, \mathcal{F}_{\tau_0}\right]=O(1+C)\delta^3\mathbf{E}[n\,|\,\mathcal{F}_{\tau_0}]. \tag{4.19}$$

Now fix some $j\in\mathbb{N}$. On the event $j+1<n$, we have $(\Delta_j W)^2=\delta^2$ or $s_{j+1}-s_j=\delta^2$. Therefore,

$$\mathbf{E}[((\Delta_j W)^2+s_{j+1}-s_j)1_{\{j<n\}}\,|\,\mathcal{F}_{s_j}]\geqslant\delta^2\mathbf{P}[j+1<n\,|\,\mathcal{F}_{s_j}].$$

By (4.9), this gives

$$\mathbf{E}[(1+q_1(W_{s_j}))(s_{j+1}-s_j)1_{\{j<n\}}\,|\,\mathcal{F}_{s_j}]\geqslant\delta^2\mathbf{P}[j+1<n\,|\,\mathcal{F}_{s_j}]-C\delta^3 1_{\{j<n\}}.$$

We take expectation conditioned on $\mathcal{F}_{\tau_0}$ and use the fact that q_1 is bounded, to obtain

$$O(1)\mathbf{E}[(s_{j+1}-s_j)1_{\{j<n\}}\,|\,\mathcal{F}_{\tau_0}]\geqslant\delta^2\mathbf{P}[j+1<n\,|\,\mathcal{F}_{\tau_0}]-C\delta^3\mathbf{P}[j<n\,|\,\mathcal{F}_{\tau_0}].$$

We sum this over all $j\in\mathbb{N}$, to get

$$O(1)\mathbf{E}[s_n-\tau_0\,|\,\mathcal{F}_{\tau_0}]\geqslant\delta^2(1-C\delta)\mathbf{E}[n-1\,|\,\mathcal{F}_{\tau_0}]-C\delta^3.$$

By our assumption that $C\delta<\frac{1}{2}$, this implies that

$$O(1)\mathbf{E}[\delta^2+s_n-\tau_0\,|\,\mathcal{F}_{\tau_0}]\geqslant\delta^2\mathbf{E}[n\,|\,\mathcal{F}_{\tau_0}].$$

When combined with (4.19), this gives

$$\mathbf{E}\left[f(W_{s_n})-f(W_{\tau_0})-\int_{\tau_0}^{s_n} Lf(W_s)\,ds \,\middle|\, \mathcal{F}_{\tau_0}\right]=O(1+C)\delta\mathbf{E}[\delta^2+s_n-\tau_0\,|\,\mathcal{F}_{\tau_0}].$$

By (4.18), this completes the proof with $\tau_1'=s_n$. □

The next lemma bounds the expected time that W_s spends close to ±1.

LEMMA 4.7. *Let W be a (C,δ,η)-approximate (q_1,q_2)-diffusion $W:[0,S]\to[-1,1]$, where q_1 and q_2 are given by (4.6), $C\delta<\frac{1}{2}$ and $b>-\lambda$. Suppose that $S\geqslant1$. Given any $\varepsilon>0$ there is some $x_0<1$, $\delta'>0$ and $\eta'>0$ all depending only on ε, a, b and S such that if $\delta<\delta'$ and $\eta<\eta'$, then*

$$\mathbf{E}\left[\int_0^S 1_{\{W_s>x_0\}}\,ds\right]<\varepsilon.$$

A similar statement holds for the set of times such that W_s is near -1, provided that $a>-\lambda$.

Proof. Set $\mu(A):=\mathbf{E}\big[\int_0^S 1_{\{W_s\in A\}}\,ds\big]$. Note that $q_2(1)<0$. Fix some $y_0\in[0,1)$ such that $q_2(x)\leqslant\frac{1}{2}q_2(1)$ throughout $[y_0,1]$ and set $y_n:=1-(1-y_0)2^{-n}$. Let $f(x)$ be the twice continuously differentiable function that is zero on $[-1,y_0]$ and satisfies

$$f''(x)=\begin{cases}\min\{x-y_0,y_1-x\}, & \text{on } [y_0,y_1],\\ 0, & \text{on } [y_1,1].\end{cases}$$

We apply Lemma 4.6 to f with $\tau_0=0$ and $\tau_1=S$. Clearly $Lf(x)=0$ in $[-1,y_0]$. On the interval $[y_0,y_1]$, we have $f''(x)\leqslant\frac{1}{2}(y_1-y_0)$, $f'(x)\geqslant 0$ and $q_2(x)<0$. Consequently,

$$Lf(x)\leqslant\tfrac{1}{4}q_1(x)(y_1-y_0)\leqslant(1-x)(y_1-y_0)\leqslant(1-y_0)^2\quad\text{on } [y_0,y_1].$$

On the interval $[y_1,1]$, we have $f''(x)=0$, $f'(x)=\frac{1}{4}(y_1-y_0)^2$ and $q_2(x)\leqslant\frac{1}{2}q_2(1)<0$. Consequently, $Lf(x)\leqslant -c(1-y_0)^2$, where $c>0$ depends only on $q_2(1)$. Also note that

$$|f(W_S)-f(W_0)|\leqslant\sup_x f(x)-\inf_x f(x)=f(1)-0\leqslant(1-y_0)^3.$$

Therefore, Lemma 4.6 gives

$$(1-y_0)^2\mu([y_0,y_1))-c(1-y_0)^2\mu([y_1,1])\geqslant-(1-y_0)^3+O(C+1)\delta S+O(S)\eta.$$

We assume that δ' and η' are sufficiently small so that the right-hand side is larger than $-2(1-y_0)^3$. Then we get

$$c\mu([y_1,1])\leqslant 2(1-y_0)+\mu([y_0,y_1)).$$

This implies that

$$\mu([y_1,1])\leqslant 2(1-y_0)+\frac{\mu([y_0,1])}{1+c}.$$

A similar inequality applies to y_n and y_{n+1}. Induction therefore gives

$$\mu[y_n,1]\leqslant\frac{\mu([y_0,1])}{(1+c)^n}+\sum_{j=0}^{n-1}2(1-y_j)<\frac{\mu([y_0,1])}{(1+c)^n}+4(1-y_0),$$

provided that δ' and η' are smaller than some functions of n, S, C and y_0. Consequently, we first choose y_0 such that in addition to the requirements stated in the beginning of the proof, $4(1-y_0)\leqslant\frac{1}{2}\varepsilon$. Then we take n sufficiently large so that $(1-c)^{-n}S<\frac{1}{2}\varepsilon$. Then δ' and η' are determined. This proves the first claim. The second one follows by symmetry. □

The next lemma estimates the conditional expectation and conditional second moment of the time it takes W_s to move a distance of δ beyond its location at a stopping time.

LEMMA 4.8. *Let W be a (C,δ,η)-approximate (q_1,q_2)-diffusion $W\colon[0,S]\to[-1,1]$, where q_1 and q_2 are given by (4.6) and $S>1$. Let $x_0\in(0,1)$. There is a function $\delta_0>0$, depending only on x_0, C, a and b such that the following holds if $\delta<\delta_0\wedge\frac{1}{2}$ and $\eta<\delta^5/S^2$. Let τ_0 be a stopping time for W and let $\tau_1:=\inf\{s\geqslant\tau_0:s=S$ or $|W_s-W_{\tau_0}|=\delta\}$. Let $\mathcal{A}$ denote the event $\{\tau_0<S-\frac{1}{2}\}\cap\{|W_{\tau_0}|<x_0\}$. Then*

$$\mathbf{E}\left[\left|\mathbf{E}[\tau_1-\tau_0\,|\,\mathcal{F}_{\tau_0}]-\frac{\delta^2}{q_1(W_{\tau_0})}\right|1_{\mathcal{A}}\right]=O_{x_0}(C+1)\delta^3. \tag{4.20}$$

Moreover,

$$\mathbf{E}[\mathbf{E}[(\tau_1-\tau_0)^2\,|\,\mathcal{F}_{\tau_0}]1_{\mathcal{A}}]=O_{x_0}(\delta^4). \tag{4.21}$$

Proof. Let $f(x)=\frac{1}{4}(x-W_{\tau_0})^2$, and let L be as in Lemma 4.6. Then

$$Lf(x)=\tfrac{1}{4}q_1(W_{\tau_0})+O(\delta)$$

for $x\in[W_{\tau_0}-\delta,W_{\tau_0}+\delta]$. Therefore, Lemma 4.6 gives

$$\mathbf{E}\big[f(W_{\tau_1'})-\tfrac{1}{4}q_1(W_{\tau_0})(\tau_1'-\tau_0)\,\big|\,\mathcal{F}_{\tau_0}\big]=O(C+1)\delta\mathbf{E}[\delta^2+\tau_1'-\tau_0\,|\,\mathcal{F}_{\tau_0}].$$

That is,

$$\mathbf{E}[f(W_{\tau_1'})\,|\,\mathcal{F}_{\tau_0}]-\tfrac{1}{4}(q_1(W_{\tau_0})+O(C+1)\delta)\mathbf{E}[\tau_1'-\tau_0\,|\,\mathcal{F}_{\tau_0}]=O(C+1)\delta^3. \tag{4.22}$$

By choosing δ_0 sufficiently small, we make sure that $O(C+1)\delta<\frac{1}{2}q_1(W_{\tau_0})$ on $\mathcal{A}$. Since $|f(W_{\tau_1'})|\leqslant\frac{1}{4}\delta^2$, the above gives

$$\mathbf{P}[\tau_1'=S\,|\,\mathcal{F}_{\tau_0}]1_{\mathcal{A}}=\frac{O(\delta^2)}{q_1(W_{\tau_0})}.$$

Recall that $f(W_{\tau_1'})=\frac{1}{4}\delta^2$ unless $\tau_1'=S$ or $\tau_1'<\tau_1$. Therefore, on $\mathcal{A}$,

$$\mathbf{E}[f(W_{\tau_1'})\,|\,\mathcal{F}_{\tau_0}]=\tfrac{1}{4}\delta^2+O_{x_0}(\delta^4)+O(\delta^2)\mathbf{P}[\tau_1'<\tau_1\,|\,\mathcal{F}_{\tau_0}].$$

We plug this and

$$\mathbf{E}[\tau_1'-\tau_1\,|\,\mathcal{F}_{\tau_0}]=O(S)\mathbf{P}[\tau_1'\neq\tau_1\,|\,\mathcal{F}_{\tau_0}]$$

into (4.22), simplify, and get

$$\delta^2-(q_1(W_{\tau_0})+O(C+1)\delta)\mathbf{E}[\tau_1-\tau_0\,|\,\mathcal{F}_{\tau_0}]=O_{x_0}(C+1)\delta^3+O(S)\mathbf{P}[\tau_1'<\tau_1\,|\,\mathcal{F}_{\tau_0}] \tag{4.23}$$

on $\mathcal{A}$. Now (4.20) follows by dividing (4.23) by $q_1(W_{\tau_0})+O(C+1)\delta$, taking expectation and recalling that $\mathbf{P}[\tau_1'\neq\tau_1]\leqslant\eta$.

Now define $t_n := \tau_1 \wedge (\tau_0 + 16n\delta^2/q_1(W_{\tau_0}))$. If δ_0 is sufficiently small, then $Lf(x) > \frac{1}{2}Lf(W_{\tau_0}) = \frac{1}{8}q_1(W_{\tau_0})$ throughout $[W_{\tau_0}-\delta, W_{\tau_0}+\delta]$ on the event $\mathcal{A}$. Thus, we get by applying Lemma 4.6 to the stopping times t_n and t_{n+1},

$$\tfrac{1}{8}(q_1(W_{\tau_0}) - O(C+1)\delta)\mathbf{E}[t'_{n+1} - t_n \,|\, \mathcal{F}_{t_n}] \leqslant \tfrac{1}{4}\delta^2 + O(C+1)\delta^3,$$

where t'_{n+1} is the stopping time provided by the lemma. Again, on $\mathcal{A}$ we may assume that $O(C+1)\delta < \frac{1}{2}q_1(W_{\tau_0}) \wedge \frac{1}{4}$. Thus,

$$\tfrac{1}{16}q_1(W_{\tau_0})\mathbf{E}[t'_{n+1} - t_n \,|\, \mathcal{F}_{t_n}] \leqslant \tfrac{1}{2}\delta^2,$$

which implies that

$$\mathbf{P}\left[t'_{n+1} = t_n + \frac{16n\delta^2}{q_1(W_{\tau_0})} \,\middle|\, \mathcal{F}_{t_n}\right] \leqslant \frac{1}{2}.$$

But if $t'_{n+1} \neq t_n + 16n\delta^2/q_1(W_{\tau_0})$, then $t'_{n+1} = \tau_1$ or $t'_{n+1} \neq t_{n+1}$. If $\mathcal{B}_n$ denotes the event that $t'_j = t_j$ for all $j = 1, \ldots, n$, then induction gives

$$\mathbf{P}[t_n \neq \tau_1, \mathcal{B}_n \,|\, \mathcal{F}_{\tau_0}] 1_{\mathcal{A}} \leqslant 2^{-n}.$$

Lemma 4.6 gives $\mathbf{P}[t'_{n+1} \neq t_{n+1}] \leqslant \eta$ and therefore $\mathbf{P}[\neg\mathcal{B}_n] \leqslant n\eta$. (In fact, it is not hard to get the better estimate $\mathbf{P}[\neg\mathcal{B}_n] \leqslant \eta$.) Consequently, for $n \in \mathbb{N}$,

$$\mathbf{P}\left[\tau_1 - \tau_0 > \frac{16n\delta^2}{q_1(W_{\tau_0})} \,\middle|\, \mathcal{F}_{\tau_0}\right] 1_{\mathcal{A}} \leqslant 2^{-n} + \mathbf{P}[\neg\mathcal{B}_n \,|\, \mathcal{F}_{\tau_0}].$$

The above applies with $\tilde{n} := n \wedge \lceil -\log_2 \eta \rceil$ in place of n, and hence

$$\mathbf{P}\left[\tau_1 - \tau_0 > \frac{16n\delta^2}{q_1(W_{\tau_0})} \,\middle|\, \mathcal{F}_{\tau_0}\right] 1_{\mathcal{A}} \leqslant \mathbf{P}\left[\tau_1 - \tau_0 > \frac{16\tilde{n}\delta^2}{q_1(W_{\tau_0})} \,\middle|\, \mathcal{F}_{\tau_0}\right] 1_{\mathcal{A}} \leqslant 2^{-\tilde{n}} + \mathbf{P}[\neg\mathcal{B}_{\tilde{n}} \,|\, \mathcal{F}_{\tau_0}].$$

We multiply both sides by $2(n+1)(16\delta^2/q_1(W_{\tau_0}))^2$, and sum over n from $n=0$ to the least m such that $16m\delta^2/q_1(W_{\tau_0}) \geqslant S$. The result on the left-hand side bounds

$$\mathbf{E}[(\tau_1 - \tau_0)^2 \,|\, \mathcal{F}_{\tau_0}] 1_{\mathcal{A}}.$$

Consequently, the required bound (4.21) follows by taking expectations and using our assumed upper bound for η. □

The following lemma shows that when we discretize the approximate diffusion the resulting random walk has transition probabilities that can be well estimated from q_1 and q_2 away from the boundary.

LEMMA 4.9. *Fix some $x_0 \in (0,1)$. Let $W:[0,S]\to[-1,1]$ be a (C,δ,η)-approximate (q_1,q_2)-diffusion, where q_1 and q_2 are given by (4.6), $S\geqslant 1$ and $\eta\leqslant\delta^5/S$. Set*

$$Z := \{k\delta : k\in\mathbb{Z} \text{ and } |k\delta| < x_0\},$$

$s_0 := \inf\{s\geqslant 0 : W_s \in Z \text{ or } s=S\}$ and inductively

$$s_{n+1} := \inf\{s \geqslant s_n : W_{s_n} \neq W_s \in Z \text{ or } s = S\}.$$

Also set $X_n := W_{s_n}$ and $Z^0 := Z\setminus\{\min Z, \max Z\}$. Let

$$p_n^{\pm} := \mathbf{P}[X_{n+1} = X_n \pm \delta \mid \mathcal{F}_{s_n}]$$

and

$$r_n^{\pm} := \frac{1}{2} \pm \delta \frac{q_2(X_n)}{2q_1(X_n)}.$$

There is a $\delta_0>0$, depending only on C, x_0, a and b, such that if $\delta<\delta_0$, then for all $n\in\mathbb{N}$,

$$\mathbf{E}[|p_n^{\pm} - r_n^{\pm}| 1_{\{s_n < S-1/2\}} 1_{\{X_n\in Z^0\}}] \leqslant O_{x_0}(C+1)\delta^2.$$

Proof. We now use a different test function:

$$f(x) := \alpha(x-X_n)^2 + \beta(x-X_n),$$

where $\alpha := -q_2(X_n)\beta/q_1(X_n)$ and $\beta := |q_1(X_n)/6q_2(X_n)| \wedge \frac{1}{3}$ with $\beta=\frac{1}{2}$ if $q_2(X_n)=0$. The choice of α and β above is tailored to give $Lf(X_n)=0$ and $|4\alpha|+|\beta|\leqslant 1$. The latter implies that $\|f''\|_\infty, \|f'\|_\infty \leqslant 1$ when f is restricted to the interval $[-1,1]$. We now apply Lemma 4.6 again with this f and stopping times s_n and s_{n+1}. Note that $Lf=O(\delta)$ in the interval $[X_n-\delta, X_n+\delta]$. Hence $Lf(W_s)=O(\delta)$ for $s\in[s_n, s_{n+1}]$. Lemma 4.6 gives

$$\mathbf{E}[f(W_{s'_{n+1}}) - f(W_{s_n}) \mid \mathcal{F}_{s_n}] = O(C+1)\delta\mathbf{E}[\delta^2 + s'_{n+1} - s_n \mid \mathcal{F}_{s_n}],$$

on the event $X_n\in Z^0$, where s'_{n+1} is the stopping time produced by the lemma. The above may be written

$$\begin{aligned}(\beta\delta+\alpha\delta^2)p_n^+ + (-\beta\delta+\alpha\delta^2)p_n^- = &-\mathbf{E}[1_{\{W_{s'_{n+1}}\notin Z\}} f(W_{s'_{n+1}}) \mid \mathcal{F}_{s_n}] \\ &+ O(C+1)\delta\mathbf{E}[\delta^2 + s'_{n+1} - s_n \mid \mathcal{F}_{s_n}].\end{aligned}$$

Set $p_n^0 := \mathbf{P}[W_{s'_{n+1}} \notin \{X_n-\delta, X_n+\delta\} \mid \mathcal{F}_{s_n}]$. Then $1\geqslant p_n^+ + p_n^- \geqslant 1-p_n^0$. Hence, the above gives

$$(2p_n^+ - 1)\beta + \alpha\delta = O(p_n^0) + O(C+1)\delta^2 + O(C+1)\mathbf{E}[s_{n+1} - s_n \mid \mathcal{F}_{s_n}],$$

which, by the definitions of α and r_n^+ may be rewritten

$$2\beta(p_n^+ - r_n^+) = O(p_n^0) + O(C+1)\delta^2 + O(C+1)\mathbf{E}[s_{n+1} - s_n \mid \mathcal{F}_{s_n}]. \tag{4.24}$$

By (4.20) and our assumption $\eta \leqslant \delta^5/S$, we have

$$\mathbf{E}[(s_{n+1} - s_n) 1_{X_n \in Z^0} 1_{s_n < S - 1/2}] = O_{x_0}(\delta^2), \tag{4.25}$$

provided that δ_0 is sufficiently small. Since $W_{s'_{n+1}} \notin Z$ only when $s'_{n+1} \neq s_{n+1}$ or $s'_{n+1} = s_{n+1} = S$, on the event $\{X_n \in Z^0\} \cap \{s_n < S - \frac{1}{2}\}$,

$$\begin{aligned} \mathbf{E}[p_n^0 \mid \mathcal{F}_{s_n}] &\leqslant \mathbf{P}[s'_{n+1} \neq s_{n+1} \mid \mathcal{F}_{s_n}] + \mathbf{P}\big[s'_{n+1} \geqslant s_n + \tfrac{1}{2} \,\big|\, \mathcal{F}_{s_n}\big] \\ &\leqslant \mathbf{P}[s'_{n+1} \neq s_{n+1} \mid \mathcal{F}_{s_n}] + 2\mathbf{E}[s_{n+1} - s_n \mid \mathcal{F}_{s_n}]. \end{aligned} \tag{4.26}$$

Note that $\beta^{-1} = O_{x_0}(1)$ and $\mathbf{P}[s'_{n+1} \neq s_{n+1}] \leqslant \eta$. Hence, we now obtain the result for p_n^+ by taking expectation on the event $\{X_n \in Z^0\} \cap \{s_n < S - \frac{1}{2}\}$ in (4.24) and using (4.25) and (4.26). A symmetric argument applies to p_n^- and r_n^-, and the proof is complete. □

Next, we show that the time parameterization of W can be well approximated by a function of the discretized walk trajectory.

LEMMA 4.10. *Assume the setting and notation of Lemma* 4.9 *in addition to* $\delta < \delta_0$. *For* $n \in \mathbb{N}$ *let* t_n *denote the time spent up to time* s_n *in segments* $[s_j, s_{j+1}]$ *such that* $X_j = W_{s_j} \in Z^0$; *that is*,

$$t_n := \sum_{j=0}^{n-1} 1_{\{X_j \in Z^0\}} (s_{j+1} - s_j).$$

Also let

$$\sigma_n := \sum_{j=0}^{n-1} 1_{\{X_j \in Z^0\}} \frac{\delta^2}{q_1(X_j)}.$$

Let $N_0 := \min\{n \in \mathbb{N} : s_n \geqslant S - \frac{1}{2}\}$. *Then for all* $n \in \mathbb{N}$,

$$\mathbf{E}[\max\{|\sigma_j - t_j| : j = 1, \ldots, n \wedge N_0\}] \leqslant O_{x_0}(C+1)(\delta^2 n^{1/2} + \delta^3 n). \tag{4.27}$$

Proof. Let $v_j := (s_{j+1} - s_j) 1_{\{X_j \in Z^0\}} 1_{\{j < N_0\}}$, $u_j := \mathbf{E}[v_j \mid \mathcal{F}_{s_j}]$ and

$$w_j := \frac{\delta^2}{q_1(X_j)} 1_{\{X_j \in Z^0\}} 1_{\{j < N_0\}}.$$

Now, $M_n := \sum_{j=0}^{n-1} (v_j - u_j)$ is clearly a martingale. Consequently, Doob's maximal inequality for L^2 martingales [RY, II.1.6] gives

$$\mathbf{E}[\max\{|M_j| : j = 1, \ldots, n\}]^2 \leqslant O(1)\mathbf{E}[M_n^2] = O(1) \sum_{j=0}^{n-1} \mathbf{E}[(v_j - u_j)^2].$$

Since $u_j=\mathbf{E}[v_j|\mathcal{F}_{s_j}]$, we have $\mathbf{E}[(v_j-u_j)^2]\leqslant\mathbf{E}[v_j^2]$. By Lemma 4.8, $\mathbf{E}[v_j^2]$ is bounded by the right-hand side of (4.21). Now, the right-hand side of (4.20) bounds $\mathbf{E}[|u_j-w_j|]$. The result follows by our assumption $\eta\leqslant\delta^5/S^2$, since for every $m\leqslant n\wedge N_0$,

$$|\sigma_m-t_m|=\left|\sum_{j=0}^{m-1}(v_j-w_j)\right|\leqslant|M_m|+\sum_{j=0}^{m-1}|u_j-w_j|$$
$$\leqslant\max\bigl\{|M_j|:j=1,\ldots,n\bigr\}+\sum_{j=0}^{n-1}|u_j-w_j|. \qquad \square$$

Proof of Proposition 4.5. By Lemma 4.4, W^2 is a $(C',\delta,0)$-approximate (q_1,q_2)-diffusion for some fixed constant $C'>0$ and every $\delta>0$. We may assume, with no loss of generality, that $C\geqslant C'$. Let $\varepsilon>0$. Let $x_0'\in(1-\varepsilon,1)$ satisfy Lemma 4.7 with this given ε, and assume that δ is sufficiently small so that that lemma is valid. Take $x_0=\frac{1}{2}(1+x_0')$. Let Z and Z^0 be as in Lemma 4.9, let s_j^k be the corresponding stopping times introduced there for W^k and let $p_{k,j}^{\pm}$ denote the random transition probabilities for W^k defined there. Also abbreviate $X_j^k:=W_{s_j}^k$. Let $\mathcal{F}_s^k$ denote the filtration of W^k, $k=1,2$. Let $Y_j^k=1$ if $X_{j+1}^k-X_j^k=\delta$, $Y_j^k=-1$ if $X_{j+1}^k-X_j^k=-\delta$ and $Y_j^k=0$ if $|X_{j+1}^k-X_j^k|\neq\delta$. Then $\mathbf{P}[Y_j^k=\pm1|\mathcal{F}_{s_j^k}^k]=p_{k,j}^{\pm}$ and $\mathbf{P}[Y_j^k=0|\mathcal{F}_{s_j^k}^k]=1-p_{k,j}^+-p_{k,j}^-$ if $X_j^k\in Z^0$.

For the coupling of W^1 and W^2 we use an independent identically distributed sequence U_j of uniform random variables in $[0,1]$. The coupling proceeds as follows. Up to their corresponding stopping times s_0^k, $k=1,2$, let them run independently. Inductively, we suppose that the coupling has been constructed up to their corresponding stopping times s_j^k, $k=1,2$. For each $k=1,2$, we take

$$Y_j^k=\begin{cases}1, & \text{if } U_j\leqslant p_{k,j}^+,\\ -1, & \text{if } U_j\geqslant 1-p_{k,j}^-,\\ 0, & \text{if } U_j\in(p_{k,j}^+,1-p_{k,j}^-).\end{cases}$$

(In other words, we try to match up $X_{j+1}^1-X_j^1$ with $X_{j+1}^2-X_j^2$ as much as possible.) These choices respect the correct conditional distributions for these variables. Now we sample the restriction of W^1 to $[s_j^1,s_{j+1}^1]$ and the restriction of W^2 to $[s_j^2,s_{j+1}^2]$ independently of their corresponding conditional distribution given $(\mathcal{F}_{s_j}^1,Y_j^1)$ and $(\mathcal{F}_{s_j}^2,Y_j^2)$, respectively. This completes the description of the coupling.

Let $N:=\min\bigl\{n:s_n^1\vee s_n^2\geqslant S-\frac{1}{2}\bigr\}$. Let $\mathcal{A}_j$ be the event $\{X_j^1,X_j^2\in Z^0\}$ and set

$$Q_j:=1_{\mathcal{A}_j}|Y_j^1-Y_j^2| \quad\text{and}\quad \widehat{Q}_n:=\sum_{j=0}^{n}Q_j1_{\{j<N\}}.$$

Note that on $\neg\mathcal{A}_j$ we have $|X_{j+1}^1-X_{j+1}^2|\leqslant|X_j^1-X_j^2|$ unless $s_{j+1}^1\vee s_{j+1}^2=S$. Moreover, $|X_0^1-X_0^2|\leqslant\delta$ and when $Y_j^k=0$ we have $s_{j+1}^k=S$. Consequently,

$$|X_n^1-X_n^2|1_{\{n<N\}}\leqslant\delta+\delta\widehat{Q}_{n-1}. \tag{4.28}$$

We now proceed to estimate $\theta_j := \mathbf{E}[Q_j 1_{\{j<N\}}]$. Clearly,

$$\mathbf{P}[Q_j \neq 0 \,|\, \mathcal{F}^1_{s^1_j} \vee \mathcal{F}^2_{s^2_j}] \leqslant (|p^+_{1,j} - p^+_{2,j}| + |p^-_{1,j} - p^-_{2,j}|) 1_{\mathcal{A}_j}.$$

Let $r^\pm_{k,j}$ be the $r^\pm_j$ in Lemma 4.9 corresponding to the process W^k. Then

$$|p^\pm_{1,j} - p^\pm_{2,j}| \leqslant |p^\pm_{1,j} - r^\pm_{1,j}| + |r^\pm_{1,j} - r^\pm_{2,j}| + |r^\pm_{2,j} - p^\pm_{2,j}|.$$

By that lemma,

$$\mathbf{E}[|p^\pm_{k,j} - r^\pm_{k,j}| 1_{\mathcal{A}_j} 1_{\{j<N\}}] \leqslant O_{x_0}(C+1)\delta^2.$$

Using the expression given for $r^\pm_{k,j}$, we deduce that

$$|r^\pm_{1,j} - r^\pm_{2,j}| \leqslant O_{x_0}(\delta)|X^1_j - X^2_j|.$$

Thus, we get

$$\theta_n = \mathbf{E}[Q_n 1_{\{n<N\}}] \leqslant O_{x_0}(C+1)\delta^2 + O_{x_0}(\delta)\mathbf{E}[|X^1_n - X^2_n| 1_{\{n<N\}}].$$

In conjunction with (4.28), this gives

$$\theta_n \leqslant O_{x_0}(C+1)\delta^2 \Bigg(1 + \sum_{j=0}^{n-1} \theta_j\Bigg).$$

Induction therefore implies

$$\theta_n \leqslant O_{x_0}(C+1)\delta^2 (1 + O_{x_0}(C+1)\delta^2)^n.$$

Taking note of (4.28), we infer that

$$\mathbf{E}[\max_{j<n\wedge N} |X^1_j - X^2_j|] \leqslant \delta + O_{x_0}(C+1)\delta^3 n (1 + O_{x_0}(C+1)\delta^2)^n. \tag{4.29}$$

Now let $m_0 := \lceil 4S\delta^{-2} \max\{q_1(x) : |x| \leqslant x_0\} \rceil$. Observe that at least one of every two consecutive $j \in \mathbb{N}$ satisfies $X^k_j \in Z^0$ or $s^k_j = S$. Consequently, $m_0 < N$ implies that $\sigma^k_{m_0} \geqslant 2S$ for $k = 1, 2$, where σ^k_j denotes the σ_j from Lemma 4.10 corresponding to W^k. Note that in that lemma $t_j \leqslant s_j \leqslant S$. Therefore, taking $n = m_0$ in (4.27) implies that

$$\mathbf{P}[N > m_0] \leqslant O_{x_0}(C+1)S^{-1}(\delta^2 m_0^{1/2} + \delta^3 m_0) = O_{x_0}(C+1)\delta.$$

Set $X^* := \max_{j<N} |X^1_j - X^2_j|$. The above and (4.29) with $n = m_0$ imply that

$$\mathbf{P}[X^* > \delta^{1/2} \text{ or } N > m_0] \leqslant O_{x_0,C,S}(\delta^{1/2}). \tag{4.30}$$

Now for each $s\leqslant S$ let $J(s):=\min\{j\in\mathbb{N}:s_j^1\geqslant s\}$. Then

$$\begin{aligned}\sup_{s\in[0,S-1]}|W_s^1-W_s^2|\leqslant&\sup_{s\in[0,S-1]}|W_s^1-X_{J(s)}^1|\\&+\sup_{s\in[0,S-1]}|X_{J(s)}^1-X_{J(s)}^2|+\sup_{s\in[0,S-1]}|W_s^2-X_{J(s)}^2|.\end{aligned}\tag{4.31}$$

First, it is clear that

$$\sup_{s\in[0,S-1]}|W_s^1-X_{J(s)}^1|\leqslant\varepsilon+\delta.$$

(The left-hand side is usually at most δ but can be as large as $1-x_0+\delta$ if, for example, $X_{J(s)-1}=\max Z=x_0$.) We leave aside, for now, the estimation of the second summand in (4.31) and consider the last. Set

$$t^*:=\sup_{s\in[0,S-1]}|s-s_{J(s)}^2|.$$

Since $X_{J(s)}^2=W_{s_{J(s)}^2}^2$, we have

$$\sup_{s\in[0,S-1]}|W_s^2-X_{J(s)}^2|\leqslant\sup_{s\in[0,S-1]}\sup_{t\in[0,t^*]}|W_s^2-W_{s+t}^2|.\tag{4.32}$$

Observe from (4.27) that for $k=1,2$,

$$\max_{j\leqslant N\wedge m_0}\{|\sigma_j^k-t_j^k|\}\to0$$

in probability as $\delta\to0$, where t_j^k is the t_j of Lemma 4.10 corresponding to W^k. Now the choice of x_0 (via Lemma 4.7) implies that

$$\mathbf{P}[\max_{j\in\mathbb{N}}|s_0^k+t_j^k-s_j^k|>\sqrt{\varepsilon}\,]<\sqrt{\varepsilon}.$$

Lemmas 4.7 and 4.8 imply that $\mathbf{E}[s_0^k]\leqslant\varepsilon\vee O_{x_0,C}(\delta^2)$. Consequently, we have

$$\max_{j\leqslant N\wedge m_0}\{|\sigma_j^k-s_j^k|\}\to0\tag{4.33}$$

in probability as $\varepsilon,\delta\to0$. Now,

$$\begin{aligned}|\sigma_n^1-\sigma_n^2|&\leqslant\sum_{j=0}^{n-1}\left|1_{\{X_j^1\in Z^0\}}\frac{\delta^2}{q_1(X_j^1)}-1_{\{X_j^2\in Z^0\}}\frac{\delta^2}{q_1(X_j^2)}\right|\\&\leqslant\sum_{j=0}^{n-1}\left|\frac{\delta^2}{q_1(X_j^1)}-\frac{\delta^2}{q_1(X_j^2)}\right|+\sum_{k=1}^{2}\sum_{j=0}^{n-1}1_{\{X_j^k\notin Z^0\}}\frac{\delta^2}{q_1(X_j^k)}.\end{aligned}$$

The right-hand side is monotone non-decreasing in n. When $n \leqslant N \wedge m_0$ the first sum is at most $O_{x_0}(\delta^2) m_0 X^*$. It is easy to see that for $n = N \wedge m_0$ the iterated sum on the right tends to 0 in probability: this follows from the proof of (4.33), because if we replace x_0 by $x_0 + \delta$, the terms appearing in this iterated sum are included in σ_n. We now get, from (4.30),

$$\max_{j \leqslant N} |\sigma_j^1 - \sigma_j^2| \to 0 \quad \text{in probability as } \varepsilon, \delta \to 0.$$

By (4.33), this also gives

$$\max_{j \leqslant N} |s_j^1 - s_j^2| \to 0 \quad \text{in probability as } \varepsilon, \delta \to 0.$$

In particular, (4.30) and (4.33) imply $\sup_{j \leqslant N-1} (s_{j+1}^k - s_j^k) \to 0$ in probability, because $\sigma_{j+1}^k - \sigma_j^k \leqslant O_{x_0}(\delta^2)$. One consequence is that $\mathbf{P}[J(S-1) \geqslant N] \to 0$. Furthermore, we now have

$$t^* \leqslant \sup_{s \in [0, S-1]} |s - s^1_{J(s)}| + \sup_{s \in [0, S-1]} |s^2_{J(s)} - s^1_{J(s)}| \to 0$$

in probability. Now (4.32) implies that $\sup_{s \in [0, S-1]} |W_s^2 - X^2_{J(s)}| \to 0$ in probability, because the right-hand side in (4.32) is smaller than $(t^*)^{1/3}$ with probability going to 1 as $t^* \to 0$, since W^2 is a solution of (4.7). This takes care of the last summand on the right-hand side of (4.31).

The middle summand on the right-hand side of (4.31) also tends to 0 in probability because, as we have seen, $\mathbf{P}[J(S-1) < N] \to 1$ and $X^* \to 0$ in probability. This completes the proof. □

Proof of Theorem 4.1. Let $x_1, x_2 \in [-1, 1]$ be two arbitrary points, and let

$$Y_s^1, Y_s^2 \colon [0, \infty) \to [-1, 1]$$

be two independent solutions of (4.7) (with respect to two independent Brownian motions) which start at x_1 and x_2, respectively. We claim that $s_= := \min\{s : Y_s^1 = Y_s^2\} < \infty$ almost surely. The argument is quite standard. Suppose without loss of generality that $x_2 > x_1$. By Lemma 4.7, it is unlikely that Y_s^2 stays very close to 1 for a long time and unlikely that Y_s^1 stays very close to -1 for a long time. It is therefore easy to conclude from Lemmas 4.8 and 4.9 that there are constants $s_0, c_0 > 0$ (which do not depend on x_1 or x_2) such that $\mathbf{P}[Y^2_{s_0} < 0] > c_0$ and $\mathbf{P}[Y^1_{s_0} > 0] > c_0$. This implies that $\mathbf{P}[s_= < s_0] > c_0^2$. By strong uniqueness of solutions of (4.7), it follows that the solutions are Markov and have stationary transition probabilities. Consequently, we get by induction $\mathbf{P}[s_= > n s_0] < (1 - c_0^2)^n$ for all $n \in \mathbb{N}$, which proves that $s_= < \infty$ almost surely.

We now argue that $\mathbf{P}[s_= < \infty] = 1$ and the uniqueness in law of solutions of (4.7) implies that for every Borel subset $A \subset [-1,1]$ the limit

$$\mu(A) := \lim_{r \to \infty} \mathbf{P}[Y_r^1 \in A]$$

exists. We may couple a solution started at some time $s_0 < r$ such that $r - s_0$ is a large constant to be independent of Y_s^1 until the first time in $[s_0, \infty)$ in which they meet and to agree with Y_s^1 afterwards. Because these solutions are likely to meet prior to time r, it follows that $\mathbf{P}[Y_r^1 \in A]$ is close to the probability that the solution started at time s_0 is in A at time r, proving the existence of μ. (In our setting, μ may be explicitly described. Its density with respect to the Lebesgue measure is proportional to $(1+x)^{(b-\lambda)/2\lambda}(1-x)^{(a-\lambda)/2\lambda}$.) Since solutions of (4.7) are Markov, a solution

$$\widetilde{Y} : [0, \infty) \to [-1, 1]$$

of (4.7) such that the distribution of $\widetilde{Y}_0$ is given by μ is time-stationary. To get a time-stationary solution $Y : (-\infty, \infty) \to [-1, 1]$, we may take the weak limit of time-translations of $\widetilde{Y}$.

Now let S' be much larger than S. By Proposition 4.2 with S' instead of S and Proposition 4.5 translated to start at time $-S'$ and an appropriate choice of the S appearing there, we may couple $\widetilde{W}_s$ so that with probability close to 1 it stays close to a solution W_s^2 of (4.7) starting at $\widetilde{W}_{-S'}$ throughout $[-S', S']$. We may at the same time couple W_s^2 so that with high probability it agrees with Y_s inside the interval $[-S, S]$. Then with high probability $\widetilde{W}_s$ stays close to Y_s in $[-S, S]$, which concludes the proof of the theorem. □

Remark 4.11. At this point, it may be worthwhile to point out which properties of the functions q_1 and q_2 played a part in the proof. The only properties that are essential for the above proof are that q_1 and q_2 are both Lipschitz continuous in $[-1, 1]$, that $q_1 > 0$ in $(-1, 1)$ and $q_1 = 0$ on $\{-1, 1\}$, and that $q_2(1) < 0 < q_2(-1)$.

4.5. Back to chordal

In §4.2, we described the transition from the chordal Loewner system to the setup with the points ± 1 fixed. We now describe the reverse transformation. We start with some continuous $Y : (-\infty, \infty) \to [-1, 1]$. Set

$$w_s := \frac{e^s}{2} Y_s + \frac{1}{2} \int_{-\infty}^{s} e^u Y_u \, du. \tag{4.34}$$

Also define

$$t^*(s) := \frac{1}{8}\int_{-\infty}^{s} e^{2u}(1-Y_u^2)\,du, \quad s^*(t) := \sup\{s\in(-\infty,\infty) : t^*(s)\leqslant t\}.$$

Now set $\widehat{Y}_t := w_{s^*(t)}$ for $t>0$, $\widehat{Y}_0 := 0$, and observe that $\widehat{Y}$ is continuous provided that there is no non-trivial time interval in which $Y\in\{\pm 1\}$.

LEMMA 4.12. *If $Y_s = \widetilde{W}_s$ is defined from W_t as in §4.2, then $\widehat{Y}_t = W_t$.*

Proof. By the definition of $s(t)$ in §4.2, we have $e^s = y_t - x_t$. Consequently, (4.4) implies that $\partial_t(t^*(s(t))) = 1$. Since $s(0) = -\infty$ and $t^*(-\infty) = 0$, it follows that $t^*(s(t)) = t$ for all $t\geqslant 0$ and $s^*(t^*(s)) = s$ for all $s\in(-\infty,\infty)$. Next, (1.4) and the definition of W^* give

$$\partial_t(y_t + x_t) = \frac{8W_t^*}{(y_t - x_t)(1-(W_t^*)^2)}.$$

Since $y_0 = 0 = x_0$ and $\widetilde{W}_{s(t)} = W_t^*$, this implies that

$$y_t + x_t = \int_0^t \frac{8\widetilde{W}_{s(r)}}{(y_r - x_r)(1-\widetilde{W}_{s(r)}^2)}\,dr = \int_{-\infty}^{s(t)} e^u \widetilde{W}_u\,du,$$

where the second equality follows by a change of variable. Now $\widehat{Y}_t = w_{s(t)} = W_t$ follows from the definition of W_t^*. The proof of the lemma is therefore complete. □

We now discuss the behavior of solutions of (4.7) in the chordal coordinate system, but generalize to the case $\varkappa\neq 4$.

LEMMA 4.13. *Let $\tilde{a},\tilde{b}>0$, let $Y:(-\infty,\infty)\to[-1,1]$ be a solution of*

$$dY_s = (-\tilde{a}(Y_s-1) - \tilde{b}(Y_s+1))\,ds + \sqrt{\tfrac{1}{2}\varkappa(1-Y_s^2)}\,dB_s,$$

and let $\widehat{Y}_t$ be the corresponding process, as described following (4.34). Then on any time interval which avoids $\{t : Y_{s^*(t)}\notin\{\pm 1\}\}$, the process $\widehat{Y}_t$ satisfies the SDE of the driving term for SLE$(\varkappa; 2(\tilde{a}-1), 2(\tilde{b}-1))$:

$$d\widehat{Y}_t = \frac{2(\tilde{a}-1)}{\widehat{Y}_t - x_t} + \frac{2(\tilde{b}-1)}{\widehat{Y}_t - y_t} + \sqrt{\varkappa}\,d\widehat{B}_t$$

for some Brownian motion $\widehat{B}_t$, where x_t and y_t satisfy (1.4) with $\widehat{Y}_t$ in place of W_t.

Proof. Let $\beta_s = \int_{-\infty}^{s} e^u Y_u \, du$, $y_s^* := \frac{1}{2}(\beta_s + e^s)$ and $x_s^* := \frac{1}{2}(\beta_s - e^s)$. Itô's formula and the definition of t^* give

$$\begin{aligned} dw_s &= e^s Y_s \, ds + \tfrac{1}{2} e^s \sqrt{\tfrac{1}{2}\varkappa(1-Y_s^2)}\, dB_s + \tfrac{1}{2} e^s (-\tilde{a}(Y_s-1) - \tilde{b}(Y_s+1))\, ds \\ &= \sqrt{\varkappa (t^*)'(s)}\, dB_s + 4e^{-s}\left(\frac{\tilde{a}-1}{Y_s+1} + \frac{\tilde{b}-1}{Y_s-1}\right)(t^*)'(s)\, ds \\ &= \sqrt{\varkappa (t^*)'(s)}\, dB_s + 2\left(\frac{\tilde{a}-1}{w_s - x_s^*} + \frac{\tilde{b}-1}{w_s - y_s^*}\right)(t^*)'(s)\, ds. \end{aligned}$$

Now set

$$\widehat{B}_t = \int_{-\infty}^{s^*(t)} \sqrt{(t^*)'(u)}\, dB_u.$$

Then $\widehat{B}_t$ is clearly a continuous martingale. Since also $\langle \widehat{B} \rangle_t = \int_{-\infty}^{s^*(t)} (t^*)'(u)\, du = t$, we find that $\widehat{B}$ is a Brownian motion with respect to t. The above formula for dw gives

$$d\widehat{Y}_t = \sqrt{\varkappa}\, d\widehat{B}_t + 2\left(\frac{\tilde{a}-1}{\widehat{Y}_t - x_t} + \frac{\tilde{b}-1}{\widehat{Y}_t - y_t}\right) dt,$$

where $x_t := x^*_{s^*(t)}$ and $y_t := y^*_{s^*(t)}$. Now

$$\partial_t x_t = \frac{\partial_s x_s^*}{(t^*)'(s)} = \frac{2}{x_t - \widehat{Y}_t},$$

and similarly for y_t. This concludes the proof. □

As mentioned at the beginning of this section (§4), existence and uniqueness of solutions to the usual SDE defining SLE$(\varkappa; \varrho_1, \varrho_2)$ have not been proved beyond times when the driving term W_t meets the force points. We now offer the following.

Definition 4.14. If $\tilde{a}, \tilde{b} > 0$, then the Loewner equation driven by the process $\widehat{Y}$ of Lemma 4.13 is called SLE$(\varkappa; 2(\tilde{a}-1), 2(\tilde{b}-1))$ (starting from $(0, 0_-, 0_+)$).

4.6. Loewner driving term convergence

In this section, we complete the proof of Theorem 1.3. The theorem will follow quite easily from Theorem 4.1.

Proof of Theorem 1.3. Fix $T, \varepsilon, \varepsilon_0, \varepsilon' > 0$. Let Y_s and $\widetilde{W}_s$ be coupled as in Theorem 4.1, but with ε' in place of ε. Let $\widehat{Y}_t$, $t^*(s)$ and $s^*(t)$ be defined from Y_s as in the beginning of §4.5. Since the interior of the set of times for which $Y_s \in \{-1, 1\}$ is empty,

it follows that there is some positive $t_0>0$ such that with probability at least $1-\varepsilon_0$ we have $t^*(1)-t^*(0)>t_0$. Because Y_s is stationary, it follows that

$$\mathbf{P}[t^*(s+1)-t^*(s)>e^{2s}t_0]\geqslant 1-\varepsilon_0.$$

In particular, there is some $S_0>0$ such that $\mathbf{P}[t^*(S_0)>T+1]\geqslant 1-\varepsilon_0$. Fix S_0 satisfying this and additionally $e^{-S_0}<\frac{1}{2}\varepsilon_0$.

Let

$$\tilde{t}(s):=\frac{1}{8}\int_{-\infty}^{s} e^{2u}(1-\widetilde{W}_u^2)\,du,$$

which is the equivalent of $t^*(s)$ with $\widetilde{W}$ replacing Y. It is clear that if $\varepsilon'=\varepsilon'(S_0,\varepsilon_0)$ is sufficiently small and

$$\sup\bigl\{|\widetilde{W}_s-Y_s|:s\in[-S_0,S_0]\bigr\}<\varepsilon', \tag{4.35}$$

then for every $s\in[-\infty,S_0]$ the right-hand side in (4.34) differs from the corresponding quantity where $\widetilde{W}$ replaces Y by at most ε_0. Lemma 4.12 then gives

$$\sup\{|\widehat{W}_{\tilde{t}(s)}-W_{t^*(s)}|:s\leqslant[0,S_0]\}<\varepsilon_0,$$

where W_t is the chordal driving term for $\mathrm{SLE}(4;a/\lambda-1,b/\lambda-1)$ and $\widehat{W}$ is the chordal driving term for $\phi_D\circ\gamma$. (Here, we also use the fact that $s^*(t^*(s))=s$.) If we assume (4.35) with ε' sufficiently small, we also get $\sup\{|t^*(s)-\tilde{t}(s)|:s\leqslant S_0\}<\varepsilon_0$. Let T_0 be the obvious upper bound for $\tilde{t}$ and t^* in $(-\infty,S_0]$; that is, $T_0:=\frac{1}{16}e^{2S_0}$. Also set

$$M(\varepsilon_0):=\sup\{|W_{t_0}-W_{t_1}|:0\leqslant t_0\leqslant t_1\leqslant t_0+\varepsilon_0\leqslant T_0+\varepsilon_0\}.$$

Since W_t is almost surely continuous, $M(\varepsilon_0)\to 0$ in probability as $\varepsilon_0\to 0$. Now the triangle inequality $|\widehat{W}_{\tilde{t}(s)}-W_{\tilde{t}(s)}|\leqslant|\widehat{W}_{\tilde{t}(s)}-W_{t^*(s)}|+|W_{t^*(s)}-W_{\tilde{t}(s)}|$ shows that when (4.35) holds we have $\sup\{|\widehat{W}_{\tilde{t}(s)}-W_{\tilde{t}(s)}|:s\leqslant S_0\}<\varepsilon_0+M(\varepsilon_0)$. Hence,

$$\sup\{|\widehat{W}_t-W_t|:t\in[0,\tilde{t}(S_0)]\}<\varepsilon_0+M(\varepsilon_0).$$

But we have seen that $t^*(S_0)$ is very likely to be larger than $T+1$ and that

$$|t^*(S_0)-\tilde{t}(S_0)|<\varepsilon_0$$

when (4.35) holds. By Theorem 4.1, when r_D is large (4.35) holds with high probability. This concludes the proof. □

4.7. Carathéodory convergence

For $K \subset \overline{\mathbb{H}}$ let $N_\varepsilon(K)$ denote the ε-neighborhood of K in $\overline{\mathbb{H}}$, and let $\overline{\mathbb{H}}_\varepsilon(K)$ denote the unbounded connected component of $\overline{\mathbb{H}} \setminus N_\varepsilon(K)$. Set

$$d_{\mathrm{CKC}}(K, K') := \inf\{\varepsilon > 0 : K \cap \overline{\mathbb{H}}_\varepsilon(K') = \varnothing = K' \cap \overline{\mathbb{H}}_\varepsilon(K)\}.$$

It is easy to see that d_{CKC} is a metric on the collection of compact connected $K \subset \overline{\mathbb{H}}$ such that $\mathbb{H} \setminus K$ is connected. This metric is related to the Carathéodory kernel convergence topology, which is of central importance in the theory of conformal mappings.

Let K_t and K'_t denote the evolving hulls corresponding to two Loewner evolutions generated by continuous driving terms W_t and W'_t, respectively (as defined in §1.4). Such evolving hulls are also sometimes called *Loewner chains*. We set, for $T \geqslant 0$,

$$d^T_{\mathrm{CKC}}(K, K') := \sup_{t \in [0,T]} d_{\mathrm{CKC}}(K_t, K'_t).$$

The following is a simple lemma relating uniform convergence of driving terms to d_{CKC}-convergence of the corresponding Loewner chains.

LEMMA 4.15. *The Loewner transform $W \mapsto K$ is a continuous map from the space of continuous paths W with the topology of uniform convergence to the space of Loewner chains with d^T_{CKC}-convergence.*

In other words, for every $\varepsilon > 0$, every $T > 0$ and every $W : [0,T] \to \mathbb{R}$ continuous, there is some $\delta = \delta(\varepsilon, T, W) > 0$ such that if $\widetilde{W} : [0,T] \to \mathbb{R}$ is continuous and satisfies $\sup_{t \in [0,T]} |W_t - \widetilde{W}_t| < \delta$, then the corresponding Loewner chains satisfy $d^T_{\mathrm{CKC}}(K, \widetilde{K}) < \varepsilon$.

This lemma is similar in spirit to [L4, Proposition 4.47]. As is well known, $\widetilde{K}_T \to K_T$ in the Hausdorff metric does not follow from $\widetilde{W} \to W$ uniformly in $[0,T]$.

Proof. Fix $T > 0$. Suppose that $W^n \to W$ uniformly in $[0,T]$. Let K^n denote the Loewner chain corresponding to W^n and let $g^{(n)}_t : \mathbb{H} \setminus K^n_t \to \mathbb{H}$ denote the corresponding Loewner evolution. Fix some $t_0 \in [0,T]$. Since $\operatorname{diam} K^n_{t_0}$ is clearly bounded by a function of T and $\|W^n\|_\infty$ [L4, Lemma 4.13], the closure of $\{K^n_{t_0} : n \in \mathbb{N}_+\}$ is compact with respect to the Hausdorff metric on non-empty compact subsets of $\overline{\mathbb{H}}$. Consider some integer sequence $n_j \to \infty$ for which the Hausdorff limit $K' := \lim_{j \to \infty} K^{n_j}_{t_0}$ exists. If $z \in \overline{\mathbb{H}} \setminus K_{t_0}$, then there is a neighborhood U of z in $\overline{\mathbb{H}}$ such that $U \cap K^n_{t_0} = \varnothing$ for all sufficiently large n, by the continuity of solutions of ODE's in the vector field specifying the ODE. It follows that $z \notin K'$, and hence $K' \subset K_{t_0}$.

With the intention of reaching a contradiction, suppose that there is some point $z \in \partial K_{t_0} \setminus (K' \cup \mathbb{R})$. Let z' be a point in $\mathbb{H} \setminus K_{t_0}$ satisfying $|z - z'| < \frac{1}{3} \operatorname{dist}(z', K' \cup \mathbb{R})$. The

above argument shows that $\lim_{j\to\infty} g_{t_0}^{(n_j)}(z')=g_{t_0}(z')\in\mathbb{H}$. On the other hand, for all sufficiently large j we have $\operatorname{dist}(z', K_{t_0}^{n_j})>2|z-z'|$. Now the Koebe distortion theorem (e.g., [P, Corollary 1.4]) applied to the restriction of $g_{t_0}^{(n_j)}$ to the disk of radius $2|z-z'|$ about z' (once with z' and again with z) shows that $\operatorname{Im} g_{t_0}^{(n_j)}(z)\geqslant\frac{3}{16}\operatorname{Im} g_{t_0}^{(n_j)}(z')$. However, since $z\in K_{t_0}$, for every $\varepsilon>0$ there is some first $t_1\in[0,t_0)$ such that $\operatorname{Im} g_{t_1}(z)\leqslant\varepsilon$. The convergence argument above shows that for all arbitrarily large n, $\operatorname{Im} g_{t_1}^{(n)}(z)\leqslant 2\varepsilon$. Since $\operatorname{Im} g_t^{(n)}(z)$ decreases monotonically in t, it follows that $\operatorname{Im} g_{t_0}^{(n)}(z)\leqslant 2\varepsilon$ for all sufficiently large n. This contradicts our previous conclusion

$$\operatorname{Im} g_{t_0}^{(n_j)}(z)\geqslant\tfrac{3}{16}\operatorname{Im} g_{t_0}^{(n_j)}(z')\to\tfrac{3}{16}\operatorname{Im} g_{t_0}(z')>0,$$

and proves that $K'\supset\partial K_{t_0}\backslash\mathbb{R}$.

Let $\widetilde{K}$ be the union of K' and the bounded connected components of $\overline{\mathbb{H}}\backslash K'$. The above implies that $\widetilde{K}\supset K_{t_0}\backslash\mathbb{R}$. Now note that $K_{t_0}\backslash\mathbb{R}$ is dense in K_{t_0}. (This follows from the easy direction (2) $\Rightarrow$ (1) in [LSW1, Theorem 2.6] and from the fact that $(\mathbb{H}\cap K_t)\backslash K_{t'}$ is non-empty when $t>t'$.) Consequently, $\widetilde{K}=K_{t_0}$, which implies that $d_{\mathrm{CKC}}(K_t^n, K_t)\to 0$ for every fixed $t\in[0,T]$.

Note that the above proof also gives $\lim_{s\to t} d_{\mathrm{CKC}}(K_s, K_t)=0$ for $t\in[0,T]$ and s tending to t in $[0,T]$. Thus, K_t is continuous in t with respect to d_{CKC}. Since $d_{\mathrm{CKC}}(L,\tilde{L})\leqslant d_{\mathrm{CKC}}(L',\tilde{L})\vee d_{\mathrm{CKC}}(L'',\tilde{L})$ when $L'\subset L\subset L''$, the d_{CKC}^T convergence easily follows from the pointwise convergence, from continuity of K_t and from monotonicity of K_t^n in t. □

4.8. Improving the convergence topology

In this subsection we complete the proof of Theorem 1.2. There are examples showing that the convergence of the Loewner driving term does not imply the uniform convergence of the paths parameterized by capacity. (See [LSW4, §3.4].) Therefore, we will need to apply other considerations. Before embarking on the proof, we note that when $a,b\geqslant\lambda$ the trace of SLE$(4; a/\lambda-1, b/\lambda-1)$ is a simple path that does not hit $\mathbb{R}$, except at its starting point. Indeed, note first that the force points are moving monotonically away from one another. By comparison with a Bessel process, for example, it is easy to see that the trace does not hit the real line at any time $t>0$. It also does not hit itself, since $t\mapsto g_s(\gamma(t+s))$ has law that is mutually absolutely continuous with the path of SLE(4).

LEMMA 4.16. *Let $T>0$ and let $W_n\colon[0,T]\to\mathbb{R}$ be a sequence of continuous functions converging uniformly to a function $W\colon[0,T]\to\mathbb{R}$. Suppose that each W_n is the driving term of a Loewner evolution of a path $\gamma_n\colon[0,T]\to\overline{\mathbb{H}}$, and W is the driving term of a Loewner evolution of a simple path $\gamma\colon[0,T]\to\overline{\mathbb{H}}$ satisfying $\gamma(0,T]\cap\mathbb{R}=\varnothing$. Then $\lim_{n\to\infty}\sup_{t\in[0,T]} d_{\mathrm{H}}(\gamma_n[0,t],\gamma[0,t])=0$, where d_{H} denotes the the Hausdorff metric.*

Proof. First note that $\operatorname{diam}\gamma_n[0,T]$ is bounded, because $\|W_n\|_\infty$ is bounded. Fix some $t\in[0,T]$, and let Γ denote a subsequential Hausdorff limit of $\gamma_n[0,t]$. It suffices to prove that $\Gamma=\gamma[0,t]$. By Lemma 4.15, we know that for every $\varepsilon>0$ we have $\Gamma\cap\overline{\mathbb{H}}_\varepsilon(\gamma[0,t])=\varnothing$. Since γ is a simple path satisfying $\gamma(0,T]\cap\mathbb{R}=\varnothing$, it follows that $\bigcup_{\varepsilon>0}\overline{\mathbb{H}}_\varepsilon(\gamma[0,t])=\mathbb{H}\setminus\gamma[0,t]$, which implies that $\Gamma\subset\mathbb{R}\cup\gamma[0,t]$. Fix some $z_1\in\mathbb{R}\setminus\gamma[0,t]=\mathbb{R}\setminus\{\gamma(0)\}$. By the continuity in W and z of the solutions of Loewner's equation (1.3), it follows that there is a neighborhood V of z_1 such that $V\cap\gamma_n[0,T]=\varnothing$ for all sufficiently large n. This implies that $z_1\notin\Gamma$, and hence $\Gamma\subset\gamma[0,t]$.

Now let $t'\in[0,t]$. By Lemma 4.15 again, for every $\varepsilon>0$ and every n sufficiently large, $\gamma(t')\notin\overline{\mathbb{H}}_\varepsilon(\gamma_n[0,t])$, which means that every path connecting $\gamma(t')$ to ∞ in $\overline{\mathbb{H}}$ must come within distance ε from $\gamma_n[0,t]$. Thus, every such path must intersect Γ. Since $\Gamma\subset\gamma[0,t]$ is closed, this implies that $\gamma(t')\in\Gamma$. Therefore, $\Gamma=\gamma[0,t]$; that is,

$$\lim_{n\to\infty} d_{\mathrm{H}}(\gamma_n[0,t],\gamma[0,t])=0.$$

Since $\gamma_n[0,t]$ and $\gamma[0,t]$ are monotone increasing in t and $\gamma[0,t]$ is continuous in t with respect to d_{H}, it easily follows that $\lim_{n\to\infty}\sup_{t\in[0,T]} d_{\mathrm{H}}(\gamma_n[0,t],\gamma[0,t])=0$. □

Here is an outline of the main ideas going into the proof of Theorem 1.2. Let γ^ϕ be the path $\phi\circ\gamma$ parameterized by half-plane capacity. The main step in the proof is to show that if we fix $T>0$, we have $\sup_{t\in[0,T]}|\gamma^\phi(t)-\gamma_{\mathrm{SLE}}(t)|\to0$ in probability. By Theorem 1.3 and Lemma 4.16, we get $\sup_{t\in[0,T]} d_{\mathrm{H}}(\gamma^\phi[0,t],\gamma_{\mathrm{SLE}}[0,t])\to0$ in probability (since γ_{SLE} is a simple path). We only need to rule out the possibility that γ^ϕ has significant (and fast) backtracking along γ_{SLE}. This is ruled out by invoking Lemma 3.17 and observing that the ϕ-image of the place where simple random walk (starting from a vertex near $\phi^{-1}(i)$) hits $\partial D(\gamma)$ can be close to any fixed segment of $\gamma_{\mathrm{SLE}}(0,T]$.

Proof of Theorem 1.2. Let γ^ϕ be the path $\phi\circ\gamma$ parameterized by half-plane capacity. Let $\delta,T>0$ and $r_D=\mathrm{rad}_{\phi^{-1}(i)}(D)$. Let W denote the Loewner driving term of γ_{SLE}. Since γ_{SLE} is almost surely a simple path, Lemma 4.16 implies that for every $\varepsilon>0$ there is some $\varepsilon'=\varepsilon'(\varepsilon,\gamma_{\mathrm{SLE}})>0$ such that if $\widetilde{W}$ is the driving term of a continuous path $\widetilde{\gamma}$ and $\sup_{t\in[0,T]}|\widetilde{W}_t-W_t|<\varepsilon'$, then $\sup_{t\in[0,T]} d_{\mathrm{H}}(\gamma_{\mathrm{SLE}}[0,t],\widetilde{\gamma}[0,t])<\varepsilon$. Moreover, it is not hard to see that ε' can be chosen as a measurable function of W. Hence, Theorem 1.3 implies that for every $\varepsilon_0>0$, if r_D is larger than some function of ε_0, δ, T, a and b, then there is a coupling of h and $\mathrm{SLE}(4;a/\lambda-1,b/\lambda-1)$ such that

$$\varrho:=\sup_{t\in[0,T]} d_{\mathrm{H}}(\gamma^\phi[0,t],\gamma_{\mathrm{SLE}}[0,t])<\varepsilon_0$$

with probability at least $1-\delta$. Without yet specifying ε_0, we assume that indeed $\gamma^\phi[0,T]$ and γ_{SLE} are so coupled. Let $\mathcal{A}_0$ be the event that $\varrho<\varepsilon_0$. Then $\mathbf{P}[\mathcal{A}_0]\geqslant1-\delta$.

Let $0<t_0<t_1<t_2<t_3\leqslant T$. We will show that under this coupling, if r_D is large, then with high probability $\gamma^\phi(t)$ is close to $\gamma_{\mathrm{SLE}}[t_0,T]$ for every $t\in[t_3,T]$. This will then imply that $\sup_{t\in[0,T]}|\gamma_{\mathrm{SLE}}(t)-\gamma^\phi(t)|$ is small.

Since γ_{SLE} is almost surely a simple path disjoint from $\{i\}$, there is a constant $s_0=s_0(t_0,t_1,t_2,t_3,T,\delta)>0$ such that $\mathbf{P}[\mathcal{A}_1]\geqslant 1-\delta$, where $\mathcal{A}_1$ is the event that

(1) the harmonic measure of $\gamma_{\mathrm{SLE}}[t_1,t_2]$ from i with respect to $\mathbb{H}\setminus\gamma_{\mathrm{SLE}}[0,\infty)$ is at least s_0,

(2) $\mathrm{dist}(\mathbb{R},\gamma_{\mathrm{SLE}}[t_0,T])>s_0$,

(3) $\mathrm{dist}(\gamma_{\mathrm{SLE}}[0,t_j],\gamma_{\mathrm{SLE}}[t_{j+1},T])>s_0$ for $j=0,1,2$,

(4) $\operatorname{diam}\gamma_{\mathrm{SLE}}[0,T]<1/s_0$, and

(5) $i\in\overline{\mathbb{H}}_{s_0}(\gamma_{\mathrm{SLE}}[0,T])$.

(It is tedious, but straightforward, to check that $\mathcal{A}_1$ is measurable.)

Consider a simple random walk S independent of h starting at a TG-vertex closest to $\phi^{-1}(i)$. Let τ_T be the first time t when $S(t)\in\partial D(\phi^{-1}\circ\gamma^\phi[0,T])$, and, as in §3.7, let $z_T:=S(\tau_T)$. We claim that for every $\varepsilon_1>0$, if r_D is sufficiently large and ε_0 is sufficiently small, then

$$\mathbf{P}[\mathrm{dist}(\phi(z_T),\gamma_{\mathrm{SLE}}[t_1,t_2])<\varepsilon_1\mid\gamma_{\mathrm{SLE}},\gamma^\phi[0,T]]>\tfrac{1}{2}s_0\quad\text{on }\mathcal{A}_0\cap\mathcal{A}_1. \tag{4.36}$$

To prove (4.36), first observe that conditional on γ_{SLE} such that $\mathcal{A}_1$ holds, a 2-dimensional Brownian motion $\widehat{S}$ started at i has probability at least s_0 to first hit $\gamma_{\mathrm{SLE}}\cup\mathbb{R}$ in $\gamma_{\mathrm{SLE}}[t_1,t_2]$. Moreover, if this happens, the Brownian motion is likely to stay within a compact subset $L\subset\mathbb{H}$ before hitting $\gamma_{\mathrm{SLE}}[t_1,t_2]$ and not to come arbitrarily close to γ_{SLE} far from its hitting point. On compact subsets of $\mathbb{H}$, the map $r_D^{-1}\phi^{-1}$ distorts distances by a bounded factor, by the Koebe distortion theorem [P, Theorem 1.3 and Corollary 1.4]. Since ϕ^{-1} takes a Brownian motion to a monotonically time-changed Brownian motion, by taking r_D large we may couple $r_D^{-1}S$ and $r_D^{-1}\phi^{-1}\circ\widehat{S}$ to stay arbitrarily close (until $\phi^{-1}\circ\widehat{S}$ hits ∂D) with high probability, up to a time change. Assuming that ϱ is arbitrarily small and taking (5) into account, we find that on $\mathcal{A}_1$ and given $(\gamma_{\mathrm{SLE}},\gamma^\phi[0,T])$, with conditional probability at least $\frac{5}{6}s_0$ the random walk gets to a vertex v where $\mathrm{dist}(\phi(v),\gamma_{\mathrm{SLE}}[t_1,t_2])$ is arbitrarily small before time τ_T. Now, on the event $\mathcal{A}_0$, $\gamma^\phi[0,T]$ has to be close by, and so we find from Lemma 2.1 that on $\mathcal{A}_0\cap\mathcal{A}_1$ and given $(\gamma_{\mathrm{SLE}},\gamma^\phi[0,T])$, with conditional probability at least $\frac{2}{3}s_0$ we have $\mathrm{dist}(\phi(z_T),\gamma_{\mathrm{SLE}}[t_1,t_2])$ arbitrarily small. This proves (4.36).

We take $\gamma^\phi[T,\infty)$ independent of γ_{SLE} given $\gamma^\phi[0,T]$ in the coupling of γ_{SLE} with γ. As in §3.5, let τ be the hitting time of S on $\partial D(\gamma)$. In (3.62) we choose $\varepsilon=\delta$, and get a corresponding $p>0$. Let $\mathcal{A}_2$ denote the event

$$\mathbf{P}[z_T=S_\tau\mid S,\gamma^\phi[0,T]]\geqslant p.$$

Then (3.62) reads $\mathbf{P}[\mathcal{A}_2]\geqslant 1-\delta$. In conjunction with (4.36) and $\mathbf{P}[\mathcal{A}_j]\geqslant 1-\delta$, $j=0,1$, this implies that

$$\mathbf{P}\big[\mathbf{P}[\operatorname{dist}(\phi(S_\tau),\gamma_{\mathrm{SLE}}[t_1,t_2])<\varepsilon_1 \mid \gamma_{\mathrm{SLE}},\gamma^\phi[0,T]]>\tfrac{1}{2}ps_0\big]\geqslant 1-3\delta. \tag{4.37}$$

Conditional on S_τ, on the event $\mathcal{A}_1\cap\{\operatorname{dist}(\phi(S_\tau),\gamma_{\mathrm{SLE}}[t_1,t_2])<\varepsilon_1\}$ we invoke Lemma 3.17 with S_τ translated to 0 where the p in that lemma is chosen as $\frac{1}{2}ps_0\delta$ and the R is taken to be s_2r_D, where $s_2=s_2(s_0)>0$ is a small constant depending only on s_0. Note that the assumption necessary for the lemma that $\operatorname{dist}(S_\tau,\{v_0\}\cup\partial D)>4R$ holds by (2), (4) and (5) in the definition of $\mathcal{A}_1$ and the fact that the distance distortion of $r_D^{-1}\phi^{-1}$ is bounded on compact subsets of $\mathbb{H}$. Let $\tilde{a}$ denote the a provided by the lemma, which is a function of s_0, a, b and δ. Set $r=R/(\tilde{a}+1)$. Then the lemma together with (4.37) imply that with probability $1-O(\delta)$ there is within distance ε_1 from $\gamma_{\mathrm{SLE}}[t_1,t_2]$ the ϕ-image of a vertex $v\in\partial D(\gamma)$ such that $\phi^{-1}\circ\gamma^\phi[0,T]$ has precisely two disjoint crossings of the annulus $\{z:r\leqslant|z-v|\leqslant R\}$. If this happens, let z_1 be a point in $\gamma_{\mathrm{SLE}}[t_1,t_2]$ closest to such a $\phi(v)$.

Let r' denote the lower bound we get on $\{|\phi(v)-\phi(z)|:|v-z|\geqslant r\}$ which follows from the bounded distortion of $r_D\phi$. We may also assume that $|\phi(v)-\phi(z)|\leqslant\frac{1}{8}s_0$ when $|v-z|\leqslant R$. Now take $\varepsilon_1=\frac{1}{4}r'$ and let $s_3=s_3(\delta,r')\in\big(0,\frac{1}{8}r'\big)$ be so small that with probability at least $1-\delta$ for every ball of radius $\frac{1}{4}r'$ centered at a point $z_0\in\gamma_{\mathrm{SLE}}[t_0,T]$ the distance outside of the ball $B\big(z_0,\frac{1}{4}r'\big)$ between the two connected components of $\mathbb{R}\cup(\gamma_{\mathrm{SLE}}[0,T]\setminus\{z_0\})$ is at least s_3. Note that when this is the case, every path connecting these two components outside of $B\big(z_0,\frac{1}{4}r'\big)$ has to intersect $\overline{\mathbb{H}}_{s_3/3}(\gamma_{\mathrm{SLE}}[0,T])$. Consequently, if additionally $\varrho<\frac{1}{3}s_3$, then $\gamma^\phi[0,T]$ cannot contain an arc whose endpoints are within distance $\frac{1}{4}s_3$ of these two components, unless the arc visits the ball $B\big(z_0,\frac{1}{4}r'\big)$. If ϱ is sufficiently small, then there is some $t_3'\leqslant t_3$ such that $|\gamma_{\mathrm{SLE}}(t_3)-\gamma^\phi(t_3')|<\frac{1}{4}s_3$. Now choose $z_0:=z_1$, for the previous paragraph. The path $\gamma^\phi[0,t_3']$ must pass through the ball $B\big(z_0,\frac{1}{4}r'\big)$, and therefore $\phi^{-1}\circ\gamma^\phi[0,t_3']$ contains two disjoint crossings of the annulus $\{z:r\leqslant|z-v|\leqslant R\}$. If we assume that $\phi^{-1}\circ\gamma^\phi[0,T]$ has no more than two disjoint crossings of this annulus (which happens with probability at least $1-O(\delta)$), it follows that for $t\in[t_3,T]$ the point $\gamma^\phi(t)$ is closer to $\gamma_{\mathrm{SLE}}[t_0,T]$ than to $\gamma_{\mathrm{SLE}}[0,t_0]\cup\mathbb{R}$. Since in the above δ, t_0 and t_3 are arbitrary subject to the constraint $0<t_0<t_3$ and $\delta>0$, the claimed uniform convergence in $[0,T]$ follows.

To prove convergence in law with respect to the uniform d_* metric, it suffices to show that for every radius $r_1>0$ there is some $r_2>r_1$ such that γ^ϕ is unlikely to return to $B(0,r_1)$ after its first exit from $B(0,r_2)$. For this proof, we will use the conformal invariance of extremal length (see [A]).

Fix some $r_1>0$. The extremal length of the collection of arcs in the half-annulus $A:=\{z\in\overline{\mathbb{H}}:r_1<|z|<r_2\}$, which connect $\mathbb{R}_+$ with $\mathbb{R}_-$, tends to zero as $r_2\to\infty$. Let $A':=$

$\phi^{-1}(A)$, $\partial_j := \phi^{-1}(\{z \in \overline{\mathbb{H}} : |z| = r_j\})$, $j = 1, 2$, $L := \mathrm{dist}(\partial_1, \partial_2; A')$ and $L' := \mathrm{dist}(\partial_+, \partial_-; A')$. By conformal invariance of extremal length, it follows that $L'/L \to 0$ as $r_2 \to \infty$, uniformly in D. (Otherwise, the metric which is equal to the Euclidean metric in the ball of radius $3L$ centered on a point in an arc of length at most $2L$ from ∂_1 to ∂_2 in A', and is zero outside this ball, contradicts the extremal length going to zero.)

Let $\beta \subset A'$ be an arc of diameter at most $2L'$ connecting ∂_+ and ∂_-. Let

$$L_1 := \mathrm{dist}(\beta, \partial_1; A') \quad \text{and} \quad L_2 := \mathrm{dist}(\beta, \partial_2; A').$$

Since $L'/L \to 0$ as $r_2 \to \infty$, we have $L_1 \geqslant \frac{1}{3}L$ or $L_2 \geqslant \frac{1}{3}L$ if r_2 is sufficiently large. Suppose first that $L_2 \geqslant \frac{1}{3}L$. Let s_0 denote the first time such that $|\phi \circ \gamma(s_0)| = r_2$. Then there are two connected components β_1 and β_2 of $\beta \setminus \gamma[0, s_0]$ such that $\gamma[0, s_0] \cup \beta_1 \cup \beta_2$ separates $\phi^{-1}(B(0, r_1))$ from y_∂ in D. Now the proof of Theorem 3.22 shows that

$$\mathbf{P}[\gamma \cap \beta_1 \neq \varnothing \mid \gamma[0, s_0]] < \delta$$

if L_2/L' is sufficiently large. (Hence when $L_2 \geqslant \frac{1}{3}L$ and r_2 is sufficiently large.) A similar estimate holds with β_2. Thus, with probability at most $O(\delta)$, γ^ϕ contains two disjoint crossings of the annulus $r_1 \leqslant |z| \leqslant r_2$. On the other hand, if $L_2 < \frac{1}{3}L$ and $L_1 \geqslant \frac{1}{3}L$, then we may apply the same argument to the reversal of γ (or else slightly modify the way the analog of Theorem 3.22 is proved) to reach the same conclusion. This completes the proof. □

5. Other lattices

In this section we describe the modifications necessary to adapt the proofs of Theorems 1.3 and 1.2 to the more general framework of Theorem 1.4.

Before we go into the actual proof, a few words need to be said about the properties of the weighted random walk on $\mathfrak{G}$ and its convergence to Brownian motion. Fix some vertex v_0 in $\mathfrak{G}$, and let V_0 denote its orbit under the group generated by the two translations T_1 and T_2 preserving $\mathfrak{G}$. If the walk starts at v_0, then a new Markov chain is obtained by looking at the sequence of vertices in V_0 that the walk visits. A simple path reversal argument shows that for this new Markov chain the transition probability from v to u is the same as the transition probability from u to v, for every pair of vertices $v, u \in V_0$. Also observe that the $\mathbb{R}^2$-length of a single step has an exponential tail. This is enough to show that the Markov chain on V_0, rescaled appropriately in time and space, converges to a linear image of Brownian motion (and it is not hard to verify that the linear transformation is non-singular). Moreover, the few properties of the simple random walk on TG that we have used in the course of the paper are easily verified for this Markov chain on V_0 and easily translated to the weighted walk on $\mathfrak{G}$.

Proof of Theorem 1.4. Very few changes are needed to adapt the proof. Let $\mathfrak{G}$ denote the original lattice consisting of only edges of positive weight, and let $\overline{\mathfrak{G}}$ denote the triangulation of $\mathfrak{G}$, as described in §1.5. Let $\overline{\mathfrak{G}}^*$ denote the planar dual of $\overline{\mathfrak{G}}$.

The statement and proof of Lemma 3.1 requires some changes, because in the more general setup it is not true that every vertex adjacent to an interface on the right has a $\mathfrak{G}$-neighbor on the left of the interface (and similarly in the other direction). Thus, in the revised version of the lemma, the assumption that each vertex in V_+ neighbors with a vertex in $V_-\cup V_\partial$ and every vertex in V_- neighbors with a vertex in $V_+\cup V_\partial$ needs to be replaced by the assumption that for some constant m, depending on the lattice, for every vertex $v\in V_+$ the $\mathfrak{G}$-graph-distance from v to $V_-\cup V_\partial$ is at most m, and symmetrically for vertices in V_-. This change requires a few extra lines in the proof of (3.1). Let M_j be the maximum of $\mathbf{E}[e^{h(v)}|\mathcal{K}]$ for vertices in V_+ at $\mathfrak{G}$-distance at most j from $V_-\cup V_\partial$. Every vertex at $\mathfrak{G}$-distance $j>0$ from $V_-\cup V_\partial$ has a $\mathfrak{G}$-neighbor at $\mathfrak{G}$-distance $j-1$ from $V_-\cup V_\partial$. Therefore, the proof of (3.8) now gives

$$M_j \leqslant O(1)M_m^c M_{j-1}^{1-c}+O(1),$$

where $c<1$ is some constant depending only on the lattice and its edge weights. We can certainly drop the trailing additive $O(1)$. Induction on j now gives

$$M_j \leqslant O(1)^{q_j} M_m^{c\,q_j} M_0^{(1-c)^j},$$

where $q_j=1+(1-c)+\ldots+(1-c)^{j-1}=(1-(1-c)^j)/c$. When $j=m$, this reads

$$M_m^{(1-c)^m} \leqslant O(1)^{q_m} M_0^{(1-c)^m},$$

Clearly, $M_0\leqslant e^{\bar{\Lambda}}$, and the bound $M_m=O_{m,\bar{\Lambda}}(1)$ follows. A corresponding bound clearly also holds for $\mathbf{E}[e^{-h(v)}|\mathcal{K}]$ when $v\in V_-$. The remainder of the proof of the analog of Lemma 3.1 proceeds without difficulty.

The proof of Lemma 3.2 needs to be similarly adapted, but essentially the same argument works.

The next point which requires adaptation is the definition of $\mathcal{Z}_0^\sigma$ in §3.5. Let Σ denote the set of pairs (v,e^*), where v is a vertex in $\mathfrak{G}$ and e^* is an edge in $\overline{\mathfrak{G}}^*$ that is dual to one of the edges incident with v in $\overline{\mathfrak{G}}$. If $\sigma=(v,e^*)\in\Sigma$, let $\mathcal{Z}^\sigma$ denote the event that the first vertex adjacent to γ that S hits is v and moreover $e^*\in\gamma$. Let Σ' be a collection of elements of Σ, one from each orbit under the group generated by the translations T_1 and T_2 preserving $\mathfrak{G}$. Let $\mathcal{Z}_0:=\bigcup_{\sigma\in\Sigma'}\mathcal{Z}^\sigma$. The proof then proceeds essentially unchanged, with $\mathcal{Z}^\sigma$ in place of $\mathcal{Z}_0^\sigma$ and with the modified definition for $\mathcal{Z}_0$. □

References

[A] Ahlfors, L. V., *Conformal Invariants: Topics in Geometric Function Theory.* McGraw-Hill Series in Higher Mathematics. McGraw-Hill, New York, 1973.

[B] Beffara, V., The dimension of the SLE curves. *Ann. Probab.*, 36:4 (2008), 1421–1452.

[BPZ] Belavin, A. A., Polyakov, A. M. & Zamolodchikov, A. B., Infinite conformal symmetry in two-dimensional quantum field theory. *Nuclear Phys. B*, 241 (1984), 333–380.

[BEF] Bricmont, J., El Mellouki, A. & Fröhlich, J., Random surfaces in statistical mechanics: roughening, rounding, wetting,.... *J. Stat. Phys.*, 42:5–6 (1986), 743–798.

[CN] Camia, F. & Newman, C. M., Two-dimensional critical percolation: the full scaling limit. *Comm. Math. Phys.*, 268 (2006), 1–38.

[Ca] Cardy, J., SLE for theoretical physicists. *Ann. Physics*, 318 (2005), 81–118.

[Co] Coniglio, A., Fractal structure of Ising and Potts clusters: exact results. *Phys. Rev. Lett.*, 62:26 (1989), 3054–3057.

[DMS] Di Francesco, P., Mathieu, P. & Sénéchal, D., *Conformal Field Theory.* Graduate Texts in Contemporary Physics. Springer, New York, 1997.

[DoSn] Doyle, P. G. & Snell, J. L., *Random Walks and Electric Networks.* Carus Mathematical Monographs, 22. Mathematical Association of America, Washington, DC, 1984.

[D] Duplantier, B., Two-dimensional fractal geometry, critical phenomena and conformal invariance. *Phys. Rep.*, 184 (1989), 229–257.

[DS1] Duplantier, B. & Saleur, H., Exact critical properties of two-dimensional dense self-avoiding walks. *Nuclear Phys. B*, 290 (1987), 291–326.

[DS2] — Winding-angle distributions of two-dimensional self-avoiding walks from conformal invariance. *Phys. Rev. Lett.*, 60:23 (1988), 2343–2346.

[F] Foltin, G., An alternative field theory for the Kosterlitz–Thouless transition. *J. Phys. A*, 34:26 (2001), 5327–5333.

[FS] Fröhlich, J. & Spencer, T., The Kosterlitz–Thouless transition in two-dimensional abelian spin systems and the Coulomb gas. *Comm. Math. Phys.*, 81 (1981), 527–602.

[Ga] Gawędzki, K., Lectures on conformal field theory, in *Quantum Fields and Strings: a Course for Mathematicians*, Vol. 1, 2 (Princeton, NJ, 1996/1997), pp. 727–805. Amer. Math. Soc., Providence, RI, 1999.

[Gi] Giacomin, G., Limit theorems for random interface models of Ginzburg–Landau $\nabla\phi$ type, in *Stochastic Partial Differential Equations and Applications* (Trento, 2002), Lecture Notes in Pure and Appl. Math., 227, pp. 235–253. Dekker, New York, 2002.

[GJ] Glimm, J. & Jaffe, A., *Quantum Physics.* Springer, New York, 1987.

[HK] Huber, G. & Kondev, J., Passive-scalar turbulence and the geometry of loops. *Bull. Amer. Phys. Soc.*, DCOMP Meeting 2001, Q2.008.

[I] Isichenko, M. B., Percolation, statistical topography, and transport in random media. *Rev. Modern Phys.*, 64:4 (1992), 961–1043.

[Ka] Kadanoff, L. P., Lattice Coulomb gas representations of two-dimensional problems. *J. Phys. A*, 11:7 (1978), 1399–1417.

[KN] Kager, W. & Nienhuis, B., A guide to stochastic Löwner evolution and its applications. *J. Stat. Phys.*, 115:5–6 (2004), 1149–1229.

[KS] Karatzas, I. & Shreve, S. E., *Brownian Motion and Stochastic Calculus.* Graduate Texts in Mathematics, 113. Springer, New York, 1988.

[Ke] KENYON, R., Dominos and the Gaussian free field. *Ann. Probab.*, 29:3 (2001), 1128–1137.

[KDH] KONDEV, J., DU, S. & HUBER, G., Two-dimensional passive-scalar turbulence and the geometry of loops. *Bull. Amer. Phys. Soc.*, March Meeting 2002, U4.003.

[KH] KONDEV, J. & HENLEY, C. L., Geometrical exponents of contour loops on random Gaussian surfaces. *Phys. Rev. Lett.*, 74:23 (1995), 4580–4583.

[KHS] KONDEV, J., HENLEY, C. L. & SALINAS, D. G., Nonlinear measures for characterizing rough surface morphologies. *Phys. Rev. E*, 61 (2000), 104–125.

[Ko] KOSTERLITZ, J. M., The d-dimensional Coulomb gas and the roughening transition. *J. Phys. C*, 10:19 (1977), 3753–3760.

[KT] KOSTERLITZ, J. M. & THOULESS, D. J., Ordering, metastability and phase transitions in two-dimensional systems. *J. Phys. C*, 6:7 (1973), 1181–1203.

[L1] LAWLER, G. F., *Intersections of Random Walks*. Probability and its Applications. Birkhäuser, Boston, MA, 1991.

[L2] — Strict concavity of the intersection exponent for Brownian motion in two and three dimensions. *Math. Phys. Electron. J.*, 4 (1998), Paper 5, 67 pp.

[L3] — An introduction to the stochastic Loewner evolution, in *Random Walks and Geometry*, pp. 261–293. de Gruyter, Berlin, 2004.

[L4] — *Conformally Invariant Processes in the Plane*. Mathematical Surveys and Monographs, 114. Amer. Math. Soc., Providence, RI, 2005.

[LSW1] LAWLER, G. F., SCHRAMM, O. & WERNER, W., Values of Brownian intersection exponents. I. Half-plane exponents. *Acta Math.*, 187 (2001), 237–273.

[LSW2] — Values of Brownian intersection exponents. III. Two-sided exponents. *Ann. Inst. H. Poincaré Probab. Statist.*, 38 (2002), 109–123.

[LSW3] — Conformal restriction: the chordal case. *J. Amer. Math. Soc.*, 16 (2003), 917–955.

[LSW4] — Conformal invariance of planar loop-erased random walks and uniform spanning trees. *Ann. Probab.*, 32:1B (2004), 939–995.

[M] MANDELBROT, B., How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science*, 156 (1967), 636–638.

[NS] NADDAF, A. & SPENCER, T., On homogenization and scaling limit of some gradient perturbations of a massless free field. *Comm. Math. Phys.*, 183 (1997), 55–84.

[N1] NIENHUIS, B., Exact critical point and critical exponents of O(n) models in two dimensions. *Phys. Rev. Lett.*, 49:15 (1982), 1062–1065.

[N2] — Critical behavior of two-dimensional spin models and charge asymmetry in the Coulomb gas. *J. Stat. Phys.*, 34:5–6 (1984), 731–761.

[dN] DEN NIJS, M., Extended scaling relations for the magnetic critical exponents of the Potts model. *Phys. Rev. B*, 27:3 (1983), 1674–1679.

[P] POMMERENKE, C., *Boundary Behaviour of Conformal Maps*. Grundlehren der Mathematischen Wissenschaften, 299. Springer, Berlin–Heidelberg, 1992.

[RY] REVUZ, D. & YOR, M., *Continuous Martingales and Brownian Motion*. Grundlehren der Mathematischen Wissenschaften, 293. Springer, Berlin–Heidelberg, 1999.

[RS] ROHDE, S. & SCHRAMM, O., Basic properties of SLE. *Ann. of Math.*, 161 (2005), 883–924.

[SD] SALEUR, H. & DUPLANTIER, B., Exact determination of the percolation hull exponent in two dimensions. *Phys. Rev. Lett.*, 58:22 (1987), 2325–2328.

[Sch] SCHRAMM, O., Scaling limits of loop-erased random walks and uniform spanning trees. *Israel J. Math.*, 118 (2000), 221–288.

[SS] SCHRAMM, O. & SHEFFIELD, S., Harmonic explorer and its convergence to SLE_4. *Ann. Probab.*, 33:6 (2005), 2127–2148.

[Sh] SHEFFIELD, S., Gaussian free fields for mathematicians. *Probab. Theory Related Fields*, 139 (2007), 521–541.

[Sm] SMIRNOV, S., Critical percolation in the plane: conformal invariance, Cardy's formula, scaling limits. *C. R. Acad. Sci. Paris Sér. I Math.*, 333 (2001), 239–244.

[Sp] SPENCER, T., Scaling, the free field and statistical mechanics, in *The Legacy of Norbert Wiener: A Centennial Symposium* (Cambridge, MA, 1994), Proc. Sympos. Pure Math., 60, pp. 373–389. Amer. Math. Soc., Providence, RI, 1997.

[W] WERNER, W., Random planar curves and Schramm–Loewner evolutions, in *Lectures on Probability Theory and Statistics*, Lecture Notes in Math., 1840, pp. 107–195. Springer, Berlin–Heidelberg, 2004.

ODED SCHRAMM
was at the Theory Group
of Microsoft Research
One Microsoft Way
Redmond, WA 98052-6399
U.S.A.

SCOTT SHEFFIELD
Courant Institute
New York University
251 Mercer Street
New York, NY 10012
U.S.A.
sheff@math.nyu.edu
sheffield@math.mit.edu

Received October 18, 2006

Commun. Math. Phys. 288, 43–53 (2009)
Digital Object Identifier (DOI) 10.1007/s00220-009-0731-6

Communications in
Mathematical Physics

Conformal Radii for Conformal Loop Ensembles

Oded Schramm[1], **Scott Sheffield**[2,3,⋆], **David B. Wilson**[1]

[1] Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA
[2] Courant Institute, New York University, 251 Mercer Street, New York, NY 10021, USA
[3] Department of Mathematics, M. I. T., 77 Massachusetts Ave., Cambridge, MA 02139, USA

Received: 20 September 2007 / Accepted: 10 November 2008
Published online: 26 February 2009 – © Springer-Verlag 2009

Abstract: The conformal loop ensembles CLE_κ, defined for $8/3 \leq \kappa \leq 8$, are random collections of loops in a planar domain which are conjectured scaling limits of the $O(n)$ loop models. We calculate the distribution of the conformal radii of the nested loops surrounding a deterministic point. Our results agree with predictions made by Cardy and Ziff and by Kenyon and Wilson for the $O(n)$ model. We also compute the expectation dimension of the CLE_κ **gasket**, which consists of points not surrounded by any loop, to be

$$2 - \frac{(8-\kappa)(3\kappa-8)}{32\kappa},$$

which agrees with the fractal dimension given by Duplantier for the $O(n)$ model gasket.

1. Introduction

The conformal loop ensembles CLE_κ, defined for all $8/3 \leq \kappa \leq 8$, are random collections of loops in a simply connected planar domain $D \subsetneq \mathbb{C}$. They were defined and constructed from branching variants of SLE_κ in [She06], where they were conjectured to be the scaling limits of various random loop models from statistical physics, including the so-called $O(n)$ loop models with

$$n = -2\cos(4\pi/\kappa), \tag{1}$$

see e.g. [KN04] for an exposition. This paper is a sequel to [She06]. We will state the results about CLE_κ from [She06] that we need for this paper (namely Propositions 1 and 2), but we will not repeat the definition of CLE_κ here.

When $8/3 < \kappa < 8$, CLE_κ is almost surely a countably infinite collection of loops. CLE_8 is a single space-filling loop almost surely and $\mathrm{CLE}_{8/3}$ is almost surely empty.

⋆ Partially supported by NSF grant DMS0403182.

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_33, © Springer Science+Business Media, LLC 2011

CLE_6 is the scaling limit of the cluster boundaries of critical site percolation on the triangular lattice [CN06,Smi01,CN07]. We will henceforth assume $8/3 < \kappa < 8$.

For each $z \in D$, we inductively define L_k^z to be the outermost loop surrounding z when the loops $L_1^z, \ldots, L_{k-1}^z$ are removed (provided such a loop exists). For each deterministic $z \in D$, the loops L_k^z exist for all $k \geq 1$ with probability one. Define $A_0^z = D$ and let A_k^z be the component of $D \backslash L_k^z$ that contains z. The **conformal gasket** is the random closed set Γ consisting of points that are not surrounded by any loop of an instance of CLE_κ, i.e. the set of points for which L_1^z does not exist.

If D is a simply connected planar domain and $z \in D$, the **conformal radius of** D **viewed from** z is defined to be $\mathrm{CR}(D, z) := |g'(z)|^{-1}$, where g is any conformal map from D to the unit disk $\mathbb{D}$ that sends z to 0. The following is immediate from the construction in [She06]:

Proposition 1. *Let D be a simply connected bounded planar domain, and consider a CLE_κ on D for some $8/3 < \kappa < 8$. Then Γ is almost surely the closure of the set of points that lie on an outermost loop (i.e., a loop of the form L_1^z for some z). Conditioned on the outermost loops, the law of the remaining loops is given by an independent CLE_κ in each component of $D \backslash \Gamma$. For $z \in D$ and $k = 1, 2, 3, \ldots$, define*

$$B_k^z := \log \mathrm{CR}(A_{k-1}^z, z) - \log \mathrm{CR}(A_k^z, z).$$

For any fixed z, the B_k^z's are i.i.d. random variables.

Various authors in the physics literature have used heuristic arguments (based on the so-called Coulomb gas method) to calculate properties of the scaling limits of statistical physical loop models, including the $O(n)$ models, based on certain conformal invariance hypotheses of these limits. Although the scaling limits of the $O(n)$ models have not been shown to exist, there is strong evidence that if they do exist they must be CLE_κ. (For example, there is heuristic evidence that any scaling limit of the $O(n)$ models should be in some sense conformally invariant; it is shown in [She06,SW08b,SW08a] that any random loop ensemble satisfying certain hypotheses including conformal invariance and a Markov-type property must be a CLE_κ.) It is therefore natural to interpret these calculations as predictions about the behavior of the CLE_κ.

Cardy and Ziff [CZ03] predicted and experimentally verified the expected number of loops surrounding a point in the $O(n)$ model, which, in light of (1), may be interpreted as a prediction of the expectation of B_z^k:

$$\frac{1}{\mathbb{E}[B_k^z]} = \frac{(\kappa/4 - 1)\cot(\pi(1 - 4/\kappa))}{\pi}. \tag{2}$$

Kenyon and Wilson [KW04] went further and predicted the distribution of B_k^z, giving its moment generating function

$$\mathbb{E}[\exp(\lambda B_k^z)] = \frac{-\cos(4\pi/\kappa)}{\cos\left(\pi\sqrt{(1 - 4/\kappa)^2 + 8\lambda/\kappa}\right)}, \tag{3}$$

for λ satisfying $\mathrm{Re}\lambda < 1 - \frac{2}{\kappa} - \frac{3\kappa}{32}$, and density function

$$\frac{d}{dx}\Pr[B_k^z < x] = \frac{-\kappa\cos(4\pi/\kappa)}{4\pi}\sum_{j=0}^{\infty}(-1)^j(j + 1/2)\exp\left[-\frac{(j + 1/2)^2 - (1 - 4/\kappa)^2}{8/\kappa}x\right]. \tag{4}$$

The main result of this paper is Theorem 1, which confirms these predictions. In the special case $\kappa = 6$, this prediction for the law of B_k^z was independently confirmed by Dubédat [Dub05].

Theorem 1. *Let f_κ denote the density function for the first time that a standard Brownian motion started at 0 exits the interval $(-2\pi/\sqrt{\kappa}, 2\pi/\sqrt{\kappa})$. Then for $8/3 < \kappa < 8$, the density function for B_k^z is*

$$\frac{d}{dx}\Pr[B_k^z < x] = -f_\kappa(x)\cos(4\pi/\kappa)\exp\left[\frac{(\kappa-4)^2}{8\kappa}x\right]. \tag{5}$$

The equivalence of the formulations (4) and (5), and the fact that they imply (3), follows from a calculation of Ciesielski and Taylor, who showed that the exit-time distribution of a Brownian motion from the center of a 1-dimensional ball of radius r has a moment generating function of $1/\cos\sqrt{2r^2\lambda}$, and who gave two series expansions (one in powers of e^{-x} and the other in powers of $e^{-1/x}$) for its density function [CT62, Theorem 2 and Eq. 2.22] (see also [BS02, Eqs. 1.3.0.1 and 1.3.0.2]). Since the Fourier transform is invertible on $L^2(\mathbb{R})$, the equivalence of (3) and (4) follows by considering the moment generating function restricted to the imaginary line $\mathrm{Re}\lambda = 0$.

Duplantier [Dup90] predicted the fractal dimension of the gasket Γ associated with the $O(n)$ model to be

$$\frac{3\kappa}{32} + 1 + \frac{2}{\kappa},$$

where as usual $n = -2\cos(4\pi/\kappa)$. We partially confirm this prediction by giving the expectation dimension of the gasket associated with CLE_κ. The **expectation dimension** of a random bounded set A is defined to be

$$D_{\mathbb{E}}(A) = \lim_{\varepsilon\to 0}\frac{\log \mathbb{E}[\text{minimal number of balls of radius } \varepsilon \text{ required to cover } A]}{|\log\varepsilon|},$$

provided the limit exists. The expectation dimension upper bounds the Hausdorff dimension.

Theorem 2. *Let Γ be the gasket of a CLE_κ in the unit disk with $\kappa \in (8/3, 8)$. Then*

$$\mathbb{E}[\textit{minimal number of balls of radius } \varepsilon \textit{ required to cover } \Gamma] \asymp \left(\frac{1}{\varepsilon}\right)^{\frac{3\kappa}{32}+1+\frac{2}{\kappa}}.$$

In particular, the expectation dimension of the gasket Γ is $\frac{3\kappa}{32} + 1 + \frac{2}{\kappa}$.

Here, $\asymp$ denotes equivalence up to multiplicative constants. Lawler, Schramm, and Werner [LSW02] studied the percolation gasket (associated with CLE_6), effectively proving Theorem 2 in the case $\kappa = 6$. More generally, they studied how long it takes for a radial SLE_κ to surround the origin when $\kappa > 4$, and their results implicitly imply Theorem 2 when $\kappa > 4$; see the remark in Sect. 2.1 for further discussion.

We conclude our introduction by noting that the gasket dimension described above plays an important role in the physics literature, where it is related (at least heuristically) to the exponents of magnetization and multipoint correlation functions in critical lattice models. We briefly describe this connection in the case of the q-state Potts model on the square lattice. More details and references are found in [She06, Car07, Gri06].

A sample from the q-state Potts model on a connected planar graph G is a random function $\sigma : V \to \{1, 2, \ldots, q\}$, where V is the set of vertices of G and the image values $1, 2, \ldots, q$ are often called spins. If the boundary vertices (those on the boundary of the unbounded face) of G are all assigned a particular value (say b), then using the standard FK random cluster decomposition [FK72], one may construct a sample from the Potts model as follows:

1. Sample a random subgraph G' of G containing all boundary edges (edges on the boundary of the unbounded face), with probability proportional to
$$q^{\#\text{ components of } G'} \left(\frac{p}{1-p}\right)^{\#\text{ edges of } G'},$$
where $0 < p < 1$ is a parameter.The law of G' is called the FK random cluster model with parameters p and q. Call the component of G' which contains the boundary vertices of G the **FK gasket**.
2. Set $\sigma(v) = b$ for each v in the FK gasket, and independently assign one of the q states uniformly at random to each of the remaining connected components of G' (assigning all vertices in that component the corresponding state).

The "magnetization" at an interior vertex v of G (i.e., the probability that $\sigma(v) = j$ minus $1/q$) is proportional to the probability that v is in the FK gasket. Given distinct vertices v and w, the covariance of $\sigma(v)$ and $\sigma(w)$ is proportional to the probability that both v and w lie in the same component of G'.

We now restrict to the case in which G is a finite piece of the square grid in the plane and the parameter p satisfies $p/(1-p) = \sqrt{q}$. (With this choice of p, the FK model is self-dual and believed to be critical, see e.g., [Gri06, Chap. 6].)

It is shown in [She06] that if $q \leq 4$ and certain other hypotheses including conformal invariance hold, then the scaling limit of the set of boundaries between clusters and dual clusters in the critical FK random cluster models discussed above must be given by CLE_κ for the κ satisfying $q = 4\cos^2(4\pi/\kappa)$ and $4 \leq \kappa \leq 8$. Assuming these hypotheses, the scaling limit of the discrete gasket is precisely the continuum CLE_κ gasket.

A heuristic ansatz is that the law of the critical FK gasket should have similar properties as the law of the set of squares in a fine grid which intersect the continuum gasket. If this heuristic holds, then when G is a bounded domain intersected with a square grid with spacing ε, the magnetization at a vertex v of macroscopic distance from the boundary in the discrete model will be on the order of ε^{2-d}, where d is the limiting expectation dimension of the continuum gasket. Similarly, the covariance between $\sigma(v)$ and $\sigma(w)$, for two macroscopically separated vertices v and w, should be on the order of $\varepsilon^{2(2-d)}$ (since in the continuum model, the set of pairs v and w which lie in the same continuum spin cluster has dimension $2d$; see [She06]).

2. Diffusions and Martingales

2.1. Reduction to a diffusion problem. Let $B_t : [0, \infty) \to \mathbb{R}$ be a standard Brownian motion and let $\theta_t : [0, \infty) \to [0, 2\pi]$ be a random continuous process on the interval $[0, 2\pi]$ that is instantaneously reflecting at its endpoints (i.e., the set $\{t : \theta_t \in \{0, 2\pi\}\}$ has Lebesgue measure zero almost surely) and evolves according to the SDE

$$d\theta_t = \frac{\kappa - 4}{2} \cot(\theta_t/2)\, dt + \sqrt{\kappa}\, dB_t \tag{6}$$

on each interval of time for which $\theta_t \notin \{0, 2\pi\}$. In other words, θ_t is a random continuous process adapted to the filtration of B_t which almost surely satisfies

$$\frac{\partial}{\partial t}(\theta_t - \sqrt{\kappa}\, B_t) = \frac{\kappa - 4}{2} \cot(\theta_t/2)$$

for all t for which the right hand side is well defined. The law of this process is uniquely determined by θ_0 [She06], and we also have the following from [She06]:

Proposition 2. *When* $8/3 < \kappa < 8$, *the law of* B_k^z *is the same as the law of* $\inf\{t : \theta_t = 2\pi\}$ *for the diffusion* (6) *started at* $\theta_0 = 0$.

It is convenient to lift the process θ_t so that, rather than taking values in $[0, 2\pi]$, it takes values in all of $\mathbb{R}$. Let $R : \mathbb{R} \to [0, 2\pi]$ be the piecewise affine map for which $R(x) = |x|$ when $x \in [-2\pi, 2\pi]$ and $R(4\pi + x) = R(x)$ for all x. Given θ_t, we can generate a continuous process $\tilde{\theta}_t$ with $R(\tilde{\theta}_t) = \theta_t$ in such a way that for each component (t_1, t_2) of the set $\{t : \theta_t \notin 2\pi\mathbb{Z}\}$, we independently toss a fair coin to decide whether $\tilde{\theta}_t > \tilde{\theta}_{t_1}$ or $\tilde{\theta}_t < \tilde{\theta}_{t_1}$ on that component. The θ_t together with these coin tosses (for each interval of $\{t : \theta_t \notin 2\pi\mathbb{Z}\}$) determine $\tilde{\theta}_t$ uniquely.

This $\tilde{\theta}_t$ is still a solution to (6) provided we modify B_t in such a way that dB_t is replaced with $-dB_t$ on those intervals for which $d\tilde{\theta}_t = -d\theta_t$. (This modification does not change the law of B_t.) In the remainder of the text, we will drop the $\tilde{\theta}_t$ notation and write θ_t for the lifted process on $\mathbb{R}$.

Remark. A very similar diffusion process was studied by Lawler, Schramm, and Werner [LSW02], namely

$$d\theta_t = \cot(\theta_t/2)\, dt + \sqrt{\kappa}\, dB_t, \tag{7}$$

which is the same as (6) but without the factor of $(\kappa - 4)/2$, and they too studied the time for the diffusion to reach $\theta_t = 2\pi$ when started at $\theta_0 = 0$. Diffusions 6 and 7 are identical when $\kappa = 6$, and Diffusion 6 (for $4 < \kappa < 8$) is given by Diffusion 7 (for $4 < \kappa < \infty$) upon substituting $\kappa \to (2\kappa)/(\kappa - 4)$ and scaling time by $t \to t(\kappa - 4)/2$. However, Diffusion 6 is a more singular Bessel-type process when $\kappa \leq 4$, requiring additional technical analysis to deal with the times at which process is at 0 (see e.g. Lemma 3). Furthermore, only the large-time asymptotic decay rate of the hitting time distribution (used in the proof of Theorem 2) is given in [LSW02], and additional effort is required to obtain the precise hitting time distribution provided in Theorem 1.

2.2. Local martingales. Recall the hypergeometric function defined by

$$F(a, b; c; z) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n n!} z^n,$$

where $a, b, c \in \mathbb{C}$ are parameters, $c \notin -\mathbb{N}$ (where $\mathbb{N} = \{0, 1, 2, \dots\}$), and $(\ell)_n$ denotes $\ell(\ell + 1) \cdots (\ell + n - 1)$. This definition holds for $z \in \mathbb{C}$ when $|z| < 1$, and it may be

defined by analytic continuation elsewhere (though it is then not always single-valued). We define for $\lambda \in \mathbb{C}$,

$$M^e_{\kappa,\lambda}(\theta) = F\left(1 - \tfrac{4}{\kappa} + \sqrt{\left(1-\tfrac{4}{\kappa}\right)^2 + \tfrac{8\lambda}{\kappa}}, 1 - \tfrac{4}{\kappa} - \sqrt{\left(1-\tfrac{4}{\kappa}\right)^2 + \tfrac{8\lambda}{\kappa}}; \tfrac{3}{2} - \tfrac{4}{\kappa}; \sin^2\tfrac{\theta}{4}\right),$$

$$M^o_{\kappa,\lambda}(\theta) = F\left(1 - \tfrac{2}{\kappa} + \sqrt{\left(\tfrac{1}{2}-\tfrac{2}{\kappa}\right)^2 + \tfrac{2\lambda}{\kappa}}, 1 - \tfrac{2}{\kappa} - \sqrt{\left(\tfrac{1}{2}-\tfrac{2}{\kappa}\right)^2 + \tfrac{2\lambda}{\kappa}}; \tfrac{3}{2}; \cos^2\tfrac{\theta}{2}\right)\cos\tfrac{\theta}{2}$$

(where the formula for $M^e_{\kappa,\lambda}(\theta)$ makes sense whenever $\kappa \neq \frac{8}{3}, \frac{8}{5}, \frac{8}{7}, \dots$). There is some ambiguity in the choice of square root, but since $F(b,a;c;z) = F(a,b;c;z)$, as long as the same choice of square root is made for both occurrences, there is no ambiguity in these definitions.

Lemma 1. *For the diffusion (6) with $\kappa > 0$, let T be the first time at which $\theta_t \in 2\pi\mathbb{Z}$, and let $\bar{t} = \min(t, T)$. For any $\lambda \in \mathbb{C}$, both $\exp[\lambda\bar{t}]M^e_{\kappa,\lambda}(\theta_{\bar{t}})$ and $\exp[\lambda\bar{t}]M^o_{\kappa,\lambda}(\theta_{\bar{t}})$ are local martingales parameterized by t.*

Proof. Given these formulas, in principle it is straightforward to verify that the dt term of the Itô expansion of

$$d\left[e^{\lambda t} M^{e|o}_{\kappa,\lambda}(\theta_t)\right]$$

is equal to zero (where $e|o$ is either e or o). This term can be expressed as

$$e^{\lambda t}\left[\lambda\, M^{e|o}_{\kappa,\lambda}(\theta_t) + \left(M^{e|o}_{\kappa,\lambda}\right)'(\theta_t)\,\frac{\kappa-4}{2}\,\cot(\theta_t/2) + \frac{\kappa}{2}\left(M^{e|o}_{\kappa,\lambda}\right)''(\theta_t)\right]dt.$$

Since Mathematica does not simplify this to zero, we show how to do this for $M^e_{\kappa,\lambda}$; the case for $M^o_{\kappa,\lambda}$ is similar.

We abbreviate $M^e_{\kappa,\lambda}$ with M, let F denote the hypergeometric function in the definition of $M^e_{\kappa,\lambda}$, let a, b, and c denote the parameters of F in M, and change variables to $y = y(\theta) = \sin^2(\theta/4) = (1-\cos(\theta/2))/2$:

$$\begin{aligned}
M(\theta) &= F(y(\theta)),\\
M'(\theta) &= \frac{1}{4}F'(y(\theta))\sin(\theta/2),\\
M''(\theta) &= F''(y(\theta))\frac{1}{16}\sin^2(\theta/2) + F'(y(\theta))\frac{1}{8}\cos(\theta/2)\\
&= F''(y)\frac{y-y^2}{4} + F'(y)\frac{\frac{1}{2}-y}{4},
\end{aligned}$$

so that

$$\begin{aligned}
&e^{\lambda t}\left[\lambda F(y) - \frac{\kappa-4}{4}F'(y)(y-\tfrac{1}{2}) + \frac{\kappa}{8}F''(y)\left(y-y^2\right) + \frac{\kappa}{8}F'(y)\left(\tfrac{1}{2}-y\right)\right]\\
&= e^{\lambda t}\left[\lambda F(y) + (1-3\kappa/8)\,F'(y)\left(y-\tfrac{1}{2}\right) + \frac{\kappa}{8}F''(y)(y-y^2)\right]\\
&= e^{\lambda t}\sum_{n=0}^{\infty} y^n \frac{(a)_n(b)_n}{(c)_n n!}\underbrace{\left[\begin{array}{c}\lambda + (1-3\kappa/8)\left(n - \frac{1}{2}\frac{(a+n)(b+n)}{c+n}\right)\\ +\frac{\kappa}{8}\left(\frac{(a+n)(b+n)}{c+n}n - n(n-1)\right)\end{array}\right]}_{E_n/(c+n)}
\end{aligned}$$

and we define E_n to be $c+n$ times the expression in brackets. (Note that indeed $c \notin -\mathbb{N}$.) We may write E_n as a polynomial in n:

$$\begin{aligned} E_n = &[(1-3\kappa/8)(1-1/2)+(\kappa/8)(1-c+a+b)] \times n^2 \\ &+[\lambda+(1-3\kappa/8)(c-a/2-b/2)+(\kappa/8)(c+a\,b)] \times n \\ &+[c\lambda-(1-3\kappa/8)a\,b/2]. \end{aligned}$$

By our choices of a, b and c, $E_n = 0$ for each n, which proves the claim for $M^e_{\kappa,\lambda}$. □

2.3. Expected first hitting of $2\pi\mathbb{Z}$. In this subsection we obtain asymptotics for the function

$$L(\theta) := \mathbb{E}[\theta_T | \theta_0 = \theta],$$

where θ_t is the diffusion (6) and T is the first time $t \geq 0$ at which $\theta_t \in 2\pi\mathbb{Z}$. (Recall that T is finite a.s. when $\kappa < 8$.) Whenever $\theta \in 2\pi\mathbb{Z}$, trivially $L(\theta) = \theta$.

Lemma 2. *For the diffusion (6) with $8/3 < \kappa < 8$, $L(\theta_t)$ is a martingale.*

Proof. Since $L(\theta)$ is defined in terms of expected values, $L(\theta_t)$ is a local martingale whenever $\theta_t \notin 2\pi\mathbb{Z}$, and the stopped process $L(\theta_{\min(t,T)})$ is a martingale. Since the diffusion behaves symmetrically around the points $2\pi\mathbb{Z}$ and the number of intervals of $\mathbb{R} \setminus 2\pi\mathbb{Z}$ crossed before some deterministic time has exponentially decaying tails (which implies integrability), $L(\theta_t)$ is a martingale. □

Next, we express $L(\theta_t)$ in terms of the $\lambda = 0$ case of the local martingales $\exp[\lambda t] M^e_{\kappa,\lambda}(\theta_t)$ and $\exp[\lambda t] M^o_{\kappa,\lambda}(\theta_t)$. Because $M^e_{\kappa,0}(\theta) = 1$, this local martingale is uninformative, but

$$M^o_{\kappa,0}(\theta) = F\left(\tfrac{3}{2} - \tfrac{4}{\kappa}, \tfrac{1}{2}; \tfrac{3}{2}; \cos^2\tfrac{\theta}{2}\right) \cos\tfrac{\theta}{2}.$$

We have $M^o_{\kappa,0}(0) = -M^o_{\kappa,0}(2\pi) = \sqrt{\pi}\Gamma(4/\kappa - 1/2)/(2\Gamma(4/\kappa))$, and since $M^o_{\kappa,0}$ is bounded (when $\kappa \neq 8$), the stopped process $M^o_{\kappa,0}(\theta_{\min(t,T)})$ is also a martingale. This determines L, namely,

$$L(\theta) = \pi - \frac{2\sqrt{\pi}\,\Gamma\left(\frac{4}{\kappa}\right)}{\Gamma\left(\frac{4}{\kappa} - \frac{1}{2}\right)} F\left(\tfrac{3}{2} - \tfrac{4}{\kappa}, \tfrac{1}{2}; \tfrac{3}{2}; \cos^2\tfrac{\theta}{2}\right) \cos\tfrac{\theta}{2}, \qquad \theta \in [0, 2\pi] \tag{8}$$

and $L(\theta + 2\pi) = L(\theta) + 2\pi$. (It is also possible to derive (8) from the formula for Pr[SLE trace passes to left of $x + iy$] [Sch01] after applying a Möbius transformation and suitable hypergeometric identities.)

We wish to understand the behavior of L near the points in $2\pi\mathbb{Z}$, and to this end we use the formula (see [EMOT53, p. 108, Eq. 2.10.1])

$$\begin{aligned} F(a,b;c;z) = &\frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)} F(a,b;a+b-c+1;1-z) \\ &+\frac{\Gamma(c)\Gamma(a+b-c)}{\Gamma(a)\Gamma(b)} (1-z)^{c-a-b} F(c-a,c-b;c-a-b+1;1-z) \end{aligned} \tag{9}$$

which is valid when $1-c$, $b-a$, and $c-b-a$ are not integers and $|\arg(1-z)| < \pi$. In our case the nonintegrality condition is satisfied whenever $8/\kappa \notin \mathbb{N}$. For the range of κ that we are interested in, this rules out $\kappa = 4$, for which we already know $L(\theta) = \theta$, and therefore do not need asymptotics. The endpoints of the range, $\kappa = 8/3$ and $\kappa = 8$ are also ruled out, but for the remaining κ's we have

$$\begin{aligned} F\left(\tfrac{3}{2} - \tfrac{4}{\kappa}, \tfrac{1}{2}; \tfrac{3}{2}; \cos^2\tfrac{\theta}{2}\right) &= \frac{\Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{4}{\kappa} - \frac{1}{2}\right)}{\Gamma\left(\frac{4}{\kappa}\right)\Gamma(1)} F\left(\tfrac{3}{2} - \tfrac{4}{\kappa}, \tfrac{1}{2}; \tfrac{3}{2} - \tfrac{4}{\kappa}; \sin^2\tfrac{\theta}{2}\right) \\ &\quad + \frac{\Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{1}{2} - \frac{4}{\kappa}\right)}{\Gamma\left(\frac{3}{2} - \frac{4}{\kappa}\right)\Gamma\left(\frac{1}{2}\right)} \left|\sin\tfrac{\theta}{2}\right|^{\frac{8}{\kappa}-1} F\left(\tfrac{4}{\kappa}, 1; \tfrac{4}{\kappa} + \tfrac{1}{2}; \sin^2\tfrac{\theta}{2}\right) \\ &= \frac{\sqrt{\pi}\Gamma\left(\frac{4}{\kappa} - \frac{1}{2}\right)}{2\Gamma\left(\frac{4}{\kappa}\right)} \frac{1}{\left|\cos\frac{\theta}{2}\right|} + \frac{\Gamma\left(\frac{1}{2} - \frac{4}{\kappa}\right)}{2\Gamma\left(\frac{3}{2} - \frac{4}{\kappa}\right)} \left|\sin\tfrac{\theta}{2}\right|^{\frac{8}{\kappa}-1} \\ &\quad \times F\left(\frac{4}{\kappa}, 1; \frac{4}{\kappa} + \frac{1}{2}; \sin^2\frac{\theta}{2}\right), \end{aligned}$$

and by (8),

$$L(\theta) = c_\kappa \left(\sin\tfrac{\theta}{2}\right)^{\frac{8}{\kappa}-1} F\left(\tfrac{4}{\kappa}, 1; \tfrac{4}{\kappa} + \tfrac{1}{2}; \sin^2\tfrac{\theta}{2}\right) \cos\tfrac{\theta}{2}, \qquad \theta \in (0, \pi).$$

for some constant $c_\kappa > 0$. (One can use the Legendre duplication formula to show that $c_\kappa = 2^{8/\kappa - 1}\Gamma(\frac{4}{\kappa})^2/\Gamma(\frac{8}{\kappa})$, but we do not need this.) Since $L(-\theta) = -L(\theta)$, we conclude that

$$L(\theta) = A_0(\theta^2)\,|\theta|^{8/\kappa}/\theta, \qquad \theta \in (-\pi, \pi),$$

where A_0 is a analytic function (depending on κ) satisfying $A_0(0) > 0$. This implies

$$\theta^2 = A\left(|L(\theta)|^{2\kappa/(8-\kappa)}\right) \tag{10}$$

near $\theta = 0$, for some analytic A.

2.4. Starting at $\theta_0 = 0$. We will eventually need to start the diffusion at $\theta_0 = 0$, but Lemma 1 only covers what happens up until the first time that $\theta_t \in 2\pi\mathbb{Z}$. In this subsection we show

Lemma 3. *For the diffusion (6) with $8/3 < \kappa < 8$, let T be the first time at which $\theta_t \in 2\pi\mathbb{Z} \setminus 4\pi\mathbb{Z}$, and let $\bar{t} = \min(t, T)$. For any $\lambda \in \mathbb{C}$, the process $\exp[\lambda\bar{t}]M^e_{\kappa,\lambda}(\theta_{\bar{t}})$ is a martingale.*

Proof. Let M abbreviate $M^e_{\kappa,\lambda}$, and let us assume without loss of generality that $-2\pi < \theta_0 < 2\pi$. Let us define $\omega_t = L(\theta_t)$, which is a martingale and may be interpreted as a time-changed Brownian motion. We wish to argue that $e^{\lambda\bar{t}}M(L^{-1}(\omega_{\bar{t}})) = e^{\lambda\bar{t}}M(\theta_{\bar{t}})$ is a local martingale. (The definition of L implies that it is strictly monotone, and hence L^{-1} is well defined.) Note that by Lemma 1 it is a local martingale when $\theta_{\bar{t}} \notin 2\pi\mathbb{Z}$. To extend this to a neighborhood of $\theta_{\bar{t}} = 0$, one could try to use Itô's formula. To do this, it would be necessary that $f := M \circ L^{-1}$ be twice differentiable. We have $M(\theta) = A_1(\theta^2)$ in $(-2\pi, 2\pi)$, where A_1 is analytic. Consequently, (10) gives for $8/3 < \kappa < 8$ and for ω in a neighborhood of 0,

$$f(\omega) = A_2\left(|\omega|^{2\kappa/(8-\kappa)}\right), \tag{11}$$

for some analytic A_2. Though this is not necessary for the proof, one can check that $A_2'(0) \neq 0$ and therefore $f''(0)$ is not finite when $\kappa < 4$.

To circumvent the problem of $f = M \circ L^{-1}$ not being twice differentiable, we use the Itô-Tanaka Theorem ([RY99, Theorem 1.5 on p. 223]). The exponent $\frac{2\kappa}{8-\kappa}$ in (11) ranges from 1 to ∞ as κ ranges from 8/3 to 8. In particular, $f'(0) = 0$ and f' is continuous near 0. Since A_2 is analytic, near 0 the function f may be expressed as the difference of two convex functions, namely, $f(\omega) = (f(\omega) - f(0))\ 1_{\omega \geq 0} + (f(\omega) - f(0))\ 1_{\omega \leq 0} + f(0)$. Therefore, we may apply the Itô-Tanaka Theorem to conclude that $e^{\lambda \tilde{t}} f(\omega_{\tilde{t}}) = e^{\lambda \tilde{t}} M(\theta_{\tilde{t}})$ is a local martingale also when $\theta_{\tilde{t}}$ is near zero.

Now, the hypergeometric function F satisfies

$$F(a, b; c; 1) = \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)} \tag{12}$$

provided $-c \notin \mathbb{N}$ and $\mathrm{Re}\, c > \mathrm{Re}(a+b)$ (see e.g. [EMOT53, p. 104, Eq. 46]). Therefore, $M(\pm 2\pi)$ is finite. Thus, $e^{\lambda \tilde{t}} M(\theta_{\tilde{t}})$ is bounded for bounded t, and we may conclude that it is a martingale. □

For future reference, we now calculate $M^e_{\kappa,\lambda}(\pm 2\pi)$. Observe that the parameters a, b, c of the hypergeometric function in the definition of $M^e_{\kappa,\lambda}$ satisfy $2c - a - b = 1$. Consequently, the identity $\Gamma(z)\Gamma(1-z) = \pi/\sin(\pi z)$ and (12) give

$$M^e_{\kappa,\lambda}(\pm 2\pi) = \frac{\sin\left(\frac{\pi}{2} - \pi\sqrt{\left(1-\frac{4}{\kappa}\right)^2 + \frac{8\lambda}{\kappa}}\right)}{\sin\left(3\pi/2 - 4\pi/\kappa\right)} = \frac{\cos\left(\pi\sqrt{(1-4/\kappa)^2 + 8\lambda/\kappa}\right)}{\cos(\pi(1-4/\kappa))}. \tag{13}$$

3. Proofs of Main Results

We now restate and prove Theorem 1.

Theorem 3. *Suppose the diffusion process (6) (with $8/3 < \kappa < 8$) is started at $\theta_0 = 0$, and T is the first time at which $\theta_t = \pm 2\pi$. If $\mathrm{Re}\,\lambda \leq 0$, then*

$$\mathbb{E}\left[e^{\lambda T} \mid \theta_0 = 0\right] = \frac{\cos(\pi(1-4/\kappa))}{\cos\left(\pi\sqrt{(1-4/\kappa)^2 + 8\lambda/\kappa}\right)}.$$

(This is equivalent to Theorem 1 by Proposition 2 and the remarks following the statement of Theorem 1.)

Proof. Since $M^e_{\kappa,\lambda}(\theta_T) = M^e_{\kappa,\lambda}(\pm 2\pi) = M^e_{\kappa,\lambda}(2\pi)$ a.s. and $\exp[\lambda \tilde{t}] M^e_{\kappa,\lambda}(\theta_{\tilde{t}})$ is a martingale, the optional sampling theorem gives

$$M^e_{\kappa,\lambda}(2\pi)\, \mathbb{E}\left[e^{\lambda T} \mid \theta_0 = 0\right] = \mathbb{E}\left[e^{\lambda T} M^e_{\kappa,\lambda}(\theta_T) \mid \theta_0 = 0\right] = M^e_{\kappa,\lambda}(0) = 1,$$

and the proof is completed by appeal to (13). □

Proof of Theorem 2. Fix some $\varepsilon > 0$ and let $z \in \mathbb{D}$. Set $r_0 := 1 - |z|$, $r_1 := \operatorname{dist}(z, L_1^z)$, and suppose that $\varepsilon < r_0$. We seek to estimate the probability that the open disk of radius ε about z intersects the gasket; that is, the probability that $r_1 < \varepsilon$. By the Koebe 1/4 theorem, $r_1 \le \mathrm{CR}(D_1, z) \le 4 r_1$. Likewise, $r_0 \le \mathrm{CR}(\mathbb{D}, z) \le 4 r_0$. Thus, $B_1^z = \log \mathrm{CR}(\mathbb{D}, z) - \log \mathrm{CR}(D_1, z) = \log(r_0/r_1) + O(1)$. Referring to the density function of B_k^z (4), we see that

$$\Pr[r_1 < \varepsilon] \asymp \exp[-\alpha \log(r_0/\varepsilon)] = (\varepsilon/r_0)^\alpha,$$

where

$$\alpha = \frac{1/4 - (1 - 4/\kappa)^2}{8/\kappa} = -\frac{3\kappa}{32} + 1 - \frac{2}{\kappa} = \frac{(8-\kappa)(3\kappa - 8)}{32\kappa}.$$

For each $j = 1, \ldots, \lceil 1/\varepsilon \rceil$, we may cover the annulus $\{z : (j-1)\,\varepsilon \le 1 - |z| \le j\,\varepsilon\}$ by $O(1/\varepsilon)$ disks of radius ε. The total expected number of these disks that intersect the gasket is at most

$$\sum_{j=1}^{\lceil 1/\varepsilon \rceil} O(1/\varepsilon) \times O(\varepsilon/(j\varepsilon))^\alpha = O(\varepsilon^{\alpha - 2}).$$

(Here we made use of the fact that $\alpha < 1$.) Thus on average $O(\varepsilon^{\alpha-2})$ disks of radius ε suffice to cover the gasket.

On the other hand, we may pack into $\mathbb{D}$ at least $\Theta(1/\varepsilon^2)$ points so that every two of them are more than distance 4ε apart, and each of them is at least distance $1/2$ from the boundary. For each such point z there is a $\Theta(\varepsilon^\alpha)$ chance that the disk or radius ε centered at z is not surrounded by a loop, i.e., that that the gasket Γ contains a point z' that is within distance ε of z. Since the points z are sufficiently far apart, the points z' must be covered by distinct disks in any covering of Γ by disks of radius ε. Thus the expected number of disks of radius ε required to cover the gasket Γ is at least $\Theta(\varepsilon^{\alpha-2})$. □

4. Open Problems

Kenyon and Wilson [KW04] also predicted the large-k limiting distribution of another quantity, the "electrical thickness" of the loops L_k^z when $k \to \infty$. The electrical thickness of a loop compares the conformal radius of the loop to the conformal radius of the image of the loop under the map $m(w) = 1/(w - z)$, and more precisely it is

$$\vartheta_z(L_k^z) = -\log \mathrm{CR}(L_k^z, z) - \log \mathrm{CR}(m(L_k^z), z).$$

Kenyon and Wilson [KW04] predicted that the large-k moment generating function of $\vartheta_z(L_k^z)$ is

$$\lim_{k\to\infty} \mathbb{E}[\exp(\lambda \vartheta_z(L_k^z))] = \frac{\sin(\pi(1-4/\kappa))}{\pi(1-4/\kappa)} \frac{\pi\sqrt{(1-4/\kappa)^2 + 8\lambda/\kappa}}{\sin\left(\pi\sqrt{(1-4/\kappa)^2 + 8\lambda/\kappa}\right)}, \quad (14)$$

or equivalently that the limiting probability density function is given by the density function of the exit time of a standard Brownian *excursion* started in the middle of the interval $(-2\pi/\sqrt{\kappa}, 2\pi/\sqrt{\kappa})$, reweighted by a factor of const $\times \exp[(\kappa-4)^2 x/(8\kappa)]$. (This equivalence follows from [BS02, Eq. 5.3.0.1].) Recall that the density function

of B_k^z is given by the density function of the exit time of a standard Brownian *motion* started in the middle of the interval $(-2\pi/\sqrt{\kappa}, 2\pi/\sqrt{\kappa})$, also reweighted by a factor of const $\times \exp[(\kappa-4)^2 x/(8\kappa)]$. These forms are highly suggestive, but currently we do not know how to calculate the electrical thickness using CLE_κ, nor do we have a conceptual explanation for why these distributions take these forms.

References

[BS02] Borodin, A.N., Salminen, P.: *Handbook of Brownian Motion—Facts and Formulae*. Probability and its Applications. Basel: Birkhäuser Verlag, 2nd edition, 2002

[Car07] Cardy, J.: ADE and SLE. J. Phys. A **40**(7), 1427–1438 (2007)

[CN06] Camia, F., Newman, C.M.: Two-dimensional critical percolation: the full scaling limit. Commun. Math. Phys. **268**(1), 1–38 (2006)

[CN07] Camia, F., Newman, C.M.: Critical percolation exploration path and SLE_6: a proof of convergence. Probab. Theory Related Fields **139**(3-4), 473–519 (2007)

[CT62] Ciesielski, Z., Taylor, S.J.: First passage times and sojourn times for Brownian motion in space and the exact Hausdorff measure of the sample path. Trans. Amer. Math. Soc. **103**(3), 434–450 (1962)

[CZ03] Cardy, J., Ziff, R.M.: Exact results for the universal area distribution of clusters in percolation, Ising, and Potts models. J. Stat. Phys. **110**(1-2), 1–33 (2003)

[Dub05] Dubédat, J.: 2005, Personal communication

[Dup90] Duplantier, B.: Exact fractal area of two-dimensional vesicles. Phys. Rev. Lett. **64**(4), 493 (1990)

[EMOT53] Erdélyi, A., Magnus, W., Oberhettinger, F., Tricomi, F.G.: *Higher Transcendental Functions*. Vol. I., New York: McGraw-Hill Book Company, 1953, based, in part, on notes left by Harry Bateman

[FK72] Fortuin, C.M., Kasteleyn, P.W.: On the random-cluster model. I. Introduction and relation to other models. Physica **57**, 536–564 (1972)

[Gri06] Grimmett, G.: *The Random-Cluster Model*, Volume **333** of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Berlin: Springer-Verlag, (2006)

[KN04] Kager, W., Nienhuis, B.: A guide to stochastic Löwner evolution and its applications. J. Stat. Phys. **115**(5-6), 1149–1229 (2004)

[KW04] Kenyon, R.W., Wilson, D.B.: *Conformal radii of loop models*, 2004. Manuscript

[LSW02] Lawler, G.F., Schramm, O., Werner, W.: *One-arm exponent for critical 2D percolation*. Electron. J. Probab. **7**: Paper No. 2, 13 pp. (2002)

[RY99] Revuz, D., Yor, M.: *Continuous Martingales and Brownian Motion*, Volume **293** of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Berlin: Springer-Verlag, third edition, 1999

[Sch01] Schramm, O.: *A percolation formula*. Electron. Comm. Probab. **6**, 115–120 (2001)

[She06] Sheffield, S.: *Exploration trees and conformal loop ensembles*. http://arxiv.org/abs/math.PR/0609167, 2006 Duke Math. J., to appear

[Smi01] Smirnov, S.: Critical percolation in the plane: conformal invariance, Cardy's formula, scaling limits. C. R. Acad. Sci. Paris Sér. I Math. **333**(3), 239–244 (2001)

[SW08a] Sheffield, S., Werner, W.: *Conformal loop ensembles: Construction via loop-soups*, 2008, in preparation

[SW08b] Sheffield, S., Werner, W.: *Conformal loop ensembles: The Markovian characterization*, 2008, in preparation

Communicated by M. Aizenman

Conformally invariant scaling limits: an overview and a collection of problems

Oded Schramm

Abstract. Many mathematical models of statistical physics in two dimensions are either known or conjectured to exhibit conformal invariance. Over the years, physicists proposed predictions of various exponents describing the behavior of these models. Only recently have some of these predictions become accessible to mathematical proof. One of the new developments is the discovery of a one-parameter family of random curves called stochastic Loewner evolution or SLE. The SLE curves appear as limits of interfaces or paths occurring in a variety of statistical physics models as the mesh of the grid on which the model is defined tends to zero.

The main purpose of this article is to list a collection of open problems. Some of the open problems indicate aspects of the physics knowledge that have not yet been understood mathematically. Other problems are questions about the nature of the SLE curves themselves. Before we present the open problems, the definition of SLE will be motivated and explained, and a brief sketch of recent results will be presented.

Mathematics Subject Classification (2000). Primary 60K35; Secondary 82B20, 82B43, 30C35.

Keywords. Statistical physics, conformal invariance, stochastic Loewner evolutions, percolation.

1. Introduction

In the past several years, many predictions from physics regarding the large-scale behavior of random systems defined on a lattice in two dimensions have become accessible to mathematical study and proof. Of central importance is the asymptotic conformal invariance of these systems. It turns out that paths associated with these random configurations often fall into a one-parameter family of conformally invariant random curves called stochastic Loewner evolutions, or SLE. We start by motivating SLE through a simple mathematical model of percolation. After giving the definition of SLE, we present a narrative of recent developments. However, since there are good surveys on the subject in the literature [94], [36], [21], [52], [95], this introductory part of the paper will be short and cursory. The rest of the paper will consist of an annotated list of open problems in the subject.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006
© 2007 European Mathematical Society

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_34, © Springer Science+Business Media, LLC 2011

1.1. Motivation and definition of SLE. To motivate SLE, we now discuss percolation. More specifically, we define one particular model of percolation in two dimensions. Fix a number $p \in [0, 1]$. Let ω be a random subset of the set of vertices in the triangular grid TG, where for vertices $v \in V(\mathrm{TG})$ the events $v \in \omega$ are independent and have probability p. In percolation theory one studies the connected components (a.k.a. clusters) of the random subgraph of TG whose vertex set is ω and whose edges are the edges in TG connecting two elements of ω. Equivalently, one may study the connected components of the set of white hexagons in Figure 1, where each hexagon in the hexagonal grid dual to TG represents a vertex of TG and hexagons corresponding to vertices in ω are colored white. The reasons for considering this dual representation are that the figures come out nicer and that it makes some important definitions more concise.

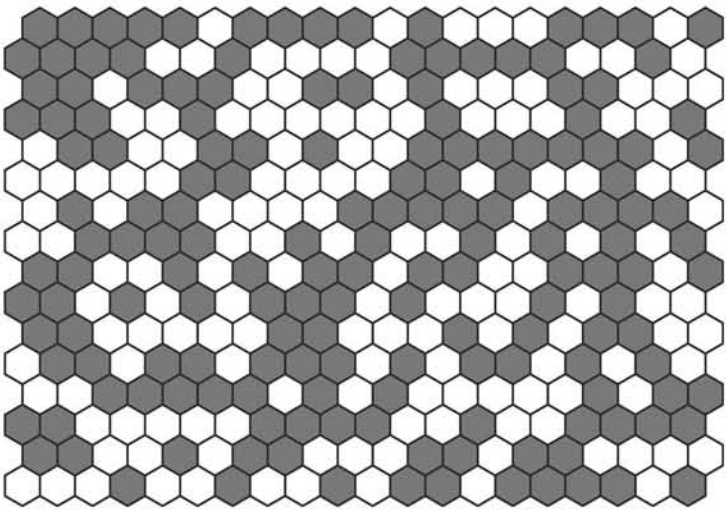


Figure 1. Site percolation on the triangular grid as represented by colored hexagons.

The above percolation model is site (or vertex) percolation on the triangular grid. Likewise, there is a bond (or edge) model, where one considers a random subgraph of a grid whose vertex set is the set of all vertices of the grid, but where each edge of the grid is in the percolation subgraph with probability p, independently. Additionally, there are various percolation models which are not based on a lattice. Some of these will be discussed in later sections.

There is an important value p_c of the parameter p, which is the threshold for the existence of an unbounded cluster and is called the critical value of p. The actual value of p_c varies depending on the particular percolation model. For site percolation on the triangular grid, as well as for bond percolation on the square grid, we have $p_c = 1/2$. This is a theorem of Kesten [43], based on earlier work by Harris [33], Russo [78] and Seymour and Welsh [84]. The underlying reason for this nice value of p_c is a duality which these two models have, though the precise form of the duality they exhibit is

different. For bond percolation on the triangular grid $p_c = 2\sin(\pi/18)$ [98], while for site percolation on the square grid there is not even a prediction for the value of p_c, though rigorous and experimental estimates exist. As p increases beyond p_c, the large scale behavior of percolation undergoes a rapid change. This is perhaps the mathematically simplest model of a phase transition. From now on, we will focus our attention on critical percolation, that is, percolation with $p = p_c$, which is in many ways the most interesting value of p.

We now define and discuss the percolation interface curve indicated in Figure 2. Consider a bounded domain D in the plane $\mathbb{R}^2 = \mathbb{C}$ whose boundary is a simple closed curve. Let $\partial_+ \subset \partial D$ be a proper arc on the boundary of D. Given $\varepsilon > 0$ we may consider the collection of hexagons in a hexagonal grid of mesh ε which intersect $\overline{D}$. Each of these hexagons which meets ∂_+ we color white, each of the hexagons which meet ∂D but not ∂_+ we color black, and each of the hexagons contained in D we color white or black with probability $1/2$, independently. In addition to white clusters (connected components of white hexagons) sometimes, black clusters are also considered. Percolation theory is the study of connected components of random sets, such as these clusters.

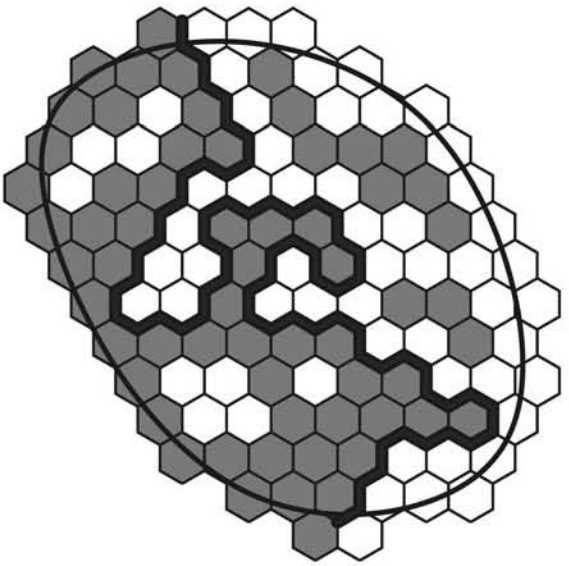


Figure 2. The interface associated with percolation.

For simplicity, we assume that ∂D is sufficiently smooth and ε is sufficiently small so that the union of hexagons intersecting ∂D but not ∂_+ is connected. There is a unique (random) path β, which is the common boundary of the white cluster meeting ∂_+ and the black cluster meeting ∂D.

The law $\mu_{D,\partial_+,\varepsilon}$ of β is a probability measure on the space of closed subsets of $\overline{D}$ with the Hausdorff metric. Smirnov [87] proved that as $\varepsilon \searrow 0$ the measure

$\mu_{D,\partial_+,\varepsilon}$ converges weakly to a measure μ_{D,∂_+}, and that μ is conformally invariant, in the following sense. If $f\colon \overline{D} \to \overline{D'} \subset \mathbb{R}^2$ is a homeomorphism that is analytic in D, then the push forward of μ_{D,∂_+} under f is $\mu_{f(D),f(\partial_+)}$. In other words, if ε is small, then $f(\beta)$ is a good approximation for the corresponding path defined using a hexagonal grid of mesh ε in $f(D) = D'$. This type of conformal invariance was believed to hold for many "critical" random systems in two dimensions. However, the only previous result establishing conformal invariance for a random scaling limit is Lévy's theorem [64] stating that for two-dimensional Brownian motion, the scaling limit of simple random walk on $\mathbb{Z}^2$, is conformally invariant up to a time-change.

Smirnov's proof is very beautiful, and the result is important, but describing the proof will throw us too far off course (since for this paper percolation is just an example model, not the primary topic). The interested reader is encouraged to consult [12], [30], [70], [45] for background in percolation and highlights of percolation theory. An elegant simplification of parts of Smirnov's proof has been discovered by Vincent Beffara [11]. A more detailed version of other parts of Smirnov's argument appears in [19].

Though this was not the original inspiration, we will now use Smirnov's result to motivate the definition of SLE. By conformal invariance, we may venture to understand β in the domain of our choice. The simplest situation turns out to be when $D = \mathbb{H}$ is the upper half plane and ∂_+ is the positive real ray $\mathbb{R}_+$, as in Figure 3.

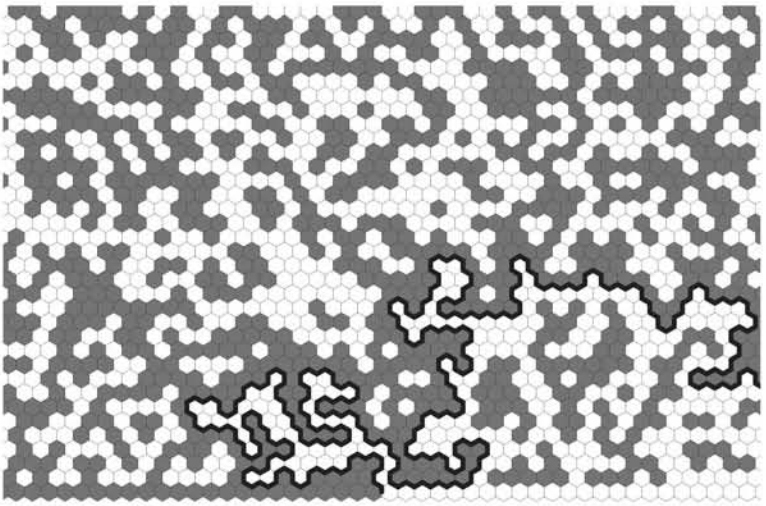


Figure 3. The percolation interface in the upper half plane.

(Though this domain D is unbounded, that does not cause any problems.) We may consider the discrete path β as a simple path $\beta\colon [0, T) \to \overline{\mathbb{H}}$ starting near 0 and satisfying $\lim_{t\to T} |\beta(t)| = \infty$ (where T is finite or infinite).

We would like to learn about β by understanding the one-parameter family of conformal maps g_t mapping $\mathbb{H} \setminus \beta([0, t])$ onto $\mathbb{H}$. To facilitate this, we must first

recall a few basic facts and discuss Loewner's theorem. At this point, assume only that β is a simple path in $\overline{\mathbb{H}}$ with $\beta(0) \in \mathbb{R}$ and $\beta(t) \notin \mathbb{R}$ for $t > 0$. The existence of the conformal maps $g_t : \mathbb{H} \setminus \beta([0,t]) \to \mathbb{H}$ is guaranteed by Riemann's mapping theorem. However, g_t is not unique. In order to choose a specific g_t for every t, we first require that $g_t(\infty) = \infty$. Schwarz reflection in the real axis implies that g_t is analytic in a neighborhood of ∞, and therefore admits a power series representation in $1/z$,

$$g_t(z) = a_1 z + a_0 + a_{-1} z^{-1} + a_{-2} z^{-2} + \cdots ,$$

valid for all z sufficiently large. Since g_t maps the real line near ∞ into the real line, it follows that $a_j \in \mathbb{R}$ for all j, and because $g_t : \mathbb{H} \setminus \beta([0,t]) \to \mathbb{H}$, we find that $a_1 > 0$. We now pick a specific g_t by imposing the so-called hydrodynamic normalization at ∞, namely $a_1 = 1$ and $a_0 = 0$. This can clearly be achieved by post-composing with a map of the form $z \mapsto a z + b$, $a > 0$, $b \in \mathbb{R}$.

The coefficients a_j of the series expansion of g_t are now functions of t. It is not hard to verify that $a_{-1}(t)$ is a continuous, strictly increasing function of t. Clearly, $g_0(z) = z$ and hence $a_{-1}(0) = 0$. We may therefore reparametrize β so as to have $a_{-1}(t) = 2t$ for all $t > 0$. This is called the half-plane capacity parametrization of β. With this parametrization, a variant of Loewner's theorem [65] states that the maps g_t satisfy the differential equation

$$\frac{dg_t(z)}{dt} = \frac{2}{g_t(z) - W(t)}, \tag{1}$$

where $W(t) := g_t\big(\beta(t)\big)$ is called the Loewner driving term. A few comments are in order.

1. Although g_t is defined in $\mathbb{H} \setminus \beta([0,t])$, it does extend continuously to $\beta(t)$, and therefore $W(t)$ is well defined.

2. If $z = \beta(s)$ for some s, then (1) makes sense only as long as $t < s$. That is to be expected. The domain of definition of g_t is shrinking as t increases. A point z falls out of the domain of g_t at the first time $\tau = \tau_z$ such that $\liminf_{t \nearrow \tau} g_t(z) - W(t) = 0$.

3. The main point here is that information about the path β is encoded in $W(t)$, which is a path in $\mathbb{R}$.

4. The proof of (1) is not too hard. In [53, Theorem 2.6] a proof (of a generalization) may be found.

We now return to the situation where β is the percolation interface chosen according to $\mu_{\mathbb{H},\mathbb{R}_+,\varepsilon}$, parametrized by half-plane capacity. It is easy to see that in this case $T = \infty$. Fix some $s > 0$. Suppose that we examine the colors of only those hexagons that are necessary to determine $\beta([0,s])$. This can be done by sequentially testing the hexagons adjacent to β starting from $\beta(0)$ as follows. Each time the already

determined arc of β meets a hexagon whose color has not yet been examined, we test the color (which permits us to extend the determined initial arc of β by at least one segment), until $\beta([0, s])$ has been determined. See Figure 4.

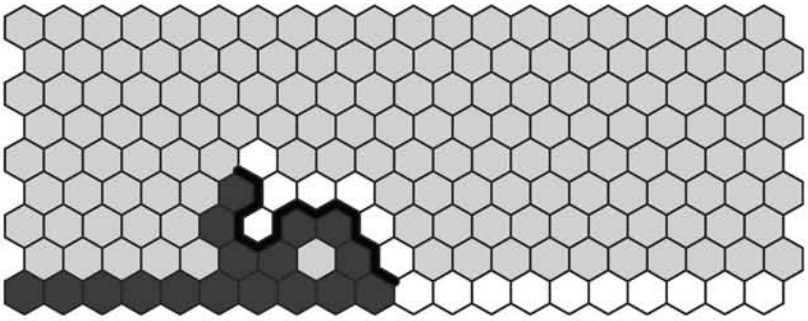


Figure 4. Initial segment of interface.

Now comes the main point. Let D_s be the unbounded component of the collection of hexagons of undetermined color in $\mathbb{H}$, and let ∂_+^s be the subset of ∂D_s lying on the boundary of hexagons of determined color white. Then the distribution of the continuation $\beta([s, \infty))$ of the interface given $\beta([0, s])$ is $\mu_{D_s, \partial_+^s, \varepsilon}$. By Smirnov's theorem, if $G\colon D_s \to \mathbb{H}$ is the conformal map satisfying the hydrodynamic normalization, then the image under G of $\mu_{D_s, \partial_+^s, \varepsilon}$ is close to $\mu_{\mathbb{H}, G(\partial_+^s), \varepsilon}$. (Actually, to justify this, one needs a slightly stronger "uniform" version of Smirnov's theorem. But here we want to convey the main ideas, and do not bother about being entirely precise.) Now, since D_s approximates $\mathbb{H} \setminus \beta([0, s])$, it follows that G is very close to g_s and $G(\partial_+^s)$ is close to $[W(s), \infty) = [g_s(\beta(s)), \infty)$. Therefore, in the limit as $\varepsilon \searrow 0$, we have for β sampled according to $\mu_{\mathbb{H}, \mathbb{R}_+}$ that given $\beta([0, s])$ the distribution of $g_s \circ \beta([s, \infty))$ (which is the conformal image of the continuation of the path) is $\mu_{\mathbb{H}, \mathbb{R}_+}$ translated by $W(s)$.

The Loewner driving term of the path $t \mapsto g_s \circ \beta(s + t)$ is $W(s + t)$, because $g_{s+t} \circ g_s^{-1}$ maps $\mathbb{H} \setminus g_s\big(\beta([s, s + t])\big)$ onto $\mathbb{H}$. The conclusion of the previous paragraph therefore implies that given $(W(t) : t \in [0, s])$ the distribution of the continuation of W is identical to the original distribution of W translated to start at $W(s)$. This is a very strong property. Indeed, for every $n \in \mathbb{N}$ and $t > 0$ we may write $W(t) = \sum_{j=1}^n \big(W(j\,t/n) - W((j-1)t/n)\big)$, which by the above is a sum of n independent identically distributed random variables. If we assume that the variance of $W(t)$ is finite, then it is also the sum of the variances of the summands. By the central limit theorem, $W(t)$ is therefore a Gaussian random variable. By symmetry, $W(t)$ has the same distribution as $-W(t)$, and so $W(t)$ is a centered Gaussian. It now easily follows that there is some constant $\kappa \geq 0$ such that $W(t)$ has the same distribution as $B(\kappa\, t)$, where B is one-dimensional Brownian motion starting at $B(0) = 0$. Using results from the theory of stochastic processes (e.g., the characterization of continuous martingales as time-changed Brownian motion), the same conclusion can be reached

while replacing the assumption that $W(t)$ has finite variance with the continuity of $W(t)$ in t.

We have just seen that Smirnov's theorem implies that the Loewner driving term of a sample from $\mu_{\mathbb{H},\mathbb{R}_+}$ is $B(\kappa t)$ for some $\kappa \geq 0$. This should serve as adequate motivation for the following definition from [79].

Definition 1.1. Fix some $\kappa \geq 0$, and let g_t be the solution of Loewner's equation (1) satisfying $g_0(z) = z$ with $W(t) = B(\kappa t)$, where B is standard one-dimensional Brownian motion starting at $B(0) = 0$. Then $(g_t : t \geq 0)$ is called *chordal stochastic Loewner evolution* with parameter κ or SLE_κ.

Of course, SLE_κ is a random one-parameter family of maps; the randomness is entirely due to the Brownian motion.

It has been proven [76], [59] that with probability 1 there is a (unique) random continuous path $\gamma(t)$ such that for each $t \geq 0$ the domain of definition of g_t is the unbounded component of $\mathbb{H} \setminus \gamma([0, t])$. The path is given by $\gamma(t) = g_t^{-1}(W(t))$, but proving that $g_t^{-1}(W(t))$ is well defined is not easy. It is also known [76] that a.s. $\gamma(t)$ is a simple path if and only if $\kappa \leq 4$ and is space-filling if and only if $\kappa \geq 8$. Sometimes the path γ itself is called SLE_κ. This is not too inconsistent, because g_t can be reconstructed from $\gamma([0, t])$ and vice versa.

If D is a simply connected domain in the plane and $a, b \in \partial D$ are two distinct points (or rather prime ends), then chordal SLE from a to b in D is defined as the image of γ under a conformal map from $\mathbb{H}$ to D taking 0 to a and ∞ to b. Though the map is not unique, the choice of the map does not effect the law of the SLE in D. This follows from the easily verified fact that up to a rescaling of time, the law of the SLE path is invariant under scaling by a positive real constant, as is the case for Brownian motion.

The reason for calling the SLE "chordal" is that it connects two boundary points of a domain D. There is another version of SLE, which connects a boundary point to an interior point, called *radial* SLE. Actually, there are a few other variations, but they all have similar definitions and analogous properties.

1.2. A historical narrative. In this subsection we list some works and discoveries related to SLE and random scaling limits in two dimensions. The following account is not comprehensive. Some of the topics not covered here are discussed in Wendelin Werner's [97] contribution to this ICM proceedings. We start by very briefly discussing the historical background.

In the survey paper [49], Langlands, Pouliot and Saint-Aubin present a collection of intriguing predictions from statistical physics. They have discussed these predictions and some simulation data with Aizenman, which prompted him to conjecture that the critical percolation crossing probabilities are asymptotically conformally invariant (see [49]). This means that the probability $Q_\varepsilon(D, \partial_1, \partial_2)$ that there exists a critical percolation cluster in a domain $D \subset \mathbb{C}$ connecting two boundary arcs ∂_1 and ∂_2 on a lattice with mesh ε has a limit $Q(D, \partial_1, \partial_2)$ as $\varepsilon \searrow 0$ and that the limit is conformally

invariant, namely, $Q(D, \partial_1, \partial_2) = Q\big(f(D), f(\partial_1), f(\partial_2)\big)$ if f is a homeomorphism from $\overline{D}$ to $f(\overline{D})$ that is conformal in D. This led John Cardy [20] to propose his formula (involving hypergeometric functions) for the asymptotic crossing probability in a rectangle between two opposite edges. The survey [49] highlighted these predictions and the role of the conjectured conformal invariance in critical percolation, as well as several other statistical physics models in two dimensions.

Prior to SLE there were attempts to use compositions of conformal slit mappings and even Loewner's equation in the study of diffusion limited aggregation (DLA). DLA is a random growth process, which produces a random fractal and is notoriously hard to analyse mathematically. (See [100], [4] for a definition and discussion of DLA.) Makarov and Carleson [22] used Loewner's equation to study a much simplified deterministic variant of DLA, which is not fractal, and Hastings and Levitov [34] have used conformal mapping techniques for a non-rigorous study of more realistic versions of DLA. Given that the fractals produced by DLA are not conformally invariant, it is not too surprising that it is hard to faithfully model DLA using conformal maps. Harry Kesten [44] proved that the diameter of the planar DLA cluster after n steps grows asymptotically no faster than $n^{2/3}$, and this appears to be essentially the only theorem concerning two-dimensional DLA, though several very simplified variants of DLA have been successfully analysed.

The original motivation for SLE actually came from investigating the Loop-erased random walk (a.k.a. LERW), which is a random curve introduced by Greg Lawler [51]. Consider some bounded simply-connected domain D in the plane. Let $G = G(D, \varepsilon)$ be the subgraph of a square grid of mesh ε that falls inside D and let V_∂ be the set of vertices of G that have fewer than 4 neighbors in D. Suppose that $0 \in D$, and let o be some vertex of G closest to 0. Start a simple random walk on G from o (at each step the walk jumps to any neighbor of the current position with equal probability). We keep track of the trajectory of the walk at each step, except that every time a loop is created, it is erased from the trajectory. The walk terminates when it first reaches V_∂, and the loop-erased random walk from o to V_∂ is the final trajectory. See Figure 5, where D is a disk.

The LERW is intimately related to the uniform spanning tree. In particular, if we collapse V_∂ to a single vertex v_∂ and take a random spanning tree of the resulting projection of G, where each possible spanning tree is chosen with equal probability, then the unique path in the tree joining o to v_∂ (as a set of edges) has precisely the same law as the LERW from o to V_∂ [74]. This is not a particular property of the square grid, the corresponding analog holds in an arbitrary finite graph. In the other direction, there is a marvelous algorithm discovered by David Wilson [99] which builds the uniform spanning tree by successively adding loop-erased random walks. The survey [66] is a good window into the beautiful theory of uniform spanning trees and forests.

Using sophisticated determinantcalculations and Temperley's bijection between the collection of spanning trees and a certain collection of dimer tilings (which is

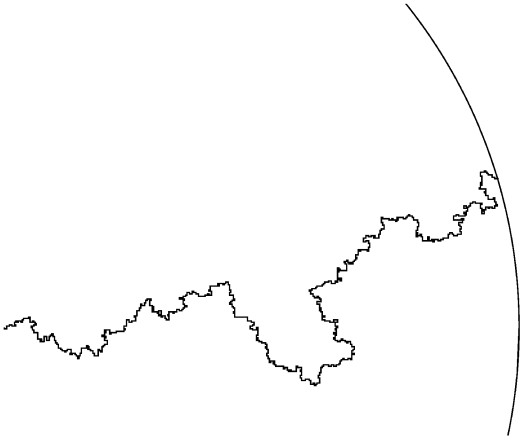


Figure 5. The LERW in a disk.

special to the planar setting), Richard Kenyon [39], [41], [40] was able to calculate several properties of the LERW. For example, it was shown that the variance of the winding number of the above LERW in $G(D, \varepsilon)$ is $(2+o(1))\log(1/\varepsilon)$ as $\varepsilon \searrow 0$, and that the growth exponent for the number of edges in a LERW is $5/4$.

In [79] it was shown that *if* the limit of the law of the LERW as $\varepsilon \searrow 0$ exists and is conformally invariant, then it is a radial SLE_2 path, when parametrized by capacity. (See subsection 2.1 for a description of two alternative topologies on spaces of probability measures on curves, for which this convergence may be stated.) In broad strokes, the reason why it should be an SLE path is basically the same as the argument presented above for the percolation interface. Two important properties of the percolation interface scaling limit were crucial in the above argument: conformal invariance and the following Markovian property. If we condition on an initial segment of the path, the remainder is an instance of the path in the domain slitted by the initial segment starting from the endpoint of the initial segment. Conformal invariance was believed to hold for the LERW scaling limit, while the Markovian property does hold for the reversal of the LERW.

The identification of the correct value of the parameter κ as 2 is based on Kenyon's calculated LERW winding variance growth rate and a calculation of the variance of the winding number of the radial SLE_κ path truncated at distance ε from the interior target point. The latter grows like $(\kappa + o(1))\log(1/\varepsilon)$.

It was also conjectured in [79] that the percolation interface discussed above converges to SLE_6. The identification of the parameter κ as 6 in this case was based on Cardy's formula [20] and the verification that the corresponding formula holds for SLE_κ if and only if $\kappa = 6$.

The percolation interface satisfies the following locality property. The evolution of the path (given its past) does not depend on the shape of the domain away from

the current location of the endpoint of the path. Though this is essentially obvious, it should be noted that other interesting paths (such as the LERW scaling limit) do not satisfy locality.

Another process that clearly satisfies locality is Brownian motion. Greg Lawler and Wendelin Werner [61], [62] studied the intersection exponents of planar Brownian motion and the relations between them. An example of an intersection exponent is the unique number $\xi(1, 1)$ such that the probability that the paths of two independent Brownian motions started at distance 1 apart within the unit disk $\mathbb{U}$ and stopped when they first hit the circle $R\,\partial\mathbb{U}$ do not intersect one another is $R^{-\xi(1,1)+o(1)}$ as $R \to \infty$. At the time, there were conjectures [27] for the values of many of these exponents, which were rational numbers, but only two of these could be proved rigorously (not accidentally, those had values 1 and 2). These exponents encode many fundamental properties of Brownian motion. For example, Lawler [50] showed that the dimension of the outer boundary of planar Brownian motion stopped at time $t = 1$, say, is $2(1 - \alpha)$ for a certain intersection exponent α. Lawler and Werner [61], [62] have proved certain relations between the intersection exponents, and have shown that any process which like Brownian motion satisfies conformal invariance and a certain version of the locality property necessarily has intersection exponents that are very simply related to the Brownian exponents.

Since SLE_6 was believed to be the scaling limit of the percolation interface, it should satisfy locality. It is also conformally invariant by definition. Thus, the Brownian exponents should apply to SLE_6. Indeed, in a series of papers [53], [54], [55] Lawler, Werner and the present author proved the conjectured values of the Brownian exponents by calculating the corresponding exponents for SLE_6 (and using the previous work by Lawler and Werner). Very roughly, one can say that the reason why the exponents of SLE are easier to calculate than the Brownian exponents is that the SLE path, though it may hit itself, does not cross itself. Thus, the outer boundary of the SLE path is drawn essentially in chronological order.

Later [58] it became clear that the relation between SLE_6 and Brownian motion is even closer than previously apparent: the outer boundary of Brownian motion started from 0 and stopped on hitting the unit circle $\partial\mathbb{U}$ has the same distribution as the outer boundary of a variant of SLE_6.

Lennart Carleson observed that, assuming conformal invariance, Cardy's formula is equivalent to the statement that $Q(D, \partial_1, \partial_2) = \mathrm{length}(\partial_2)$ when D is an equilateral triangle of sidelength 1, ∂_1 is its base, and $\partial_2 \subset \partial D$ is a line segment having the vertex opposite to ∂_1 as one of its endpoints. Smirnov [87] proved Carleson's form of Cardy's formula for critical site percolation on the triangular lattice (that is, the same percolation model we have described above) and showed that crossing probabilities between two arcs on the boundary of a simply connected domain are asymptotically conformally invariant. As a corollary, Smirnov concluded that the scaling limit of the percolation interface exists and is equal to the SLE_6 path. This connection enabled proving many conjectures about this percolation model. For example, the prediction [24], [73] that the probability that the cluster of the origin has diameter larger

than R decays like

$$\mathbf{P}\big[\text{the origin is in a cluster of diameter} \geq R\big] = R^{-5/48+o(1)} \tag{2}$$

as $R \to \infty$ was proved [56]. This value $5/48$ is an example of what is commonly referred to as a critical exponent. Building on earlier work by Kesten and others, as well as on Smirnov's theorem and SLE, Smirnov and Werner [88] were able to determine many useful percolation exponents. Julien Dubédat [25] has used SLE to prove Watts' [92] formula for the asymptotic probability that in a given rectangle there are both a white horizontal and vertical crossing (for the above percolation model at $p = p_c = 1/2$).

The next process for which conformal invariance and convergence to SLE was established is the LERW [59]. Contrary to Smirnov's proof for percolation, where convergence to SLE was a consequence of conformal invariance, in the case of the LERW the proof establishes conformal invariance as a consequence of the convergence to SLE_2. More specifically, the argument in [59] proceeds by considering the Loewner driving term of the discrete LERW (before passing to the limit) and proving that the driving term converges to an appropriately time-scaled Brownian motion. The same paper also shows that the uniform spanning tree scaling limit is conformally invariant, and the Peano curve associated with it (essentially, the boundary of a thickened uniform spanning tree) converges to SLE_8. Another difference between the results of [87] and [59] is that while the former is restricted to site percolation on the triangular lattice, the results in [59] are essentially lattice independent. Figure 5 above shows a fine LERW, which gives an idea of what an SLE_2 looks like. Likewise, Figure 6 shows a sample of an initial segment of the uniform spanning tree Peano curve in a rectangular domain. Note that the curve is space filling, as is SLE_8.

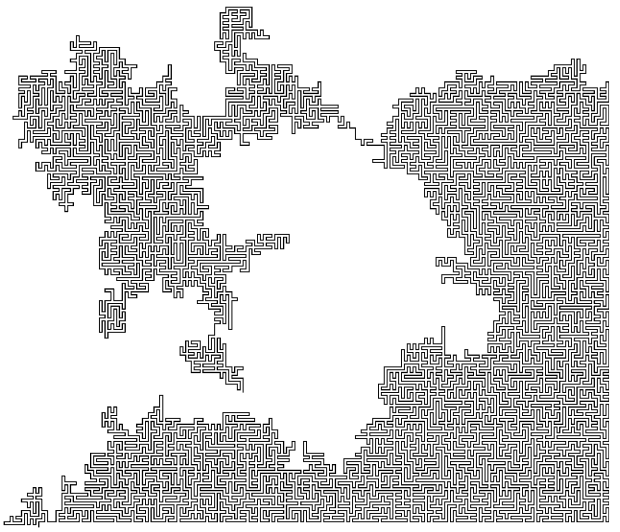


Figure 6. An initial segment of the uniform spanning tree Peano path.

Meanwhile, Gady Kozma [47] came up with a different proof that the LERW scaling limit exists. Although Kozma's proof does not identify the limit, it has the advantage of generalizing to three dimensions [48].

There are two discrete models for which convergence to SLE_4 has been established by Scott Sheffield and the present author. These models are the harmonic explorer [80] and the interface of the discrete Gaussian free field [81]. The discrete and continuous Gaussian free fields (a.k.a. the harmonic crystal) play an important role in the heuristic physics analysis of various statistical physics models. The discrete Gaussian free field is a probability measure on real valued functions defined on a graph, often a piece of a lattice. If, for example, the graph is a triangulation of a domain in the plane, an interface is a curve in the dual graph separating vertices where the function is positive from vertices where the function is negative. See [85] or [81] for further details, and see Figure 7 for a simulation of the harmonic explorer, and therefore an approximation of SLE_4.

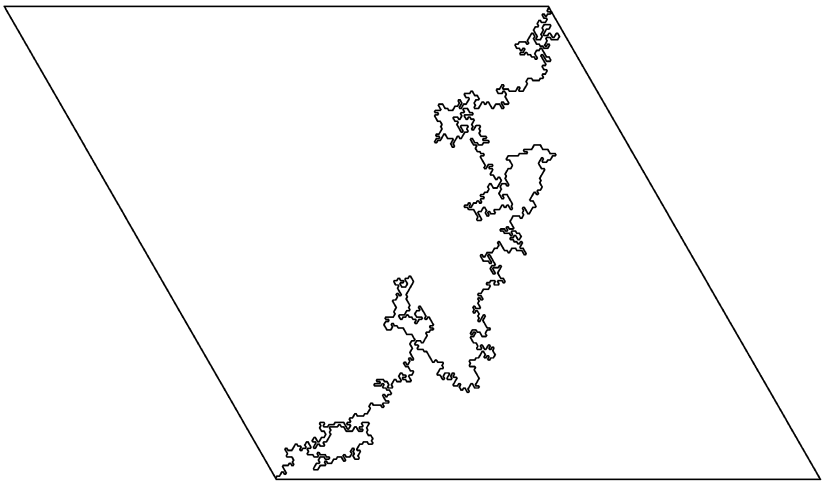


Figure 7. The harmonic explorer path.

Sheffield also announced work in progress connecting the Gaussian free field with SLE_κ for other values of κ. The basic idea is that while SLE_4 may be thought of as a curve solving the equation $h = 0$, where h is the Gaussian free field, for other κ, the SLE_κ curve may be considered as a solution of

$$c\,\mathrm{winding}(\gamma[0,t]) = h(\gamma(t)), \tag{3}$$

where c is a constant depending on κ. When $\kappa = 4$, the corresponding constant c is zero, which reduces to the setting of [81]. Alternatively, (3) can be heuristically written as $c\,\gamma'(s) = \exp(i\,h(\gamma(t))$, where s is the length parameter of γ. However, we stress that it is hard to make sense of these equations, for the Gaussian free field

is not a smooth function (in fact, it is not even a function but rather a distribution). Likewise, the SLE path is not rectifiable and its winding at most points is infinite.

As mentioned above, there are several different variants of SLE in simply connected domains: chordal, radial, as well as a few others, which we have not mentioned. These variants are rather closely related to one another [83]. There are also variants defined in the multiply-connected setting [101], [6], [5], [7]. One motivation for this study comes from statistical physics models, which are easy to define on multiply connected domains. Since one can easily vary the boundary conditions on different boundary components of the domain, it is clear that there is often more than one reasonable choice for the definition of the SLE path.

Finally, we mention an intriguing connection between Brownian motion and SLE_κ for $\kappa \in (8/3, 4]$. There is the notion of the Brownian loop soup [63], which is a Poisson measure on the space of Brownian motion loops. According to [93], the boundaries of clusters of a sample from the loop soup measure with intensity c are SLE_κ-like paths, where $\kappa = \kappa(c) \in (8/3, 4]$. The proof is to appear in a future joint work of Sheffield and Werner.

The above account describes some of the highlights of the developments in the field in the past several years. The rest of the paper will be devoted to a description of some problems where we hope to see some future progress. Some of these problems are obvious to anyone working in the field (though the solution is not obvious), while others are borrowed from several different sources. A few of the problems appear here for the first time. The paper [76] contains some additional problems.

2. Random processes converging to SLE

As we have seen, paths associated with several random processes have been proved to converge to various SLE paths. However, the list of processes where the convergence is expected but not proved yet is longer. This section will present questions of this sort, most of which have previously appeared in the literature.

The strategy of the proofs of convergence to SLE in the papers [59], [80], [81] is very similar. In these papers, a collection of martingales with respect to the filtration given by the evolution of the curve is used to gain information about the Loewner driving term of the discrete curve. Although such a proof is also possible for the percolation interface (using Cardy's formula), this technique was not available at the time, and Smirnov used instead an argument which uses the independence properties of percolation in an essential way and is therefore not likely to be applicable to many other models. Thus, it seems that presently the most promising technique is the martingale technique from [59].

2.1. Notions of convergence. To be precise, we must describe the meaning of these scaling limits. In fact, there are at least two distinct reasonable notions of convergence, which we now describe. Suppose that γ_n are random paths in the closed upper half

plane $\overline{\mathbb{H}}$ starting from 0. Consider the one-point compactification $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ of $\mathbb{C} = \mathbb{R}^2$, which may be thought of as the sphere S^2. The law of γ_n may be thought of as a Borel probability measure on the Hausdorff space of closed nonempty subsets of $\hat{\mathbb{C}}$. Since that Hausdorff space is compact, the space of Borel probability measures on it is compact with respect to weak convergence of measures [26]. We may say that γ_n converges in the Hausdorff sense to a random set $\gamma \subset \overline{\mathbb{H}} \cup \{\infty\}$ if the law of γ_n converges weakly to the law of γ. A similar definition applies to curves in the closure of a bounded domain $D \subset \mathbb{C}$. The above stated instances of convergence to SLE hold with respect to this notion. However, in the case of convergence to SLE_8, this does not mean very much, for SLE_8 fills up the domain.

The second notion of convergence is stronger. Suppose that each γ_n is a.s. continuous with respect to the half-plane capacity parametrization from ∞. (This holds, in particular, if γ_n is a.s. a (continuous) simple path.) If d is a metric on $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ compatible with its topology, then we may consider the metric

$$d^*(\beta_1, \beta_2) := \sup_{t \in [0,\infty)} d\big(\beta_1(t), \beta_2(t)\big)$$

on the space of continuous paths defined on $[0, \infty)$. We may say that γ_n converges to a random path γ weakly-uniformly if the law of γ_n converges weakly to the law of γ in the space of Borel measures with respect to the metric d^*. This implies Hausdorff convergence. Since d^* is finer than the Hausdorff metric, there are more functions from the space of paths to $\mathbb{R}$ that are continuous with respect to d^* than with respect to the Hausdorff metric. Consequently, weakly-uniform convergence is stronger than Hausdorff convergence. In all the results stated above saying that some random path converges to SLE, the convergence is weakly-uniform when the paths are parametrized by capacity or half-plane capacity (depending on whether the convergence is to radial or chordal SLE, respectively).

In the following, when we ask for convergence to SLE, we will mean weakly-uniform convergence. However, weaker nontrivial forms of convergence would also be very interesting.

2.2. Self avoiding walk. Let G be either the square, the hexagonal or the triangular grid in the plane, positioned so that 0 is some vertex in G. For $n \in \mathbb{N}$ consider the uniform measures on all self avoiding n-step walks in G that start at 0 and stay in the upper half plane. It has been shown in [60] that when $G = \mathbb{Z}^2$ the limiting measure as $n \to \infty$ exists. (The same proof probably applies for the other alternatives for G, provided that G is positioned so that horizontal lines through vertices in G do not intersect the relative interior of edges of G which they do not contain.)

Problem 2.1 ([60]). Let γ be a sample from the $n \to \infty$ limit of the uniform measure on n-step self avoiding paths in the upper half plane described above. Prove that the limit as $s \searrow 0$ of the law of $s\,\gamma$ exists and that it is $\mathrm{SLE}_{8/3}$.

The convergence may be considered with respect to either of the two topologies discussed in subsection 2.1.

In [60] some consequences of this convergence are indicated, as well as support for the conjecture.

There are some indications that the setting of the hexagonal lattice is easier: the rate of growth of the number of self avoiding paths on the hexagonal grid is predicted [72] to be $\left(2+\sqrt{2}+o(1)\right)^{n/2}$; no such prediction exists for the square grid or triangular lattice.

In dimensions $d > 4$ Takashi Hara and Gordon Slade [32] proved that the scaling limit of self-avoiding random walk is Brownian motion. This is also believed to be the case for $d = 4$. See [67] for references and further background on the self-avoiding walk.

2.3. Height models. There is a vast collection of height model interfaces that should converge to SLE_4. The one theorem in this regard is the convergence of the interface of the Gaussian free field [81]. This was motivated by Kenyon's theorem stating that the domino tiling height function converges to the Gaussian free field [42] and by Kenyon's conjecture that the double domino interface converges to SLE_4 (see [76] for a statement of this problem).

The domino height function is a function on $\mathbb{Z}^2$ associated with a domino tiling (see [42]). Its distribution is roughly (ignoring boundary issues) the uniform measure on functions $h\colon \mathbb{Z}^2 \to \mathbb{Z}$ such that $h(0,0) = 0$,

$$h(x,y) \mod 4 = \begin{cases} 0 & x, y \text{ even}, \\ 1 & x \text{ odd } y \text{ even}, \\ 2 & x, y \text{ odd}, \\ 3 & x \text{ even } y \text{ odd}, \end{cases}$$

and $|h(z) - h(z')| \in \{1,3\}$ if $|z - z'| = 1$, $z, z' \in \mathbb{Z}^2$.

Let D be a bounded domain in the plane whose boundary is a simple path in the triangular lattice, say. Let ∂_+ and ∂_- be complementary arcs in ∂D such that the two common endpoints of these arcs are midpoints of edges. Consider the uniform measure on functions h taking odd integer values on vertices in $\overline{D}$ such that $h = 1$ on ∂_+, $h = -1$ on ∂_-, and $|h(v) - h(u)| \in \{0,2\}$ for neighbors v, u. We may extend such a function h to $\overline{D}$ by affine interpolation within each triangle, and this interpolation is consistent along the edges. There is then a unique connected path γ that is the connected component of $h^{-1}(0)$ that contains the two endpoints of each of the two arcs $\partial_\pm$.

Problem 2.2. Is it true that the path γ tends to SLE_4? Does the law of h converge to the Gaussian free field?

The convergence we expect for h is in the same sense as in [42].

Note that if we restrict in the above the image of h to be $\{1,-1\}$, we obtain critical site percolation on the triangular grid, and the limit of the corresponding interface is in this case SLE_6.

Now suppose that $(D, \partial_+, \partial_-)$ is as above. Let $\lambda \in (0, 1/2]$ be some constant and consider the uniform measure on functions h taking real values on vertices in $\overline{D}$ such that $h = \lambda$ on ∂_+, $h = -\lambda$ on ∂_- and $|h(v) - h(u)| \leq 1$ for every edge $[v, u]$.

Problem 2.3. Is it true that for some value of λ the corresponding interface converges to SLE_4? Does the law of h converge in some sense to the Gaussian free field?

In the case of the corresponding questions for the discrete Gaussian free field, there is just one constant λ such that the interface converges to SLE_4. For other choices of λ the interface converges to a well-known variant of SLE_4 [81].

There are some restricted classes of height models for which convergence to the Gaussian free field is known [71]. It may still be very hard to prove that the corresponding interface converges to SLE_4. One interesting problem of this sort is the following.

Problem 2.4 ([81]). If we project the Gaussian free field onto the subspace spanned by the eigenfunctions of the Dirichlet Laplacian with eigenvalues in $[-r, r]$ and add the harmonic function with boundary values $\pm\lambda$ on $\partial_\pm$, does the corresponding interface converge to SLE_4 as $r \to \infty$ when λ is chosen appropriately?

The problem is natural, because the Gaussian free field is related to the Dirichlet Laplacian. In particular, the projections of the field onto the spaces spanned by eigenfunctions with eigenvalues in two disjoint intervals are independent.

2.4. The Ising, FK, and $O(n)$ loop models. The Ising model is a fundamental physics model for magnetism. Consider again a domain D adapted to the triangular lattice and a partition $\partial D = \partial_+ \cup \partial_-$ as in subsection 2.3. Now consider a function h that take the values ± 1 on vertices in $\overline{D}$ such that h is 1 on ∂_+ and -1 on ∂_-. On the collection of all such functions we put a probability measure such that the probability for a given h is proportional to $e^{-2\beta k}$, where β is a parameter and k is the number of edges $[v, u]$ such that $h(v) \neq h(u)$. This is known as the Ising model and the value associate to a vertex is often called a spin. It is known that the critical value β_c (which we do not define here in the context of the Ising model) for β satisfies $e^{2\beta} = \sqrt{3}$ (see [69], [35]). Again, the interface at the critical $\beta = \beta_c$ is believed to converge to an SLE path, this time SLE_3. For $\beta \in [0, \beta_c)$ fixed, the interface should converge to SLE_6. Note that when $\beta = 0$, the model is again identical to critical site percolation on the triangular grid, and the interface does converge to SLE_6.

Problem 2.5. Prove that when $\beta = \beta_c$, the interface converges to SLE_3 and when $\beta \in (0, \beta_c)$ to SLE_6.

When $\beta > \beta_c$ we do not expect convergence to SLE, and do not expect conformal invariance. The interface scaling limit is in this case a straight line segment if the domain is convex [75] (see also [29]).

The Fortuin–Kasteleyn [28] (FK) model (a.k.a. the random cluster model) is a probability measure on the collection of all subsets of the set of edges E of a finite

graph $G = (V, E)$. In the FK model, the measure of each $\omega \subset E$ is proportional to $\big(p/(1-p)\big)^{|\omega|} q^c$, where $q > 0$ and $p \in (0, 1)$ are parameters, $|\omega|$ is the cardinality of ω, and c is the number of connected components of the subgraph (V, ω). The FK model is very closely related to the well known Potts model [8], which is a generalization of the Ising model. Many questions about the Potts model can be translated to questions about the FK model and vice versa.

On the grid $\mathbb{Z}^2$, when $p = \sqrt{q}/(1+\sqrt{q})$ the FK model satisfies a form of self-duality.

Problem 2.6 ([76]). Prove that when $q \in (0, 4)$ and $p = \sqrt{q}/(1+\sqrt{q})$, the interface of the FK model on $\mathbb{Z}^2$ with appropriate boundary conditions converges to SLE_κ, where $\kappa = 4\pi/\cos^{-1}\big(-\sqrt{q}/2\big)$. (See [76] for further details.)

The $O(n)$ loop model on a finite graph $G = (V, E)$ is a measure on the collection of subgraphs of G where the degree of every vertex in the subgraph is 2. (The subgraph does not need to contain all the vertices.) The probability of each such subgraph is proportional to $x^e n^c$, where c is the number of connected components, e is the number of edges in the subgraph, and $x, n > 0$ are parameters. When n is a positive integer, the $O(n)$ loop model is derived from the $O(n)$ spin model, which is a measure on the set of functions which associate to every vertex a unit vector in $\mathbb{R}^n$.

In order to pin down a specific long path in the $O(n)$ loop model, we pick two points on the boundary of the domain and require that in the random subgraph the degrees of two boundary vertices near these two points be 1, while setting the degrees of all other boundary vertices to 0, say. Then the measure is supported on configurations with one simple path and a collection of loops.

Now we specialize to the hexagonal lattice. Set $x_c(n) := \big(2+\sqrt{2-n}\big)^{-1/2}$, which is the conjectured critical parameter [72].

Problem 2.7 ([36]). Prove that when $n \in [0, 2]$ and $x = x_c(n)$ (respectively, $x > x_c(n)$) the scaling limit of the path containing the two special boundary vertices is chordal SLE_κ, where $\kappa \in [8/3, 4]$ (respectively, $\kappa \in [4, 8]$) and $n = -2\cos(4\pi/\kappa)$.

When $x < x_c(n)$, we expect the scaling limit to be a straight line segment (if the domain is convex). The fact that at $n = 1$ we get the same limits as for the Ising model is no accident. It is not hard to see that the $O(1)$ loop measure coincides with the law of boundaries of Ising clusters. See [36] for further details.

Similar conjectures should hold in other lattices. However, the values of the critical parameters are expected to be different.

2.5. Lattice trees. We now present an example of a discrete model where we suspect that perhaps conformal invariance might hold. However, we do not presently have a candidate for the scaling limit.

Fix $n \in \mathbb{N}_+$, and consider the collection of all trees contained in the grid G that contain the origin and have n vertices. Select a tree T from this measure, uniformly at random.

Problem 2.8. What is the growth rate of the expected diameter of such a tree? If we rescale the tree so that the expected (or median) diameter is 1, is there a limit for the law of the tree as $n \to \infty$? What are its geometric and topological properties? Can the limit be determined?

It would be good to be able to produce some pictures. However, we presently do not know how to sample from this measure.

Problem 2.9. Produce an efficient algorithm which samples lattice trees approximately uniformly, or prove that such an algorithm does not exist.

See [86] for background on lattice trees and for results in high dimensions and [17] for an analysis of a related continuum model.

2.6. Percolation interface. It is natural to try to extend the understanding of percolation at p_c to percolation at a parameter p tending to p_c. One possible framework is as follows. Fix a parameter $q \in (0, 1)$. Suppose that in Figure 3 with small mesh $\varepsilon > 0$ we choose $p(\varepsilon)$ so that the probability to have a left to right crossing of white hexagons in some fixed 1×1 square in the upper half plane is q at percolation parameter $p = p(\varepsilon)$. The corresponding interface will still be an unbounded path starting at 0, but its distribution will be different from the interface at $p = 1/2$ if $q \neq 1/2$. Thus, it is natural to ask

Problem 2.10 (Lincoln Chayes (personal communication)). What is the scaling limit of the interface as $\varepsilon \searrow 0$ and $p = p(\varepsilon)$ if $q \neq 1/2$ is fixed?

This problem is also very closely related to a problem formulated by F. Camia, L. Fontes and C. Newman [18].

Site percolation on the triangular lattice is only one of several different models for percolation in the plane. Among discrete models, widely studied is bond percolation on the square grid. As for site percolation on the triangular lattice, the critical probability is again $p_c = 1/2$.

At present, Smirnov's proof does not work for bond percolation on the square grid. The proof uses the invariance of the model under rotation by $2\pi/3$. Thus, the following problem presents itself.

Problem 2.11. Prove Smirnov's theorem for critical bond percolation on $\mathbb{Z}^2$.

Some progress on this problem has been reported by Vincent Beffara [9].

There are other natural percolation models which have been studied. Among them we mention Voronoi percolation and the boolean model. In Voronoi percolation one has two independent Poisson point processes in the plane, W and B, with intensities p and $1-p$, respectively. Let $\hat{W}$ be the closure of the set of points in $\mathbb{R}^2$ closer to W than to B, and let $\hat{B}$ be the closure of the set of points closer to B. The set $\hat{W}$ is a sample from Voronoi percolation at parameter p. Some form of conformal invariance was proved for Voronoi percolation [14], but the version proved does not imply convergence

to SLE. It is neither stronger nor weaker than the conformal invariance proved by Smirnov. Notable recent progress has been made for Voronoi percolation by Bollobás and Riordan [15], who established the very useful Russo–Seymour–Welsh theorem, as well as $p_c = 1/2$ for Voronoi percolation.

Problem 2.12. Prove Smirnov's theorem for Voronoi percolation.

The boolean percolation model (a.k.a. continuum percolation) can be defined by taking a Poisson set of points $W \subset \mathbb{R}^2$ of intensity 1 and letting $\hat{W}$ be the set of points in the plane at distance at most r from W. Here, r is the parameter of the model. (Alternatively, one may fix $r = 1$, say, and let the intensity of the Poisson process be the parameter, but this is essentially the same, by scaling.) The Russo–Seymour–Welsh theorem is known for this model [1], [77], but the critical value of the parameter has not been identified. A nice feature which the model shares with Voronoi percolation is invariance under rotations.

Problem 2.13. Prove Smirnov's theorem for boolean percolation.

3. Critical exponents

The determination of critical exponents has been one motivation to prove conformal invariance for discrete models. For example, den Nijs and Nienhuis predicted [24], [73] that the probability that the critical percolation cluster of the origin has diameter larger than R is $R^{-5/48+o(1)}$ as $R \to \infty$. Likewise, the probability that a given site in the square $[-R, R]^2$ is pivotal for a left-right crossing of the square $[-2R, 2R]^2$ was predicted to be $R^{-5/4+o(1)}$. (Here, pivotal means that the occurrence or non-occurrence of a crossing would be modified by flipping the status of the site.) These and other exponents were proved for site percolation on the triangular grid using Smirnov's theorem and SLE [56], [88]. The determination of the exponents is very useful for the study of percolation.

Richard Kenyon [39] calculated by enumeration techniques involving determinants the asymptotics of the probability that an edge belongs to a loop-erased random walk. The probability decays like $R^{-3/4+o(1)}$ when the distance from the edge to the endpoints of the walk is R. However, Kenyon's estimate is much more precise; he shows that, in a specific domain, $R^{3/4}$ times the probability is bounded away from zero and infinity, and in fact estimates the probability as $f\, R^{-3/4}\big(1 + o(1)\big)$ as $R \to \infty$, where f is an explicit function of the position of the edge.

Thus, it is natural to ask for such precise estimates for the important percolation events as well. Namely,

Problem 3.1. Improve the estimates $R^{-5/48+o(1)}$ and $R^{-5/4+o(1)}$ mentioned above (as well as other similar estimates) to more precise formulas. It would be especially nice to obtain estimates that are sharp up to multiplicative constants.

In addition to the case of the loop-erased random walk mentioned above, estimates up to constants are known for events involving Brownian motions [57].

The difficulty in getting more precise estimates is not in the analysis of SLE. Rather, it is due to the passage between the discrete and continuous setting. Consequently, the above problem seems to be related to the following.

Problem 3.2. Obtain reasonable estimates for the speed of convergence of the discrete processes which are known to converge to SLE.

There are still critical exponents which do not seem accessible via an SLE analysis. For example, we may ask

Problem 3.3. Calculate the number α such that on the event that there is a left-right crossing in critical percolation in the square $[0, R]^2$, the expected length of the shortest crossing is $R^{\alpha+o(1)}$.

Ziff [102] predicts an exponent which is related to this α, but it seems that there is currently no prediction for the exact value of α.

4. Quantum gravity

Consider the uniform measure μ_n on equivalence classes of n-vertex triangulations of the sphere, where two triangulations are considered equivalent if there is a homeomorphism of the sphere taking one to the other. One may view a sample from this measure with the graph metric as a random geometry on the sphere. Such models go under the name "quantum gravity" in physics circles. One may also impose statistical physics models on such random triangulations. For example, the sample space may include such a triangulation (or rather, equivalence class of triangulations) together with a map h from the vertices to $\{-1, 1\}$. The measure of such a pair may be taken proportional to a^k, where k is the number of edges $[v, u]$ in the triangulation for which $h(v) \neq h(u)$ and $a > 0$ is a parameter. Thus, we are in effect considering a triangulation weighted by the Ising model partition function. (The partition function is in this case the sum of all the weights of such functions h on the given triangulation.) Likewise, one may weight the triangulation by other kinds of partition functions.

In some cases it is easier to make a heuristic analysis of such statistical physics models in the quantum gravity world than in the plane. The enigmatic KPZ formula of Knizhnik, Polyakov and Zamolodchikov [46] was used in physics to predict properties of statistical physics in the plane from the corresponding properties in quantum gravity. Basically, the KPZ formula is a formula relating exponents in quantum gravity to the corresponding exponents in plane geometry.

To date, there has been progress in the mathematical (as well as physical) understanding of the statistical physics in the plane as well as in quantum gravity [2], [3], [16]. However, there is still no mathematical understanding of the KPZ formula. In fact, the author's understanding of KPZ is too weak to even state a concrete problem.

However, we may ask about the scaling limit of μ_n. There has been significant progress lately describing some aspects of the geometry of samples from μ_n [2], [23]. In particular, it has been shown by Chassaing and Schaeffer [23] that if D_n is the graph-metric diameter of a sample from μ_n, then $D_n/n^{-1/4}$ converges in law to some random variable in $(0, \infty)$. However, the scaling limit of samples from μ_n is not known. On the collection of compact metric spaces, we may consider the Gromov–Hausdorff distance $d_{\mathrm{GH}}(X, Y)$, which is the infimum of the Hausdorff distance between subsets X^* and Y^* in a metric space Z^* over all possible triples (X^*, Y^*, Z^*) such that Z^* is a metric space, $X^*, Y^* \subset Z^*$, X is isometric with X^* and Y is isometric with Y^*. Let X_n be a sample from μ_n, considered as a metric space with the graph metric scaled by $n^{-1/4}$, and let μ_n^* denote the law of X_n.

Problem 4.1. Show that the weak limit $\lim_{n\to\infty} \mu_n^*$ with respect to the Gromov–Hausdorff metric exists. Determine the properties of the limit.

Note that [68] proves convergence of samples from μ_n, but the metric used there on the samples from μ_n is very different and consequently a solution of Problem 4.1 does not seem to follow.

5. Noise sensitivity, Fourier spectrum, and dynamical percolation

The indicator function of the event of having a percolation crossing in a domain between two arcs on the boundary is a boolean function of boolean variables. Some fundamental results concerning percolation are based on general theorems about boolean functions. (One can mention here the BK inequality, the Harris-FKG inequality and the Russo formula. See, e.g. [30].) Central to the theory of boolean functions is the Fourier expansion. Basically, if $f\colon \{-1, 1\}^n \to \mathbb{R}$ is any function of n bits, the Fourier–Walsh expansion of f is

$$f(x) = \sum_{S\subset[n]} \hat{f}(S)\, \chi_S(x),$$

where $\chi_S(x) = \prod_{j\in S} x_j$ for $S \subset [n] = \{1, \ldots, n\}$. When we consider $\{-1, 1\}^n$ with the uniform probability measure, the collection $\{\chi_S : S \subset [n]\}$ forms an orthonormal basis for $L^2(\{-1, 1\}^n)$. (Often other measures are also considered.) If we suppose that $\|f\|_2 = 1$ (in particular, this holds if $f\colon \{-1, 1\}^2 \to \{-1, 1\}$), then the Parseval identity gives $\sum_{S\subset[n]} \hat{f}(S)^2 = 1$. Thus, we get a probability measure μ_f on $2^{[n]} = \{S : S \subset [n]\}$ for which $\mu_f(\{S\}) = \hat{f}(S)^2$. The map $S \mapsto |S|$, assigning to each $S \subset [n]$ its cardinality pushes forward the measure μ_f to a measure $\tilde{\mu}_f$ on $\{0, 1, \ldots, n\}$. This measure $\tilde{\mu}_f$ may be called the Fourier spectrum of f. The Fourier spectrum encodes important information about f, and quite a bit of research on the subject exists [37]. For instance, one can read off from $\tilde{\mu}_f$ the sensitivity of f to noise (see [13]).

When f is a percolation crossing function (i.e., 1 if there is a crossing, -1 otherwise), the corresponding index set $[n]$ is identified with the collection of relevant sites or bonds, depending if it is a site or bond model. Though there is some partial understanding of the Fourier spectrum of percolation [82], the complete picture is unclear. For example, if the domain is approximately an $\ell \times \ell$ square in the triangular lattice, then for the indicator function f_ℓ of a crossing in critical site percolation it is known [82] that for every $\alpha < 1/8$

$$\tilde{\mu}_{f_\ell}([1, \ell^\alpha]) \to 0 \tag{4}$$

as $\ell \to \infty$. This is proved using the critical exponents for site percolation as well as an estimate for the Fourier coefficients of general functions (based on the existence of an algorithm computing the function which is unlikely to examine any specific input variable). On the other hand, using the percolation exponents one can show that (4) fails if $\alpha > 3/4$. This is based on calculating the expected number of sites pivotal for a crossing (i.e., a change of the value of the corresponding input variable would change the value of the function) as well as showing that the second moment is bounded by a constant times the square of the expectation. The expected number of pivotals is known [88] to be $\ell^{3/4+o(1)}$ as $\ell \to \infty$. It is reasonable to conjecture that (4) holds for every $\alpha < 3/4$.

Problem 5.1. Is it true that $\lim_{\ell\to\infty} \tilde{\mu}_{f_\ell}([\ell^{\alpha_1}, \ell^{\alpha_2}]) = 1$ if $\alpha_1 < 3/4 < \alpha_2$? Determine the asymptotic behavior of $\tilde{\mu}_{f_\ell}([\ell^{\alpha_1}, \ell^{\alpha_2}])$ as $\ell \to \infty$ for arbitrary $0 \le \alpha_1 < \alpha_2 \le 2$.

Estimates on the Fourier coefficients of percolation crossings (in an annulus) play a central part in the proof [82] that dynamical percolation has exceptional times. Dynamical percolation (introduced in [31]) is a model in which at each fixed time one sees an ordinary percolation configuration, but the random bits determining whether or not a site (or bond) is open undergo random independent flips at a uniform rate, according to independent Poisson processes. The main result of [82] is that dynamical critical site percolation on the triangular lattice has exceptional times at which there is an infinite percolation component. These set of times are necessarily of zero Lebesgue measure. A better understanding of the Fourier coefficients may lead to sharper results about dynamical percolation, such as the determination of the dimension of exceptional times. Some upper and lower bounds for the dimension are known [82].

One would hope to understand the measure μ_f geometrically. Gil Kalai (personal communication) has suggested the problem of determining the scaling limit of μ_f. More specifically, the Fourier index set S for critical percolation crossing of a square is naturally identified with a subset of the plane. If we rescale the square to have edge length 1 while refining the mesh, then μ_f may be thought of as a probability measure on the Hausdorff space $\mathcal{H}$ of closed subsets of the square. It is reasonable to expect that μ_f converges weakly to some probability measure μ on $\mathcal{H}$. We really do not know what samples from μ look like. Could it be that μ is supported on singletons? Alternatively, is it possible that $\mu[S = \text{entire square}] = 1$? Is S a Cantor set μ-a.s.?

Problem 5.2 (Gil Kalai, personal communication). Prove that the limiting measure μ exists and determine properties of samples from μ.

Kalai suspects (personal communication) that the set S is similar to the set of pivotal sites (which is a.s. a Cantor set in the scaling limit). This is supported by the easily verified fact that $\mathbf{P}\big[i \in S\big] = \mathbf{P}\big[i \text{ pivotal}\big]$ and $\mathbf{P}\big[i, j \in S\big] = \mathbf{P}\big[i, j \text{ pivotal}\big]$ hold for arbitrary boolean functions. Examples of functions where the scaling limit of S has been determined are provided by Tsirelson [90], [91].

One may try to study a scaling limit of dynamical percolation. Consider dynamical critical site percolation on a triangular lattice of mesh ε, where the rate at which the sites flip is $\lambda > 0$. We choose $\lambda = \lambda(\varepsilon)$ so that the correlation between having a left-right crossing of a fixed square at time 0 and at time 1 is $1/2$, say. Noise sensitivity of percolation [13] shows that $\lim_{\varepsilon \searrow 0} \lambda(\varepsilon) = 0$ and the results of [82] imply that $\lambda(\varepsilon) = \varepsilon^{O(1)}$ and $\varepsilon = \lambda(\varepsilon)^{O(1)}$ for $\varepsilon \in (0, 1]$. It is not hard to invent (several different) notions in which to take the limit of dynamical percolation as $\varepsilon \searrow 0$.

Problem 5.3. Prove that the scaling limit of dynamical critical percolation exists. Prove that correlations between crossing events at different times $t_1 < t_2$ decay to zero as $t_2 - t_1 \to \infty$ and that a change in a crossing event becomes unlikely if $t_2 - t_1 \to 0$.

Since the correlation between events occurring at different times can be expressed in terms of the Fourier coefficients [13], [82], it follows that the second statement in Problem 5.3 is very much related to strong concentration of the measure $\tilde{\mu}_{f,\ell}$, in the spirit of Problem 5.1.

Because of the dependence of λ on ε, it is not reasonable to expect the dynamical percolation scaling limit to be invariant under maps of the form $f \times$ identity, where $f\colon D \to D'$ is conformal and the identity map is applied to the time coordinate. In particular, in the case where $f(z) = a\,z$, $a > 0$, one should expect dynamical percolation to be invariant under the map $f \times (t \mapsto a^{\beta}\, t)$, where $\beta := -\lim_{\varepsilon \searrow 0} \log \lambda(\varepsilon)/\log \varepsilon$ (and this limit is expected to exist). It is not too hard to see that $\beta = 3/4$ if the answer to the first question in Problem 5.1 is yes.

This suggests a modified form of conformal invariance for dynamical percolation. Suppose that $F(z, t)$ has the form $F(z, t) = \big(f(z), g(z, t)\big)$, where $f\colon D \to D'$ is conformal and g satisfies $\partial_t g(z, t) = |f'(z)|^{\beta}$, with the above value of β. Is dynamical percolation invariant under such maps? If such invariance is to hold, it would be in a "relativistic" framework, in which one does not consider crossings occurring at a specific time slice, but rather inside a space-time set. It is not clear if one can make good sense of that.

6. LERW and UST

The loop-erased random walk and the uniform spanning tree are models where very detailed knowledge exists. They may be studied using random walks and electrical

network techniques, and in the two-dimensional setting also by SLE as well as domino tiling methods. (See [66] and the references cited there.) However, some open problems still remain.

One may consider the random walk in a fine mesh lattice in the unit disk, which is stopped when it hits the boundary of the disk. The random walk converges to Brownian motion while its loop-erasure converges to SLE_2. It is therefore reasonable to expect that the law of the pair (random walk, its loop-erasure) converges to a coupling of Brownian motion and SLE_2. (If not, a subsequential limit will converge.)

Problem 6.1. In this coupling, is the SLE_2 determined by the Brownian motion?

It seems that this question occurred to several researchers independently, including Wendelin Werner (personal communication).

Of course, one cannot naively loop-erase the Brownian motion path, because there is no first loop to erase and there are cases where the erasure of one loop eliminates some of the other loops.

It is also interesting to try to extend some of the understanding of probabilistic statistical physics models beyond the planar setting to higher genus. The following problem in this direction was proposed by Russell Lyons (personal communication).

Consider the uniform spanning tree on a fine square grid approximation of a torus. There is a random graph dual to the tree, which consists of the dual edges perpendicular to primal edges not in the tree. It is not hard to see that this random dual of the tree contains precisely three edge-simple closed paths (i.e., no repeating edges), and that these paths are not null-homotopic.

Problem 6.2. Determine the distribution of the triple of homotopy classes containing these three closed paths.

The problem would already be interesting for a square torus, but one could hope to get the answer as a function of the geometry of the torus.

7. Non-discrete problems

In this section we mention some problems about the behavior of SLE itself, which may be stated without relation to any particular discrete model.

The parametrization of the SLE path by capacity is very convenient for many calculations. However, in some situations, for example when you consider the reversal of the path, this parametrization is not so useful. It would be great if we had an understanding of a parametrization by a kind of Hausdorff measure. Thus we are led to

Problem 7.1. Define a Hausdorff measure on the SLE path which is σ-finite.

We would expect the measure to be a.s. finite on compact subsets of the plane.

That the Hausdorff dimension of the SLE path is $\min\{2, 1+\kappa/8\}$ has been established by Vincent Beffara [10]. When $\kappa < 8$ (in which case the path has zero area), we expect the σ-finite Hausdorff measure to be the Hausdorff measure with respect to the gauge function $\phi(r) = r^d \log\log(1/r)$, where $d = 1+\kappa/8$ is the Hausdorff dimension. This is based on past experience with similar random paths [89]. However, in order to prove that this Hausdorff measure is σ-finite, one should probably find alternative constructions of the measure. In the case $\kappa \le 4$, where the SLE path is a simple path a.s., one could try to use conformal maps from the unit disk to the two components in the complement of the curve in the upper half plane. If f is such a map, it might be possible to show that the limit of the length measure of the image of the circle $r\,\partial\mathbb{U}$, rescaled appropriately, has a limit as $r \nearrow 1$. Another approach, which was discussed by Tom Kennedy [38], would be to study the α-variation of the SLE path, though this seems hard to handle.

One may also consider other measures of growth for the SLE path. For example, when $\kappa > 4$, we may study the area of the SLE hull. It would be interesting to study the various relations between different measures of growth.

It is also natural to ask what kind of sets are visited by the SLE path. More precisely:

Problem 7.2. Fix $\kappa < 8$. Find necessary or sufficient conditions on a deterministic compact set $K \subset \overline{\mathbb{H}}$ to satisfy $\mathbf{P}\big[K \cap \gamma \neq \emptyset\big] > 0$, where γ is the SLE_κ path.

The case $K \subset \mathbb{R}$ is of particular interest.

When $\kappa = 8/3$ and $\mathbb{H} \setminus K$ is simply connected, there is a simple explicit formula [58] for $\mathbf{P}\big[\gamma \cap K \neq \emptyset\big]$. It is not clear if such formulas are also available for other values of κ. Wendelin Werner [96] proved the existence of a random collection of $\mathrm{SLE}_{8/3}$-like loops with some wonderful properties. In particular, the expected number of loops which separate two boundary components of an annulus is conformally invariant, and therefore a function of the conformal modulus of the annulus. However, this function is not known explicitly.

Many of the random interfaces which are known or believed to converge to SLE are reversible, in the sense that the reversed path has the same law as the original path (with respect to a slightly modified setup). This motivates the following problem from [76].

Problem 7.3. Let γ be the chordal SLE_κ path, where $\kappa \le 8$. Prove that up to reparametrization, the image of γ under inversion in the unit circle (that is, the map $z \mapsto 1/\bar{z}$) has the same law as γ itself.

The reason that we restrict to the case $\kappa \le 8$ is that this is false when $\kappa > 8$ (as will be proved in a forthcoming joint paper with Steffen Rohde). Indeed, there are no known models from physics that are believed to be related to SLE_κ when $\kappa > 8$. Sheffield (personal communication) expects that at least in the case $\kappa < 4$ Problem 7.3 can be answered by studying the relationship between the Gaussian free field and SLE.

Acknowledgments. Greg Lawler, Wendelin Werner and Steffen Rohde have collaborated with me during the early stages of the development of SLE. Without them the subject would not be what it is today. I wish to thank Itai Benjamini, Gil Kalai, Richard Kenyon, Scott Sheffield, Jeff Steif and David Wilson for numerous inspiring conversations. Thanks are also due to Yuval Peres for useful advice, especially concerning Problem 7.1.

References

[1] Alexander, K. S., The RSW theorem for continuum percolation and the CLT for Euclidean minimal spanning trees. *Ann. Appl. Probab.* **6** (2) (1996), 466–494.

[2] Angel, O., Growth and percolation on the uniform infinite planar triangulation. *Geom. Funct. Anal.* **13** (5) (2003), 935–974.

[3] ——, Scaling of Percolation on Infinite Planar Maps, I. Preprint, 2005; arXiv:math.Pr/0501006.

[4] Barlow, M. T., Fractals, and diffusion-limited aggregation. *Bull. Sci. Math.* **117** (1) (1993), 161–169.

[5] Bauer, R. O., and Friedrich, R. M., On radial stochastic Loewner evolution in multiply connected domains. *J. Funct. Anal.* **237** (2) (2006), 565–588.

[6] ——, Stochastic Loewner evolution in multiply connected domains. *C. R. Math. Acad. Sci. Paris* **339** (8) (2004), 579–584.

[7] ——, On Chordal and Bilateral SLE in multiply connected domains. Preprint, 2005; arXiv:math.Pr/0503178.

[8] Baxter, R. J., Kelland, S. B., and Wu, F. Y., Equivalence of the Potts model or Whitney polynomial with an ice-type model. *J. Phys. A* **9** (1976), 397–406.

[9] Beffara, V., Critical percolation on other lattices. Talk at the Fields Institute, 2005; http://www.fields.utoronto.ca/audio/05-06/.

[10] ——, The dimension of the SLE curves. Preprint, 2002; arXiv:math.Pr/0211322.

[11] ——, Cardy's formula on the triangular lattice, the easy way. Preprint, 2005; http://www.umpa.ens-lyon.fr/~vbeffara/files/Proceedings-Toronto.pdf.

[12] Beffara, V., and Sidoravicius, V., Percolation theory. Preprint, 2005; arXiv:math.Pr/0507220.

[13] Benjamini, I., Kalai, G., and Schramm, O., Noise sensitivity of Boolean functions and applications to percolation. *Inst. Hautes Études Sci. Publ. Math.* **90** (1999), 5–43.

[14] Benjamini, I., and Schramm, O., Conformal invariance of Voronoi percolation. *Comm. Math. Phys.* **197** (1) (1998), 75–107.

[15] Bollobás, B., and Riordan, O., The critical probability for random Voronoi percolation in the plane is 1/2. *Probab. Theory Related Fields* **136** (2006), 417–468.

[16] Bousquet-Melou, M., and Schaeffer, G., The degree distribution in bipartite planar maps: applications to the Ising model. Preprint, 2002; arXiv:math.CO/0211070.

[17] Brydges, D. C., and Imbrie, J. Z., Branched polymers and dimensional reduction. *Ann. of Math.* (2) **158** (3) (2003), 1019–1039.

[18] Camia, F., Fontes, L. R. G., and Newman, C. M., The Scaling Limit Geometry of Near-Critical 2D Percolation. Preprint, 2005; cond-mat/0510740.

[19] Camia, F., and Newman, C. M., The Full Scaling Limit of Two-Dimensional Critical Percolation. Preprint, 2005; arXiv:math.Pr/0504036.

[20] Cardy, J., Critical percolation in finite geometries. *J. Phys. A* **25** (4) (1992), L201–L206.

[21] ——, SLE for theoretical physicists. *Ann. Physics* **318** (1) (2005), 81–118.

[22] Carleson, L., and Makarov, N., Aggregation in the plane and Loewner's equation. *Comm. Math. Phys.* **216** (3) (2001), 583–607.

[23] Chassaing, P., and Schaeffer, G., Random planar lattices and integrated superBrownian excursion. *Probab. Theory Related Fields* **128** (2) (2004), 161–212.

[24] den Nijs, M., A relation between the temperature exponents of the eight-vertex and the q-state potts model. *J. Phys. A* **12** (1979), 1857–1868.

[25] Dubédat, J., Excursion decompositions for SLE and Watts' crossing formula. *Probab. Theory Related Fields* **134** (3) (2006), 453–488.

[26] Dudley, R. M., *Real analysis and probability*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1989.

[27] Duplantier, B., and Kwon, K.-H., Conformal invariance and intersection of random walks. *Phys. Rev. Lett.* **61** (1988), 2514–2517.

[28] Fortuin, C. M., and Kasteleyn, P. W., On the random-cluster model. I. Introduction and relation to other models. *Physica* **57** (1972), 536–564.

[29] Greenberg, L., and Ioffe, D., On an invariance principle for phase separation lines. *Ann. Inst. H. Poincaré Probab. Statist.* **41** (5) (2005), 871–885.

[30] Grimmett, G., *Percolation*. Second ed., Grundlehren Math. Wiss. 321, Springer-Verlag, Berlin 1999.

[31] Häggström, O., Peres, Y., and Steif, J. E., Dynamical percolation. *Ann. Inst. H. Poincaré Probab. Statist.* **33** (4) (1997), 497–528.

[32] Hara, T., and Slade, G., Self-avoiding walk in five or more dimensions. I. The critical behaviour. *Comm. Math. Phys.* **147** (1) (1992), 101–136.

[33] Harris, T. E., A lower bound for the critical probability in a certain percolation process. *Proc. Cambridge Philos. Soc.* **56** (1960), 13–20.

[34] Hastings, M. B., and Levitov, L. S., Laplacian growth as one-dimensional turbulence. *Physica D* **116** (1998), 244–252.

[35] Itzykson, C., and Drouffe, J.-M., *Statistical field theory*. Vol. 1. From Brownian motion to renormalization and lattice gauge theory, Cambridge Monogr. Math. Phys., Cambridge University Press, Cambridge 1989.

[36] Kager, W., and Nienhuis, B., A guide to stochastic Löwner evolution and its applications. *J. Statist. Phys.* **115** (5–6) (2004), 1149–1229.

[37] Kalai, G., and Safra, S., Threshold phenomena and influence. In *Computational Complexity and Statistical Physics* (A. G. Percus, G. Istrate and C. Moore, eds.), St. Fe Inst. Stud. Sci. Complex., Oxford University Press, New York 2006, 25–60.

[38] Kennedy, T., Monte Carlo comparisons of the self-avoiding walk and SLE as parameterized curves. Preprint, 2005; arXiv:math.Pr/0510604.

[39] Kenyon, R., The asymptotic determinant of the discrete Laplacian. *Acta Math.* **185** (2) (2000), 239–286.

[40] ——, Conformal invariance of domino tiling. *Ann. Probab.* **28** (2) (2000), 759–795.

[41] ——, Long-range properties of spanning trees. Probabilistic techniques in equilibrium and nonequilibrium statistical physics. *J. Math. Phys.* **41** (3) (2000), 1338–1363.

[42] ——, Dominos and the Gaussian free field. *Ann. Probab.* **29** (3) (2001), 1128–1137.

[43] Kesten, H., The critical probability of bond percolation on the square lattice equals 1/2. *Comm. Math. Phys.* **74** (1) (1980), 41–59.

[44] ——, Upper bounds for the growth rate of DLA. *Phys. A* **168** (1) (1990), 529–535.

[45] ——, Some highlights of percolation. In *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. I, Higher Ed. Press, Beijing 2002, 345–362.

[46] Knizhnik, V. G., Polyakov, A. M., and Zamolodchikov, A. B., Fractal structure of 2D-quantum gravity. *Modern Phys. Lett. A* **3** (8) (1988), 819–826.

[47] Kozma, G., Scaling limit of loop erased random walk – a naive approach. Preprint 2002, arXiv:math.Pr/0212338.

[48] ——, The scaling limit of loop-erased random walk in three dimensions. Preprint 2005, arXiv:math.Pr/0508344.

[49] Langlands, R., Pouliot, P., and Saint-Aubin, Y., Conformal invariance in two-dimensional percolation. *Bull. Amer. Math. Soc. (N.S.)* **30** (1) (1994), 1–61.

[50] Lawler, G. F., The dimension of the frontier of planar Brownian motion. *Electron. Comm. Probab.* **1** (5) (1996), 29–47 (electronic).

[51] ——, A self-avoiding random walk. *Duke Math. J.* **47** (3) (1980), 655–693.

[52] ——, *Conformally invariant processes in the plane*. Math. Surveys Monogr. 114, Amer. Math. Soc., Providence, RI, 2005.

[53] Lawler, G. F., Schramm, O., and Werner, W., Values of Brownian intersection exponents. I. Half-plane exponents. *Acta Math.* **187** (2) (2001), 237–273.

[54] ——, Values of Brownian intersection exponents. II. Plane exponents. *Acta Math.* **187** (2) (2001), 275–308.

[55] ——, Values of Brownian intersection exponents. III. Two-sided exponents. *Ann. Inst. H. Poincaré Probab. Statist.* **38** (1) (2002), 109–123.

[56] ——, One-arm exponent for critical 2D percolation. *Electron. J. Probab.* **7** (2) (2002), 13 pp. (electronic).

[57] ——, Sharp estimates for Brownian non-intersection probabilities. *In and out of equilibrium* (Mambucaba, 2000), Progr. Probab. 51, Birkhäuser Boston, Boston, MA, 2002, pp. 113–131.

[58] ——, Conformal restriction: the chordal case. *J. Amer. Math. Soc.* **16** (4) (2003), 917–955 (electronic).

[59] ——, Conformal invariance of planar loop-erased random walks and uniform spanning trees. *Ann. Probab.* **32** (1B) (2004), 939–995.

[60] ——, On the scaling limit of planar self-avoiding walk. In *Fractal geometry and applications: a jubilee of Benoît Mandelbrot*, Part 2, Proc. Sympos. Pure Math. 72, Amer. Math. Soc., Providence, RI, 2004, 339–364.

[61] Lawler, G. F., and Werner, W., Intersection exponents for planar Brownian motion. *Ann. Probab.* **27** (4) (1999), 1601–1642.

[62] ——, Universality for conformally invariant intersection exponents. *J. Eur. Math. Soc. (JEMS)* **2** (4) (2000), 291–328.

[63] ——, The Brownian loop soup. *Probab. Theory Related Fields* **128** (4) (2004), 565–588.

[64] Lévy, P., *Processus Stochastiques et Mouvement Brownien. Suivi d'une note de M. Loève*. Gauthier-Villars, Paris 1948.

[65] Löwner, K. (C. Loewner), Untersuchungen über schlichte konforme Abbildungen des Einheitskreises, I. *Math. Ann.* **89** (1923), 103–121.

[66] Lyons, R., A bird's-eye view of uniform spanning trees and forests. In *Microsurveys in discrete probability* (Princeton, NJ, 1997), Amer. Math. Soc., Providence, RI, 1998, 135–162.

[67] Madras, N., and Slade, G., *The self-avoiding walk*. Probab. Appl., Birkhäuser, Boston, MA, 1993.

[68] Marckert, J. F., and Mokkadem, A., Limit of Normalized Quadrangulations: the Brownian map. *Ann. Probab.* **34** (2006), 2144–2202.

[69] McCoy, B. M., and Wu, T. T., *The two-dimensional Ising model*. Harvard University Press, Cambridge, Mass., 1973.

[70] Meester, R., and Roy, R., *Continuum percolation*. Cambridge Tracts in Math. 119, Cambridge University Press, Cambridge 1996.

[71] Naddaf, A., and Spencer, T., On homogenization and scaling limit of some gradient perturbations of a massless free field. *Comm. Math. Phys.* **183** (1) (1997), 55–84.

[72] Nienhuis, B., Exact critical point and critical exponents. *Phys. Rev. Lett.* **49** (1982), 1062–1065.

[73] ——, Coulomb gas description of 2-d critical behaviour. *J. Statist. Phys.* **34** (1984), 731–761.

[74] Pemantle, R., Choosing a spanning tree for the integer lattice uniformly. *Ann. Probab.* **19** (4) (1991), 1559–1574.

[75] Pfister, C.-E., and Velenik, Y., Interface, surface tension and reentrant pinning transition in the 2D Ising model. *Comm. Math. Phys.* **204** (2) (1999), 269–312.

[76] Rohde, S., and Schramm, O., Basic properties of SLE. *Ann. of Math.* (2) **161** (2) (2005), 883–924.

[77] Roy, R., The Russo-Seymour-Welsh theorem and the equality of critical densities and the "dual" critical densities for continuum percolation on $\mathbf{R}^2$. *Ann. Probab.* **18** (4) (1990), 1563–1575.

[78] Russo, L., A note on percolation. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **43** (1) (1978), 39–48.

[79] Schramm, O., Scaling limits of loop-erased random walks and uniform spanning trees. *Israel J. Math.* **118** (2000), 221–288.

[80] Schramm, O., and Sheffield, S., Harmonic explorer and its convergence to SLE_4. *Ann. Probab.* **33** (6) (2005), 2127–2148.

[81] ——, Contour lines of the 2D Gaussian free field. In preparation, 2006.

[82] Schramm, O., and Steif, J. E., Quantitative noise sensitivity and exceptional times for percolation. Preprint, 2005; arXiv:math.Pr/0504586.

[83] Schramm, O., and Wilson, D. B., SLE coordinate changes. *New York J. Math.* **11** (2005), 659–669.

[84] Seymour, P. D., and Welsh, D. J. A., Percolation probabilities on the square lattice. Advances in graph theory (Cambridge Combinatorial Conf., Trinity College, Cambridge, 1977); *Ann. Discrete Math.* **3** (1978), 227–245.

[85] Sheffield, S., Gaussian free fields for mathematicians. Preprint, 2003; arXiv:math. Pr/0312099.

[86] Slade, G., Lattice trees, percolation and super-Brownian motion. In *Perplexing problems in probability*, Progr. Probab. 44, Birkhäuser, Boston, MA, 1999, 35–51.

[87] Smirnov, S., Critical percolation in the plane: conformal invariance, Cardy's formula, scaling limits. *C. R. Acad. Sci. Paris Sér. I Math.* **333** (3) (2001), 239–244.

[88] Smirnov, S., and Werner, W., Critical exponents for two-dimensional percolation. *Math. Res. Lett.* **8** (5–6) (2001), 729–744.

[89] Taylor, S. J., The measure theory of random fractals. *Math. Proc. Cambridge Philos. Soc.* **100** (3) (1986), 383–406.

[90] Tsirelson, B., Fourier-Walsh coefficients for a coalescing flow (discrete time). TAU RP-SOR-99-02, 1999; arXiv:math.Pr/9903068.

[91] ——, Scaling limit of Fourier-Walsh coefficients (a framework). TAU RP-SOR-99-04, 1999; arXiv:math.Pr/9903121.

[92] Watts, G. M. T., A crossing probability for critical percolation in two dimensions. *J. Phys. A* **29** (14) (1996), L363–L368.

[93] Werner, W., SLEs as boundaries of clusters of Brownian loops. *C. R. Math. Acad. Sci. Paris* **337** (7) (2003), 481–486.

[94] ——, Random planar curves and Schramm-Loewner evolutions. In *Lectures on probability theory and statistics*, Lecture Notes in Math. 1840, Springer-Verlag, Berlin 2004, 107–195.

[95] ——, Conformal restriction and related questions. *Probab. Surv.* **2** (2005), 145–190 (electronic).

[96] ——, The conformally invariant measure on self-avoiding loops. Preprint, 2005; arXiv:math. Pr/0511605.

[97] ——, Conformal restriction properties. In *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume III, EMS Publishing House, Zürich 2006, 741–762.

[98] Wierman, J. C., Bond percolation on honeycomb and triangular lattices. *Adv. in Appl. Probab.* **13** (2) (1981), 298–313.

[99] Wilson, D. B., Generating random spanning trees more quickly than the cover time. In *Proceedings of the Twenty-eighth Annual ACM Symposium on the Theory of Computing* (Philadelphia, PA, 1996), ACM , New York 1996, 296–303.

[100] Witten, T. A., and Sander, L. M., Diffusion-limited aggregation. *Phys. Rev. B* (3) **27** (9) (1983), 5686–5697.

[101] Zhan, D., Stochastic Loewner evolution in doubly connected domains. *Probab. Theory Related Fields* **129** (3) (2004), 340–380.

[102] Ziff, R. M., Exact critical exponent for the shortest-path scaling function in percolation. *J. Phys. A* **32** (43) (1999), L457–L459.

Microsoft Corporation, Redmond, Washington, U.S.A.

Submitted to the Annals of Probability
arXiv: math.PR/0000000

ON THE SCALING LIMITS OF PLANAR PERCOLATION

By Oded Schramm[*,‡] and Stanislav Smirnov[†,§]
with an Appendix by Christophe Garban [¶]

Microsoft Research[‡] and Université de Genève[§]
CNRS, ENS Lyon[¶]

We prove Tsirelson's conjecture that any scaling limit of the critical planar percolation is a black noise. Our theorems apply to a number of percolation models, including site percolation on the triangular grid and any subsequential scaling limit of bond percolation on the square grid. We also suggest a natural construction for the scaling limit of planar percolation, and more generally of any discrete planar model describing connectivity properties.

Contents.

1 Introduction 2
1.1 Motivation 2
1.2 Percolation basics and notation 7
1.3 Definition of the scaling limits 13
1.4 Statement of main results 15
2 The Discrete Gluing Theorem 23
3 The space of lower sets 27
4 Uniformity in the mesh size 32
5 Factorization 40
A Continuity of crossing events 42
B Multi-scale bound on the four-arm event on $\mathbb{Z}^2$ (by Christophe Garban) 49
References 53
Author's addresses 55

[*]December 10, 1961 – September 1, 2008

[†]Supported by the European Research Council AG CONFRA, the Swiss National Science Foundation, and by the Chebyshev Laboratory (Faculty of Mathematics and Mechanics, St. Petersburg State University) under the grant of the government of the Russian Federation

AMS 2000 subject classifications: Primary 60K35; secondary 28C20, 82B43, 60G60
Keywords and phrases: percolation, noise, scaling limit

I. Benjamini, O. Häggström (eds.), *Selected Works of Oded Schramm*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-9675-6_35, © Springer Science+Business Media, LLC 2011

1. Introduction.

1.1. *Motivation.* This paper has a two-fold motivation: to propose a new construction for the (subsequential) scaling limits of the critical and near-critical percolation in the plane, and to show that such limits are two-dimensional *black noises* as suggested by Boris Tsirelson.

Percolation is perhaps the simplest statistical physics model exhibiting a phase transition. We will be interested in planar percolation models, the archetypical examples being site percolation on the triangular lattice and bond percolation on the square lattice. More generally, our theorems apply to a wide class of models on planar graphs, satisfying the assumptions outlined below.

Consider a planar graph and fix a *percolation probability* $p \in [0,1]$. In *site percolation*, each site is declared *open* or *closed* with probabilities p and $1-p$, independently of others. In the pictures we represent open and closed sites by blue and yellow colors correspondingly. One then studies open *clusters*, which are maximal connected subgraphs of open sites. In *bond percolation*, each bond is similarly declared open or closed independently of others, and one studies clusters of bonds.

Percolation as a mathematical model was originally introduced by Broadbent and Hammersley to model the flow of liquid through a random porous medium. It is assumed that the liquid can move freely between the neighboring open sites, and so it can flow between two regions if there is an open path between them. Fixing a quad Q (= a topological quadrilateral), we say that there is an *open crossing* (between two distinguished opposite sides) for a given percolation configuration, if for its restriction to the quad, there is an open cluster intersecting both opposite sides. One then defines the crossing probability for Q as the probability that such a crossing exists.

We will be interested in the (subsequential) scaling limits of percolation, as the lattice mesh tends to zero. It is expected that for a large class of models the *sharp threshold* phenomenon occurs: there is a critical value p_c such that for a fixed quad Q the crossing probability tends to 1 for $p > p_c$ and to 0 for $p < p_c$ in the scaling limit. For bond percolation on the square lattice and site percolation on the triangular lattice, it is known that $p_c = 1/2$ by the classical theorem of Kesten [Kes80].

In the critical case $p = p_c$ the crossing probabilities are expected to have non-trivial scaling limits. Moreover, those are believed to be universal, conformally invariant, and given by Cardy's formula [LPSA94, Car92]. For the models we are concerned with, this was established only for site percolation on the triangular lattice [Smi01b, Smi01a]. However, the Russo-Seymour-

Welsh theory provides existence of non-trivial subsequential limits, which is enough for our purposes. Our theory also applies to near-critical scaling limits like in [Nol08], when the mesh tends to zero and the percolation probability to its critical value in a coordinated way, so as the crossing probabilities have non-trivial limits, different from the critical one. The Russo-Seymour-Welsh theory still applies below any given scale.

Having a (subsequential) scaling limit of the (near) critical percolation, we can ask how to describe it. The discrete models have finite σ-fields, but in the scaling limit the cardinality blows up, so we have to specify how we pass to a limit. First we remark that we restrict ourselves to the *crossing events* (which form the original physical motivation for the percolation model), disregarding events of the sort "number of open sites in a given region is even" or "number of open sites in a given region is bigger than the number of closed ones." While such events could be studied in another context, they are of little interest in the framework motivated by physics. Furthermore, we restrict ourselves to the *macroscopic* crossing events, i.e. crossings of quads of fixed size. Otherwise one could add to the σ-field crossing events for microscopic quads (of the lattice mesh scale or some intermediate scale), which have infinitesimal size in the scaling limit. The resulting construction would be an extension of ours, similar to a non-standard extension of the real numbers. But because of the locality of the percolation, different scales are independent, so it won't yield new non-trivial information (as our results in fact show). However, we would like to remark that in dependent models such extensions could provide new information. For example, Kenyon has calculated probabilities of local configurations appearing in domino tilings and loop erased random walks [Ken09], while such information is lost in describing the latter by an SLE curve.

Summing it up, we restrict ourselves only to the scaling limit of the *macroscopic crossing events* for percolation – it describes the principal physical properties (though adding microscopic events can be of interest). The question then arises, how to describe such a scaling limit, and several approaches were proposed. We list some of them along with short remarks:

(1). *Random coloring of a plane* into yellow and blue colors would be a logical candidate, since in the discrete setting we deal with random colorings of graphs. However, it is difficult to axiomatize the allowed class of "percolation-induced colorings," and connectivity properties are hard to keep track of. In particular, in the scaling limit almost every point does not belong to a cluster, but is rather surrounded by an infinite number of nested open and closed clusters, and so has no well-defined color.

(2). *Collection of open clusters* as random compact subsets of the plane would satisfy obvious pre-compactness (for Borel probability measures on collections of compacts endowed with some version of the Hausdorff distance), guaranteeing existence of the subsequential scaling limits. However, crossing probabilities cannot be extracted: as a toy model imagine two circles, O and C, intersecting at two points, x and y. Consider two different configurations: in the first one, $O \setminus \{x\}$ is an open crossing and $C \setminus \{y\}$ is a closed one; in the second one, $O \setminus \{y\}$ is an open crossing and $C \setminus \{x\}$ is a closed one. In both configurations the cluster structure is the same (with O being the open and C being the closed cluster), while some crossing events differ. Thus one has to add the state of the "pivotal" points (like x and y) to the description, making it more complicated. Constructing such a limit, proving its uniqueness, and studying it would require considerable technical work.

(3). *Aizenman's web* represents percolation configuration by a collection of all curves (= crossings) running inside open clusters. Here pre-compactness follows from the Aizenman-Burchard work [AB99], based on the Russo-Seymour-Welsh theory, and crossing events are easy to extract. Moreover, it turns out that the curves involved are almost surely Hölder-continuous, so this provides a nice geometric description of the connectivity structure. However, there is a lot of redundancy in taking *all* the curves inside clusters, and there are deficiencies similar to those of the previous approach.

(4). *Loop ensemble* — instead of clusters themselves, it is enough to look at the interfaces between open and closed clusters, which form a collection of nested loops. Such a limit was constructed by Camia and Newman in [CN06] from Schramm's SLE loops. However, it was not shown that such a limit is full (i.e. contains all the crossing events), and in general, extracting geometric information is far from being straight-forward.

(5). *Branching exploration tree* — there is a canonical way to explore all the discrete interfaces: we follow one, and when it makes a loop, branch into two created components. The continuum analogue was studied (in relation with the loop ensembles) by Sheffield in [She09], and this is perhaps the most natural approach for general random cluster models with dependence, see [KS10]. Construction is rather technical, however the scaling limit can be described by branching SLE curves and so is well suited for calculations.

(6). *Height function* — in the discrete setting one can randomly orient interfaces, constructing an integer-valued height function, which changes by ± 1 whenever an interface is crossed. This stores all the connectivity

information (at least in our setup of interfaces along the edges of trivalent graphs), but adds additional randomness of the height change. It is expected that after appropriate manipulations (e.g. coarse-graining and compactifying — i.e. projecting mod 1 to a unit interval) the height function would converge in the scaling limit to a random distribution, which is a version of the free field. This is the cornerstone of the "Coulomb gas" method of Nienhuis [Nie84], which led to many a prediction. It is not immediate how to connect this approach to the more geometric ones, retrieving the interfaces (= level curves) from the random distribution, but SLE theory suggests some possibilities. Also it is not clear whether we can implement this approach so that the extra randomness disappears in the limit, while the connectivity properties remain.

(7). *Correlation functions* — similarly to the usual Conformal Field Theory approach, instead of dealing with a random field as a random distribution, we can restrict our attention to its n-point correlations. For example, that would be the usual physics framework for fields arising in approach (6), but to implement it mathematically we would need to reconstruct height lines of a field, based on its correlations, which is perhaps more difficult than developing (6) by itself, and possibly requires some extra assumptions on the field. A more geometric approach would be to study for every finite collection of points $\{z_j^k\}_{j,k}$ the probability $F_r\left(z_j^k\right)_{j,k}$ that for every k all the balls $\left\{B(z_j^k, r)\right\}_j$ are connected by an open cluster. To obtain the correlation function, one takes the double limit, passing to the scaling limit first and sending $r \to 0$ afterwards. Since the probability of a radius r ball being touched by a macroscopic cluster (conjecturally) decays as $r^{5/48}$, one has to normalize accordingly, considering F as a $\prod_{j,k} |dz_j^k|^{5/48}$—form. The geometric approach seems more feasible, but along the way one would have to further develop a version of CFT corresponding to such "connectivity" correlations.

(8). *Collection of quads crossed* — in the discrete setting one can describe a percolation configuration by listing the finite collection of all discrete quads (i.e. topological quadrilaterals) having an open crossing. We propose a continuous analogue, described in detail below. The percolation configuration space thus becomes a space of quad collections, satisfying certain additional properties.

(9). *Continuous product of probability spaces*, or *noise* – this approach was advocated by Tsirelson [Tsi04a, Tsi04b, Tsi05]. In the discrete setting

the site percolation (on a set V of vertices) is given by the *product probability space* $(\Omega_V = \prod_V \{\text{open}, \text{closed}\}, \mathcal{F}_V, \mu_V)$. Note that when V is decomposed into two disjoint subsets, we obviously have

$$\mathcal{F}_V = \mathcal{F}_{V_1} \times \mathcal{F}_{V_2} . \tag{1.1}$$

Straight-forward generalization of the product space to the continuous case does not work: the product σ-field for the space $\prod_{\mathbb{C}} \{\text{open}, \text{closed}\}$ won't contain crossing events. Instead Tsirelson proposed to send the percolation measure space to a scaling limit, described by a *noise* or a *homogeneous continuous product of probability spaces*. Namely, for every smooth domain D the percolation scaling limit inside it would be given by a probability space $(\Omega_D, \mathcal{F}_D, \mu_D)$, which is translation invariant, continuous in D, and satisfies the property (1.1) for a smooth domain V cut into two smooth domains V_1, V_2 by a curve. There is a well-developed theory of such spaces, but establishing the latter property was problematic [Tsi05].

To most constructions one can also add the information about the closed clusters, though in principle it should be retrievable from the open ones by duality. This list is not exhaustive, but it shows the variety of possible descriptions of the same object. Indeed, in the discrete settings the approaches above contain all the information about macroscopic open clusters, and are expected to keep it in the scaling limit; on the other hand with an appropriate setup (one has to be careful, especially with (6)) we do not expect to pick up any extra information on the way. However, showing that most of these approaches lead to equivalent results, and in particular to isomorphic σ-fields, *is far from easy*.

We decided to proceed with the approach (8), since it follows the original physical motivation for the percolation model, and also provides us with enough pre-compactness to establish existence of subsequential scaling limits *before* we apply any percolation techniques.

It also serves well our second purpose: to provide escription of the percolation (subsequential) scaling limits, following Tsirelson's approach (9), and show that it is a *noise*, i.e. a homogeneous continuous (parameterized by an appropriate Boolean algebra of piecewise-smooth planar domains) product of probability spaces.

The question whether critical percolation scaling limit is a noise was posed by Tsirelson in [Tsi04a, Tsi04b, Tsi05], where he has also noted that such noise must be non-classical, or *black* (compared to the classical white noise). The theory of black noises was started by Tsirelson and Vershik in [TV98]. In particular, they constructed the first examples of zero and one-dimensional

black noises (and more of those were found since); we provide the first genuinely two-dimensional example.

To establish that percolation leads to noise it is enough to prove that its scaling limit, restricted to two adjacent rectangles, determines the scaling limit in their union. Following is a version of this result, perhaps more tangible (though stated slightly informally here). It is proved in Proposition 4.1, which is the most technical result of our paper. *Consider a rectangle, and a smooth path α cutting it. We show that for every $\epsilon > 0$ there is a finite number of percolation crossing events in quads disjoint from α, from which one can reliably predict whether or not the rectangle is crossed, in the sense that the probability for a mistake is less than ϵ.* The essential point is that the set of crossing events that one looks at *does not depend on the mesh of the lattice*, which allows us to deduce the needed result, thus showing that any subsequential scaling limit of the critical percolation is indeed a *black noise.*

1.2. *Percolation basics and notation.* For completeness of exposition we present most of the needed background below, interested readers can also consult the books [Kes82, Gri99, BR06, Wer09].

We start with a rather general setup, which in particular includes all planar site and bond models (as well as percolation on planar hypergraphs, cf. [BR10]). Since we will be interested in scaling limits, it is helpful from the very beginning to work with subsets of the plane, rather than graphs.

We consider a locally finite *tiling* H of the plane (or its subdomain) by topologically closed polygonal *tiles* P. Furthermore, we ask it to be trivalent, meaning that at most three tiles meet at every vertex and so any two tiles are either disjoint or share a few edges. A *percolation model* fixes a percolation probability $p(P) \in [0,1]$ for every tile P, and declares it open (or closed) with probability $p(P)$ (or $1-p(P)$) independently of the others. In principle, we can work with a random tiling H, with percolation probabilities $p(P,H)$ satisfying appropriate measurability conditions (so that crossing events are measurable).

More generally, we can consider a *random coloring* of tiles into open and closed ones, given by some measure μ on the space $\prod_{P\in H}\{\text{open, closed}\}$ with a product σ-field (which contains all events concerned with a finite number of tiles). Percolation models correspond to product measures μ.

We define open (closed) *clusters* as connected topologically closed subsets of the plane — components of connectivity of the union of open (closed) tiles taken with boundary.

All site and bond percolation models on planar graphs can be represented

in this way. Indeed, given a planar site percolation model, we replace sites by tiles, so that two of them share a side if and only if two corresponding sites are connected by a bond. It is easy to see that such a collection always exists (e.g. tiles being the faces of the dual graph). Each tile is declared open (or closed) with the same probability p (or $1-p$) as the corresponding site.

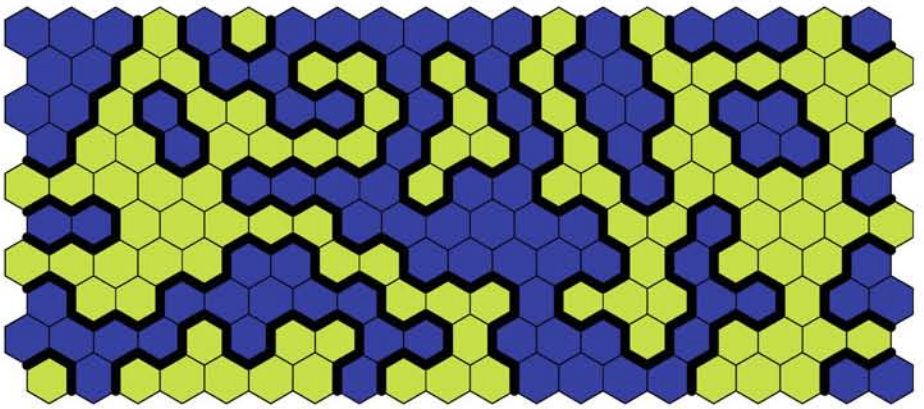


FIG 1.1. *Critical site percolation on the triangular lattice. Each site is represented by a hexagonal tile, which is open (blue) or closed (yellow) with probabilities* $1/2$ *independently of others. Interfaces between open and closed clusters are pictured in bold.*

If our original graph is a triangulation, then the dual one is trivalent and only three tiles can meet at a point. However, when our graph has a face with four or more sides, two non-adjacent sites can correspond to tiles sharing a vertex (but not an edge). To resolve the resulting ambiguity we insert a small tile P around every tiling vertex, where four or more tiles meet, and set it to be *always closed*, taking $p(P) = 0$. After the modification, at most three tiles can meet at a point, as required.

Bond percolation can also be represented by partitioning the plane into tiles. There will be a tile for every bond, site, and face of the original bond graph, chosen so that two different face-tiles or two different site-tiles are disjoint. Moreover, a bond-tile shares edges with two site-tiles and two face-tiles, corresponding to adjacent sites and faces. The bond-tiles are open (or closed) independently, with the same probability p (or $1-p$) as the corresponding bonds. The site-tiles P are then always open ($p(P) = 1$), while the face-tiles P are always closed ($p(P) = 0$). With such construction, no ambiguities arise, and at most three tiles can meet at a point. A possible construction for the bond percolation (with octagons being the bond-tiles) is illustrated in Figure 1.2.

Following the original motivation of Broadbent and Hammersley, we rep-

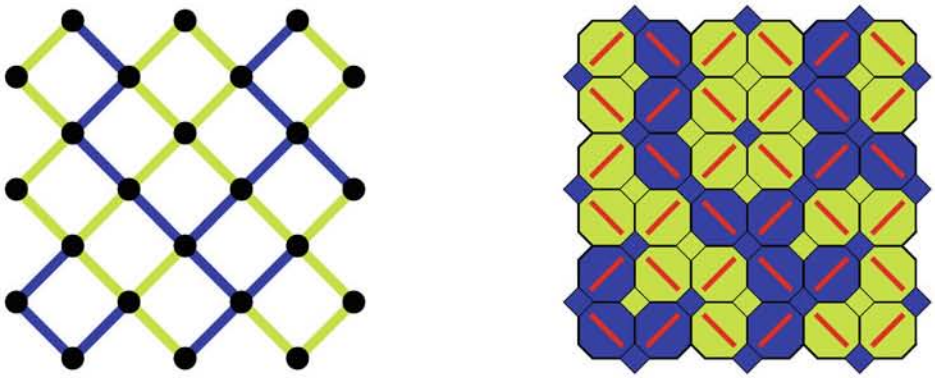


FIG 1.2. *Left: critical bond percolation on the square lattice, every bond is open (blue) or closed (yellow) with probabilities* 1/2 *independently of others. Right: representation by critical percolation on the bathroom tiling, with direction of original bonds marked on the corresponding octagonal tiles. Rhombic tiles alternate in color, with blue (always open) corresponding to the sites of the original square lattice, while yellow (always closed) — to the faces. Octagonal tiles correspond to the bonds, and are open or closed independently, with probabilities* 1/2.

resent holes in some random porous medium by open tiles. A liquid is assumed to move freely through the holes, thus it can flow between two regions if there is a path of open tiles between them. More precisely, we say that there is an *open (closed) crossing* between two sets K_1 and K_2 inside the set U, if for the percolation configuration restricted to U, there is an open (closed) cluster intersecting both K_1 and K_2. We will be mostly interested in crossings of quads (= topological quadrilaterals).

Note the following duality property: *there is an open crossing between two opposite sides of a quad if and only if there is no dual (i.e. closed) crossing between two other sides.* Exchanging the notions of open and closed, we conclude that, on a given tiling, percolation models with probabilities $p(P)$ and $p^*(P) = 1 - p(P)$ are dual to each other. In particular, the model with $p(P) \equiv 1/2$ is self-dual, which can often be used to show that it is critical, like for the hexagonal lattice. The bathroom-tile model of the square lattice bond percolation in Figure 1.2 is dual to itself translated by one lattice step (so that always open rhombi are shifted to always closed ones), which also leads to its criticality, [Kes80]. There are however dualities (e.g. arising from the star-triangle transformations) with more complicated relations between p and p^*, and so there are critical models with p away from 1/2, see e.g. [Wie81, BR10].

Though we have in mind critical models, we do not use criticality in

the proof, rather some crossing estimates, which are a part of the Russo-Seymour-Welsh theory. Those also hold (below any given scale) for the near-critical models, like those discussed in [Nol08]. In a sense the estimates we need should hold whenever the crossing probabilities do not tend to a trivial limit, see Remark 1.3.

We work with a collection of percolation models μ_η (and sometimes more general random colorings) on tilings H_η, indexed by a set $\{\eta\}$ (e.g. square lattices of different mesh), and denote by μ_η also the corresponding measure on percolation configurations (= random coloring of tiles). By $|\eta| \in \mathbb{R}_+$ we will denote a "scale parameter" of the model μ_η. We assume that all tiles have diameter at most $|\eta|$ (for a percolation on random tiling, our proof would work under the relaxed assumption that for any positive r, the probability to find in a bounded region a tile of diameter bigger than r tends to zero as $|\eta| \to 0$). This parameter also appears in the Russo-Seymour-Welsh—type estimates below, which have to hold on scales larger than $|\eta|$. Note that in most models $|\eta|$ can be taken to be the supremum of tile diameters (or the lattice mesh for the percolation on a lattice).

We will be interested in scaling limits as the scale parameter tends to zero, looking at sequences of percolation models with $|\eta| \to 0$, which do not degenerate in the limit. We will show in Remark 1.17, that in our setting working with sequences and nets yields the same results. As discussed above, all critical percolation models are conjectured to have a universal and conformally invariant scaling limit. Moreover, its SLE description allows to calculate exactly many scaling exponents and dimensions and we will use this conjectured picture in informal discussions below. However, to make our paper applicable to a wider class of models, we will work under much weaker assumptions.

The famous Russo-Seymour-Welsh theory provides universal bounds for crossing probabilities of the same shape, superimposed over lattices with different mesh. Originally established for the critical bond percolation on the square lattice, it has been since generalized to many critical (and near-critical, when the percolation probabilities tend to their critical values at appropriate speeds as the mesh tends to zero) models with several different approaches, see [Kes82, Gri99, BR06, Wer09]. We will use one of the recurrent techniques: conditioning on the lowest crossing. If a quad is crossed horizontally, one can select the lowest possible crossing (the closest to the bottom side), which will then depend only on the configuration below. If we condition on the lowest crossing, the configuration above it is unbiased, allowing for easy estimates of events there. Besides using this neat trick, we will assume certain bounds on the crossing probabilities for the annuli.

A prominent role is played by the so-called k—*arm* crossing events in annuli, since they control dimensions of several important sets. In particular, for a given percolation model μ_η we denote by $P^1_\eta(z,r,R)$ the probability of a *one-arm* event, that *there is an open crossing connecting two opposite circles of the annulus* $A(z,r,R)$. Then this is roughly the probability that the disc $B(z,r)$ intersects an open cluster of size $\approx R$, which is expected for $|\eta| < r$ to have a power law like

$$P^1_\eta(z,r,R) \;=\; (r/R)^{a_1+o(1)} \;, \quad \text{as} \;\; r/R \to 0 \;,$$

morally meaning (modulo some correlation estimates) that percolation clusters in the scaling limit have dimension $2 - a_1$.

Similarly, denote by $P^4_\eta(z,r,R)$ the probability of a *four-arm* event, that *there are alternating open—closed—open—closed crossings connecting two opposite circles of the annulus* $A(z,r,R)$. Then this is roughly the probability that changing the percolation configuration on the disc $B(z,r)$ changes the crossing events on the scale $\approx R$. Indeed, if all the tiles in $B(z,r)$ are made open, that connects two open arms; while making them closed connects two closed arms, destroying the open connection. For $|\eta| < r$ a power law

$$P^4_\eta(z,r,R) \;=\; (r/R)^{a_4+o(1)} \;, \quad \text{as} \;\; r/R \to 0 \;,$$

is expected, roughly meaning that the *pivotal* tiles (such that altering their state changes crossing events on large scale) have dimension $2 - a_4$ in the scaling limit.

Whenever the RSW theory applies, the probabilities for k arm events have power law bounds from above and below as $r \to 0$. In fact, the probabilities are conjectured to satisfy universal (independent of a particular percolation model) power laws with rational powers, predicted by the Conformal Field Theory, see discussion in [SW01]. Those predictions have been so far proved for the critical site percolation on triangular lattice only [SW01, LSW02], giving $a_1 = 5/48$ and $a_4 = 5/4$.

Other k-arm probabilities are also of interest, but estimating just the mentioned two allows to apply our methods. Roughly speaking, we need to know that the percolation clusters have dimension smaller than 2 and that pivotal points have dimension smaller than 1, or, in other words, that $a_1 > 0$ and $a_4 > 1$. Note that we do not actually need the power laws, but rather weaker versions of the corresponding upper estimates.

Namely, we need the following two estimates to hold *uniformly* for percolation models under consideration:

ASSUMPTIONS 1.1. *There exist positive functions $\Delta_1(r,R)$ and $\Delta_4(r,R)$, such that*

$$\lim_{r\to 0} \Delta_j(r,R) = 0 \text{ for any fixed } R < R_0 \ ,$$

and the following estimates hold whenever $z \in \mathbb{C}$ and $0 < |\eta| < r < R < R_0$. The probability of one open arm event and the probability of a similar event for one closed arm satisfy

$$P^1_\eta(z,r,R) \ \le \ \Delta_1(r,R) \ , \tag{1.2}$$

while the probability of a four arm event satisfies

$$P^4_\eta(z,r,R) \ \le \ \left(\frac{r}{R}\right) \cdot \Delta_4(r,R). \tag{1.3}$$

For $r \ge R$, when the annulus is empty, we set $\Delta_j(r,R) := 1$, so that the function is defined for all positive arguments. Without loss of generality we can assume that functions $\Delta_j(r,R)$ are increasing in r and decreasing in R.

REMARK 1.2 (A STRONGER RSW ESTIMATE). *For most percolation models where the RSW techniques work, their application would prove a stronger estimate, with $\Delta_1(r,R)$ replaced in (1.2) by $C(r/R)^{a_1}$ with $a_1 > 0$. However we can imagine situations where only our weaker version can a priori be established.*

REMARK 1.3 (RSW ESTIMATES AND SCALING LIMITS). *For most percolation models where the RSW techniques work, the estimate (1.2) would be equivalent to saying that crossing probability for any given quad (or for all quads) is uniformly bounded away from 0 and 1. So morally our noise characterization would apply whenever one can speak of non-trivial subsequential scaling limits of crossing probabilities.*

The first Assumption (1.2) is used several times throughout the paper. In a stronger form it was one of the original Russo-Seymour-Welsh results, and has since been shown to hold for a wide range of percolation models, see [BR06, BR10, Gri99, Kes82]. For site percolation on the triangular lattice, one can take $\Delta_1(r,R) \approx (r/R)^{5/48}$, see [LSW02].

The second Assumption (1.3) is used only once, namely in the proof of Theorem 1.5. It is known (in a sharp form with $\Delta_4(r,R) \approx (r/R)^{1/4}$) for site percolation on the triangular lattice, see [SW01]. Christophe Garban kindly allowed us to include as Appendix B his proof for bond percolation on the square lattice, which is partly inspired by the noise-sensitivity arguments in

[BKS99]. It seems possible to extend the proof to a wider range of percolation models (by changing the values of the C_j random variables so they become centered, rather than using symmetry arguments).

QUESTION 1.4 (ESTIMATING PIVOTALS). *For which models can* (1.3) *be proved? Can it be deduced from* (1.2)*? Is there a geometric argument?*

Summing up the discussion above, we study collections of percolation models satisfying Assumptions 1.1. In particular, *our results apply to critical and near-critical site percolation on the triangular lattice and bond percolation on the square lattice.*

1.3. *Definition of the scaling limits.* Every discrete percolation configuration (or a coloring of a tiling) is clearly described by the collection of quads (= topological quadrilaterals) which it crosses (and for a fixed lattice mesh in a bounded domain a finite collection of quads would suffice). We introduce a setup, which allows to work independently of the lattice mesh and pass to the scaling limit, but the idea remains the same: a percolation configuration is encoded by a collection of quads — those which are crossed by clusters. Our setup is inspired by the Dedekind's sections and uses the metric and ordering on the space of quads. The ordering comes from the obvious observation that crossing of a quad automatically contains crossings of "shorter" sub-quads, and so possible percolation configurations (= quad collections) should satisfy certain monotonicity properties.

We will work in a domain $D \subset \mathbb{C} = \mathbb{R}^2$. A *quad* in D is a topological quadrilateral, i.e. a homeomorphism $Q : [0,1]^2 \to Q([0,1]^2) \subset D$. We introduce some redundancy when considering quads as parameterized quadrilaterals, but we avoid certain technicalities. The space of all such quads will be denoted by $\mathcal{Q} = \mathcal{Q}_D$. It is a metric space under the uniform metric

$$d(Q_1, Q_2) = \sup_{z \in [0,1]^2} |Q_1(z) - Q_2(z)|.$$

A *crossing* of Q is a connected compact subset of $[Q] := Q([0,1]^2)$ that intersects both opposite sides $\partial_0 Q := Q(\{0\} \times [0,1])$ and $\partial_2 Q := Q(\{1\} \times [0,1])$. We also denote the remaining two sides by $\partial_1 Q := Q([0,1] \times \{0\})$ and $\partial_3 Q := Q([0,1] \times \{1\})$, see Figure 1.3. The whole boundary of Q we denote by ∂Q. In the discrete setting there is no difference between connected and path-connected crossings, but one can imagine lattice models where the probabilities of the two would be different in the scaling limit. However, for the models under consideration, every crossing in the scaling limit almost

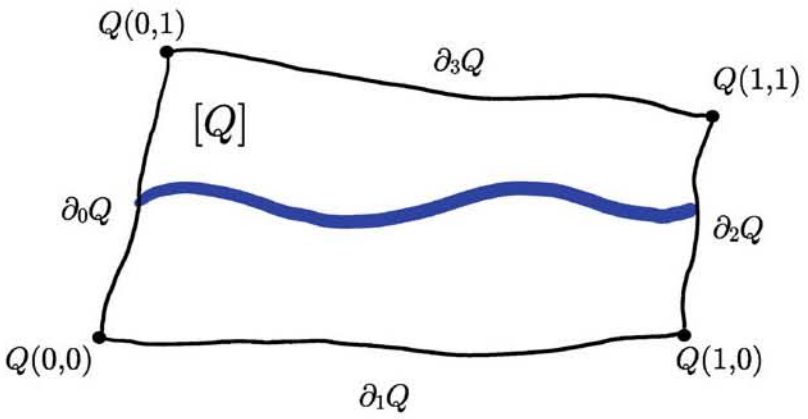


FIG 1.3. *A quad with a crossing.*

surely can be realized by a Hölder continuous curve [AB99], so the two notions are essentially the same.

In addition to the structure of $\mathcal{Q}_D$ as a metric space, we will also use the following partial order on $\mathcal{Q}_D$. If $Q_1, Q_2 \in \mathcal{Q}_D$, we write $Q_1 \leq Q_2$ if every crossing of Q_2 contains a crossing of Q_1. The simplest example is seen in Figure 1.4. Also, we write $Q_1 < Q_2$ if there are open (in the uniform metric) neighborhoods U_1 of Q_1 and U_2 of Q_2 in $\mathcal{Q}_D$, such that for every $Q \in U_1$ and $Q' \in U_2$ we have $Q \leq Q'$. Thus, the set of pairs $(Q, Q') \in \mathcal{Q}^2$ satisfying $Q < Q'$ is the interior of $\{(Q, Q') \in \mathcal{Q}^2 : Q \leq Q'\}$ in the product topology.

A subset $S \subset \mathcal{Q}_D$ is a *lower set* if, whenever $Q \in S$ and $Q' \in \mathcal{Q}_D$ satisfies $Q' < Q$, we also have $Q' \in S$. The collection of all closed lower subsets of $\mathcal{Q}_D$ will be denoted $\mathcal{H}_D$. Any discrete percolation configuration ω in $\mathbb{C}$ is naturally associated with an element S_ω of $\mathcal{H}_D$: the set of all quads for which ω contains a crossing formed by an open cluster. Thus a percolation model on the tiling H_η induces a probability measure μ_η on $\mathcal{H}_\mathbb{C}$ (and more generally on $\mathcal{H}_D$ for any domain D contained in its domain of definition).

A topology is defined on $\mathcal{H}_D$ by specifying a subbase. If $U \subset \mathcal{Q}_D$ is topologically open and $Q \in \mathcal{Q}_D$, let

$$\begin{aligned} V_U &:= \{S \in \mathcal{H}_D : S \cap U \neq \varnothing\}, \text{ and} \\ V^Q &:= \{S \in \mathcal{H}_D : Q \notin S\}. \end{aligned}$$

If we regard S as a collection of crossed quads, then

$$\begin{aligned} V_U &:= \{S : \text{ some quads from } U \text{ are crossed}\}, \text{ and} \\ V^Q &:= \{S : \text{ quad } Q \text{ is not crossed}\}. \end{aligned}$$

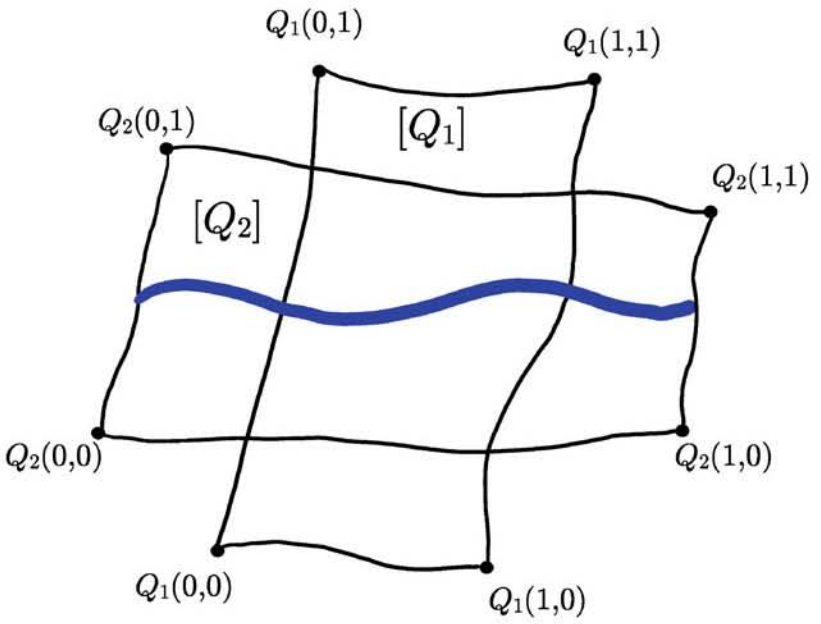


FIG 1.4. *Any crossing of Q_2 contains a crossing of Q_1, so we write $Q_1 \leq Q_2$.*

Our topology of choice on $\mathcal{H}_D$ will be the minimal topology $\mathcal{T}_D$ which contains every such V_U and V^Q. As we will see, $(\mathcal{H}_D, \mathcal{T}_D)$ is a metrizable compact Hausdorff space. The percolation scaling limit configuration will be defined as an element of $\mathcal{H}_{\mathbb{C}}$, and the scaling limit measure is a Borel probability measure on $\mathcal{H}_{\mathbb{C}}$.

Then the event of a percolation configuration crossing a quad Q corresponds to

$$\boxminus_Q \;:=\; \neg V^Q \;=\; \{S \in \mathcal{H} :\; Q \in S\} \;\subset\; \mathcal{H}.$$

We will prove that if $\mathcal{Q}^0 \subset \mathcal{Q}$ is dense in $\mathcal{Q}$, then the events $\boxminus_Q$ for $Q \in \mathcal{Q}^0$, generate the Borel σ-field of $\mathcal{H}$.

1.4. *Statement of main results.* Construction of the (subsequential) scaling limits is accomplished for all random colorings of trivalent tilings without additional assumptions. Other results of this paper apply to all *percolation models on trivalent tilings, satisfying Assumptions 1.1*, i.e.the RSW estimate (1.2) and the pivotal estimate (1.3).

As discussed above, we fix a collection of such percolation models indexed by a set $\{\eta\}$, and denote by μ_η the (locally) discrete probability measure on percolation configurations for the model corresponding to parameter η. By $|\eta| \in \mathbb{R}_+$ we denote the scale parameter of the model (which enters in the

RSW estimates). Later we will show that μ_η has subsequential scaling limits as $|\eta| \to 0$, and will work with one of those, denoted by μ_0.

First we present a simplified discrete version of the gluing theorem. Informally, the statement is that if you fix a quad Q_0 and a finite length path α cutting it into two pieces, then the discrete percolation configuration outside a small neighborhood of α reliably predicts the existence or non-existence of a crossing of Q_0 with probability close to one, uniformly in the mesh size.

Following is the precise version, estimating the random variable which is the conditional probability of observing an open crossing given the configuration s-away from α. As a regularity assumption, we need a constant $C(\alpha)$ such that for every $s > 0$ the set α can be covered by at most $C(\alpha)/s$ balls of radius s. This is the case if and only if α has finite 1-dimensional upper Minkowski content $\mathcal{M}^{*1}(\alpha) = \limsup_{s\to 0+} (\text{area}\,\{z : \text{dist}\,(z,\alpha) < s\})$.

THEOREM 1.5 (DISCRETE GLUING). *Consider a collection of percolation models satisfying Assumptions 1.1. Let $Q_0 \in \mathcal{Q}$ be some quad, and let $\alpha \subset [Q_0]$ be a finite union of finite length paths, or more generally a set of finite 1-dimensional upper Minkowski content. Let $\boxminus_{Q_0}$ denote the event that Q_0 is crossed by ω, and for each $s > 0$ let $\mathcal{F}_s$ denote the σ-field generated by the restriction of ω to the complement of the s-neighborhood of α. Then for every $\epsilon > 0$,*

$$\lim_{s\searrow 0} \sup_{|\eta|\in(0,s)} \mu_\eta\Big(\epsilon < \mu_\eta(\boxminus_{Q_0} \mid \mathcal{F}_s) < 1-\epsilon\Big) = 0\,.$$

REMARK 1.6 (GLUING IS NON-CONSTRUCTIVE). *Theorem 1.5 essentially states that event $\boxminus_{Q_0}$* for any given η *can be reconstructed with high accuracy by sampling crossing events away from α. This does not lead to the noise characterization of the scaling limit, since a priori the number of events we need to sample can grow out of control as $|\eta| \to 0$. We address this by providing a uniform bound on the number of events to be sampled in Proposition 4.1. However, we do not give a recipe, and in general the gluing procedure may still depend on η. We deal with this possibility by assuming that there is a (subsequential) scaling limit, then, since the number of possible outcomes of sampling a finite number of crossing events is finite, we can choose a sub-subsequence along which the same procedure is used. The downside is that, in the current form, our results alone cannot be used to show the uniqueness of the percolation scaling limit.*

The previous Remark leads to the following

QUESTION 1.7 (CONSTRUCTIVE GLUING). *Can one show that there is an independent of η and preferably constructive procedure to predict $\boxminus_{Q_0}$ by sampling crossing events away from α?*

A positive answer would provide an elegant proof of the uniqueness of the scaling limit (provided Cardy's formula is known), make the universality phenomenon more accessible, and lay foundation for the renormalization theory approach to percolation, similar to one discussed by Langlands, see [Lan05] and the references therein. A particularly optimistic scenario is suggested in Conjecture 1.11.

Theorem 1.5 is a toy version of our main gluing results, and the logic behind it was folklore for a long time (but strangely enough appeared few times in print, as was pointed out in [Ste09]). The proof proceeds by consequently resampling radius $2s$ balls covering the s-neighborhood of α and using the pivotal estimate (1.3). For this to work, the bounds have to sum up to something small, and so with better pivotal estimates we can relax assumptions on α (and the number of balls needed to cover it).

REMARK 1.8 (REGULARITY ASSUMPTIONS ON α). *When applying the pivotal estimate (1.3), we cannot relax the regularity assumption that α has finite 1-dimensional upper Minkowski content. However, as discussed above, most percolation models are expected to satisfy stronger estimates, and then the regularity assumption can be considerably weakened (though it cannot be dropped). The set of pivotal points conjecturally has Hausdorff dimension $3/4$ in the scaling limit (proved on the triangular lattice in [SW01]), so if the Hausdorff dimension of the set α is greater than $5/4$, the two should intersect with positive probability (needed correlation estimates are expected to hold). In such a situation, there is little hope to reliably determine the crossing without knowledge of α. To be more rigorous, instead of pivotals one has to work with the percolation spectrum, but it has the same dimension (see [GPS10a] for the definition and discussion), so α has to be of dimension at most $5/4$. On the other hand, our proof works for sets α of dimension smaller than 2 minus dimension of the pivotal points, provided also that $\alpha \cap \partial Q_0$ is finite. So $5/4$ seems to be the correct dimensional threshold, and this can be established for the critical site percolation on the triangular lattice.*

REMARK 1.9 (CROSSINGS WITH FINITE INTERSECTION NUMBER). *Suppose that a smooth curve α cuts the quad Q_0 into two quads Q_+ and Q_-. The following informal discussion concerns the critical percolation scaling*

limit, assuming its existence and conformal invariance (but it can be easily made more precise by working with the limits of discrete configurations and events). First observe, that open clusters inside $Q_\pm$ intersect the smooth boundary α on a set of Hausdorff dimension $2/3$, (and we have a control for the correlation of points inside it). Being independent, two large open clusters on opposite sides of α have a positive probability of touching. But if they touch at some point $x \in \alpha$, we cannot have a large closed cluster passing through x and separating them. Indeed, that would mean a four-arm event occurring at the pivotal *point x on the smooth curve α, and since $a_4 = 5/4$, the set of pivotal points has Hausdorff dimension $3/4$ and so almost surely does not intersect the smooth α. Thus the two open clusters in $Q_\pm$ which touch on α are in fact parts of the same large open cluster in Q_0. We conclude, that given a smooth curve α, with positive probability there is a crossing of Q_0 which intersects α only* once.

The last remark says that for a smooth curve α, if we condition on the quad being crossed, the probability of finding a crossing intersecting α only finitely many times is positive. We can ask whether this conditional probability is one, leading to the following:

CONJECTURE 1.10 (FINITE INTERSECTION PROPERTY). *Let Q be a quad and α a smooth curve. Then in the scaling limit, given existence of a crossing, there is almost surely a crossing intersecting α only a finite number of times.*

Alternatively, one can formulate a discrete version, which is a priori stronger (for an equivalent one we would have to count small scale packets of intersections):

CONJECTURE 1.11 (TIGHTNESS OF INTERSECTIONS). *Let Q be a quad and α a smooth curve. For a discrete percolation measure μ_η let the random variable $N_\eta = N_\eta(Q, \alpha)$ be the minimal possible number of intersections with α for a crossing of Q (and 0 if Q is not crossed). Then N_η is tight as $|\eta| \to 0$.*

Proving either of these conjectures would not only simplify our proof of the main theorem, but also provide an easy and constructive "gluing procedure": we take all "excursions" (= crossings in the complements of α) and check whether a crossing of Q can be assembled from finitely many of them (as discussed in Remark 1.9, the finite number of excursions will necessarily connect into one crossing). Such a procedure would give a very short proof

of the uniqueness of the full percolation scaling limit, provide an elegant setup for the renormalization theory of percolation crossings, and show that the "finite model" discussed by Langlands and Lafortune in [LL94] indeed approximates the critical percolation. Note that if Conjectures above fail, the constructive procedure for gluing crossings will have to deal with countably many excursions intersecting α on a Cantor set C. Then the criterion for possibility of gluing the excursions into one crossing is likely to be in terms of some capacitary (or dimensional) characteristics of C.

The critical percolation scaling limit is conjectured to be conformally invariant, which was established for site percolation on the triangular lattice [Smi01a, Smi01b]. In Remark 1.9 we mentioned that, for a given quad Q_0, a smooth curve α has the property that two large clusters inside two components of $[Q_0] \setminus \alpha$ touch each other (on α) with positive probability. One can ask, whether an arbitrary curve α would always enjoy this property, and the answer is negative. Moreover, under conformal invariance assumption, this property (equivalent to positive probability of finding a crossing intersecting α only finitely many times) can be reformulated in terms of multifractal properties of harmonic measure on two sides of α. Below we present an informal argument to this effect, note also similar discussions in the SLE context in [GRS08].

To simplify the matters, we discuss a related (and probably equivalent) property that the expected number of such touching points for two large clusters inside two components of $[Q_0] \setminus \alpha$ is positive (or even infinite). Let $f(a_+, a_-)$ be the two-sided dimension spectrum of harmonic measure on α, see [APS10]. Roughly speaking, $f(a_+, a_-)$ is the dimension of the set of $z \in \alpha$ where harmonic measures $\omega_\pm$ on two sides of α have power laws $a_\pm$:

$$\omega_+ \left(B\left(z,r\right)\right) \;\approx\; r^{a_+} \;, \quad \omega_- \left(B\left(z,r\right)\right) \;\approx\; r^{a_-} \;.$$

Consider a set $E \subset \alpha$ such that harmonic measures $\omega_\pm$ have power laws $a_\pm$ on E. Cover most of E by a collection of balls B_j of small radii r_j. By conformal invariance, the probability of a ball B_j to be touched by a large open cluster on the "+ side" of α is governed by its harmonic measure and Cardy's formula [Smi01b] that gives the exponent $1/3$:

$$P_+(B_j) \;\approx\; \omega_+(B_j)^{1/3} \;\approx\; r_j^{a_+/3} \;.$$

Combining with the similar estimate on the "− side," we write an estimate for the expected number of balls B_j touched by large clusters on both sides:

$$\sum_j P_+(B_j) P_-(B_j) \;\approx\; \sum_j r_j^{(a_+ + a_-)/3} \;.$$

Here the right hand side is small (large) when $(a_+ + a_-)/3$ is bigger (smaller) than $\dim(E)$, since $\sum_j r_j^{\dim(E)} \approx 1$. We conclude that the expected number of points touched by large clusters on both sides is positive if and only if we can find E such that $(a_+ + a_-)/3 < \dim(E)$.

REMARK 1.12 (SHARPNESS OF THE FINITE INTERSECTION PROPERTY). *It seems that positive probability of having a crossing intersecting α a finite number of times is roughly equivalent to the existence of positive exponents a_+ and a_- such that the two-sided multifractal spectrum of harmonic measure on α satisfies $3 f(a_+, a_-) > a_+ + a_-$, see [APS10] for discussion of such spectra. The latter property seems to fail even for some non-smooth α of dimension close to one, so we do not expect the factorization property to be equivalent to the finite intersection property, but rather to be a weaker one.*

Before proving the main theorem, we discuss a new framework for the percolation scaling limit and show that collection of discrete random colorings is precompact, so in this framework the subsequential scaling limits exist for any random model.

Recall that we work with $\mathcal{H}_D$, which is the collection of all closed lower (i.e. monotonicity of the crossing events is obeyed) subsets of $\mathcal{Q}_D$, which one should think of as sets of quads crossed. The topology $\mathcal{T}_D$ is the minimal one containing all sets V_U (of S such that some quads in U are crossed) and V^Q (of S such that Q is not crossed).

We now list some important properties of $\mathcal{H}_D$.

THEOREM 1.13 (THE SPACE OF PERCOLATION CONFIGURATIONS). *Let $D \subset \hat{\mathbb{C}}$ be topologically open and nonempty.*

1. *$(\mathcal{H}_D, \mathcal{T}_D)$ is a compact metrizable Hausdorff space.*
2. *Let A be a dense subset of $\mathcal{Q}_D$. If $S_1, S_2 \in \mathcal{H}_D$ satisfy $S_1 \cap A = S_2 \cap A$, then $S_1 = S_2$. Moreover, the σ-field generated by V^Q, $Q \in A$ is the Borel σ-field of $(\mathcal{H}_D, \mathcal{T}_D)$.*
3. *If $S_1, S_2, \ldots$ is a sequence in $\mathcal{H}_D$ and $S \in \mathcal{H}_D$, then $S_j \to S$ in $\mathcal{T}_D$ is equivalent to $S = \limsup_j S_j = \liminf_j S_j$.*
4. *Let $D' \subset D$ be a subdomain, and for $S \in \mathcal{H}_D$ let $S' := S \cap \mathcal{Q}_{D'}$. Then $S \mapsto S'$ is a continuous map from $\mathcal{H}_D$ to $\mathcal{H}_{D'}$.*

In the above, $\liminf_j S_j$ is the set of $Q \in \mathcal{Q}_D$ such that every neighborhood of Q intersects all but finitely many of the sets S_j, and $\limsup_j S_j$ is the set of $Q \in \mathcal{Q}_D$ such that every neighborhood of Q intersects infinitely many of the sets S_j.

REMARK 1.14 (APPLICABILITY OF THEOREM 1.13). *The result concerns the topological properties of the space of percolation configurations (colorings of tilings), and involves no probability measures. Thus it can be applied to* any random coloring *of a trivalent tiling (into open and closed tiles), yielding results for dependent models, like the Fortuin-Kasteleyn percolation.*

It follows from the Theorem above that the space of continuous functions on $(\mathcal{H}_D, \mathcal{T}_D)$ is separable, and hence the unit ball in its dual space $M(\mathcal{H}_D)$ of Borel measures is compact (by the Banach-Alaoglu theorem, 3.15 in [Rud73]), metrizable (by 3.16 in [Rud73]), and obviously Hausdorff in the weak-$*$ topology. The space of probability measures being a closed convex subset of the unit ball, we arrive at the following

COROLLARY 1.15 (THE SPACE OF PERCOLATION MEASURES). *The space* $\mathrm{Prob}(\mathcal{H}_D)$ *of Borel probability measures on* $(\mathcal{H}_D, \mathcal{T}_D)$ *with weak-$*$ topology is a compact metrizable Hausdorff space.*

Discrete percolation measures μ_η clearly belong to $\mathrm{Prob}(\mathcal{H}_D)$ and hence the existence of percolation subsequential scaling limits (or, more generally, for any random colorings) is an immediate corollary:

COROLLARY 1.16 (PRECOMPACTNESS). *The collection of all discrete measures* $\{\mu_\eta\}$ *(corresponding to random colorings of trivalent tilings) is precompact in the topology of weak-$*$ convergence on* $\mathrm{Prob}(\mathcal{H}_D)$.

REMARK 1.17 (NETS VS. SEQUENCES). *We can make* $\{\eta\}$ *a directed set by writing* $\eta_1 \preceq \eta_2$ *whenever* $|\eta_1| \geq |\eta_2|$. *Then the scaling limit* $\lim_{|\eta|\to 0} \mu_\eta$ *is just the limit of the net* μ_η. *Since the target space* $\mathrm{Prob}(\mathcal{H}_D)$ *is metrizable and hence first-countable, it is enough to work with the notion of convergence along sequences, rather than nets. In particular, if* $\lim_{j\to\infty} \mu_{\eta_j} = \mu_0$ *for every sequence* η_j *with* $\lim_{j\to\infty} |\eta_j| = 0$, *then* $\lim_{|\eta|\to 0} \mu_\eta = \mu_0$.

REMARK 1.18. *Our construction keeps information about all macroscopic percolation events, and so gives the* full percolation scaling limit *at or near criticality. For the critical site percolation on triangular lattice, Garban, Pete and Schramm explain in section 2 of [GPS10c] that the scaling limit is unique, appealing to the work [CN06] of Camia and Newman. Basically, we know the probabilities of individual crossing events, and it allows*

to reconstruct the whole picture, albeit in a difficult way. It would be interesting to establish the uniqueness using an appropriate modification of Proposition 4.1.

So that our theorems apply even to models where the uniqueness of the scaling limit has not been proved yet, we work with one of the subsequential limits μ_0 — a Borel probability measure on $\mathcal{H}_{\mathbb{C}}$, provided by Corollary 1.16. By Theorem 1.13, the Borel σ-field of $\mathcal{H}_D$ is $\sigma(\boxminus_Q : Q \in \mathcal{Q}_D)$, i.e. is generated by the crossing events inside D. Let $\mathcal{F}_D := \sigma(\boxminus_Q : Q \in \mathcal{Q}_D)$ also denote the corresponding subfield of the Borel σ-field of $\mathcal{H}_{\mathbb{C}}$.

THEOREM 1.19 (FACTORIZATION). *Consider a collection of percolation models, satisfying Assumptions 1.1. Let D be a domain, and let $\alpha \subset \mathbb{C}$ be a finite union of finite length paths with finitely many double points. Denote the components of $D \setminus \alpha$ by D_j. Then for any subsequential scaling limit*

$$\mathcal{F}_D \;=\; \mathcal{F}_{D\setminus\alpha} \;=\; \vee_j \mathcal{F}_{D_j}\,, \tag{1.4}$$

up to sets of measure zero.

Again, we note that this does not hold when α is a sufficiently wild path.

Fix a subsequential scaling limit of the critical percolation for the bond model on the square lattice or for the site model on the triangular lattice. By the results above for a domain D it is described by a probability space $(\mathcal{H}_D, \mathcal{F}_D, \mu_D)$, which is invariant under translations of the plane (in fact, under all conformal transformations, but it is so far established for the triangular lattice only), depends continuously on D, and satisfies (1.4). Then in the language of Tsirelson (see [Tsi04a], Definition 3d1) it is a *noise* or a *homogeneous continuous product of probability spaces* with a Boolean base given by an appropriate algebra of piecewise-smooth planar domains (e.g. generated by rectangles).

There is a well developed and beautiful theory of noises, and we refer the reader to two expositions [Tsi04a, Tsi04b] by Tsirelson. In particular, every noise can be decomposed into a "classical," or "stable" part and a "non-classical," or "sensitive" part. White noise is purely classical, and purely non-classical noises are called "black" by Tsirelson. The black noises are difficult to construct, with the first example given by Tsirelson and Vershik in [TV98]. As pointed out by Tsirelson in [Tsi05] and in Remark 8a2 of [Tsi04b], percolation noise has to be non-classical, and so our paper provides the first genuinely two-dimensional example of a black noise:

COROLLARY 1.20 (PERCOLATION IS A NOISE). *Thus we conclude that, in the language of Tsirelson [Tsi04a], any subsequential scaling limit as above is a* noise *with a Boolean base given by an appropriate algebra of piecewise-smooth planar domains (e.g. generated by rectangles). Therefore, it has to be a* black noise, *as explained in [Tsi04b], Remark 8a2.*

By the conjectured universality, all the critical percolation models should have the same scaling limit and so correspond to the same black noise. The situation with near-critical noises is less clear, so the following question was proposed by Christophe Garban (the answer is expected to be negative):

QUESTION 1.21. *Are the noises arising from the critical and near-critical models isomorphic, or do we get a family of different noises?*

The notion of isomorphism for noises is discussed by Tsirelson in Section 4a of [Tsi04a]. Roughly speaking, we ask *whether one can get a near-critical noise from the critical one by a local deterministic procedure?* Note that the current construction of the near-critical percolation scaling limit would use the critical percolation along with some extra randomness, as in [CFN06, GPS10b].

Our construction of the scaling limit applies to general random colorings, and much of what we do afterwards mostly uses "arms estimates," which are known for a wider class of models, than just percolation. Therefore it is logical to ask

QUESTION 1.22. *To what extent out results apply to the models with dependence? One has to be more careful in formulating the results, since now the configuration inside a domain depends on the boundary conditions, but this can be addressed by working with a full-plane model and stripe domains.*

Acknowledgments. We are thankful to Christophe Garban and Wendelin Werner for useful conversations. The second author is grateful to Hugo Duminil-Copin, Clément Hongler, Gábor Pete, Jeff Steif, the referee, and especially Christophe Garban and Boris Tsirelson for their comments on the manuscript.

2. The Discrete Gluing Theorem.

In this Section, we prove Theorem 1.5.

Applying several times in succession Lemma A.1, we can find a smooth (i.e. given by a diffeomorphism) quad Q, such that the crossing events for Q_0 and Q are arbitrarily close. By the same Lemma, as $\varepsilon \to 0$, the crossing event for Q is well approximated by the crossing event of its perturbation $Q^\varepsilon := Q\left([\varepsilon, 1-\varepsilon]^2\right)$. On the other hand, α has finite length and so for almost every ε the intersection $\alpha \cap \partial Q^\varepsilon$ is finite. This can be deduced, e.g., from the area theorem in the geometric measure theory [Fed69, Mat95], which implies that an orthogonal projection of a rectifiable curve on a given line covers almost every point at most finitely many times.

Thus, approximating if necessary, we can assume that α intersects ∂Q_0 at finitely many points x_i, $i = 1, \dots, k$. Fix $\epsilon_0 > 0$, then by the RSW estimate (1.2), there is an $r > 0$ such that, if $0 < |\eta| < r$, then the probability under μ_η that there is a crossing of Q_0 which intersects any of the disks $B(x_i, r)$ is less than ϵ_0. Fix such an r, remove from the curve α the disks, the rest will be away from the boundary, simplifying future estimates:

$$\begin{aligned} \alpha' &:= \alpha \setminus \cup_i B(x_i, r), \text{ and} \\ d &:= \inf\{|x-y| : x \in \alpha', y \in \partial Q_0\} > 0 . \end{aligned}$$

Fix $s \in (0, d/4)$. By the regularity assumption, we can choose $n \le C(\alpha)/s$ balls of radius s, entirely covering α. Denote their centers by $w_1, w_2, \dots, w_n$, and set $B_j := B(w_j, 2s)$.

Let M denote the union of the set of tiles of H_η that intersect the s-neighborhood of α' minus $\cup_i B(x_i, r)$. Assume $|\eta| < s$, so that M would be contained in the $2s$-neighborhood of α'. Let ω and ω' be samples from μ_η that agree on the complement of M and are independent in M. For $j = 0, 1, \dots, n$, let ω_j be the configuration that agrees with ω' on the tiles of H_η lying inside $B_1 \cup B_\cup \cdots \cup B_j$ and agrees with ω elsewhere. Then each ω_j is a fair sample from μ_η. Also observe that $\omega_0 = \omega$ and ω_n differs from ω' only inside $\cup_i B(x_i, r)$.

Our goal is to estimate

$$\mathbf{P}\left[\omega \in \boxminus_{Q_0}, \omega' \notin \boxminus_{Q_0}\right],$$

which we will do by successively comparing ω_{j-1} to ω_j. Fix $j \in \{1, 2, \dots, n\}$. In order for $\{\omega_{j-1} \in \boxminus_{Q_0}\} \cap \{\omega_j \notin \boxminus_{Q_0}\}$ to hold, there must be a crossing in ω_{j-1} from $\partial_0 Q_0$ to $\partial_2 Q_0$ that goes through B_j and there must be a dual closed crossing in ω_j from $\partial_1 Q_0$ to $\partial_3 Q_0$ that goes through B_j, see Figure 2.2. Thus, in ω_j we have the four arm event from ∂B_j to ∂Q_0. Since the radius of B_j is $2\,s$ and the distance from B_j to ∂Q_0 is at least $d - 2\,s > d/2$, by

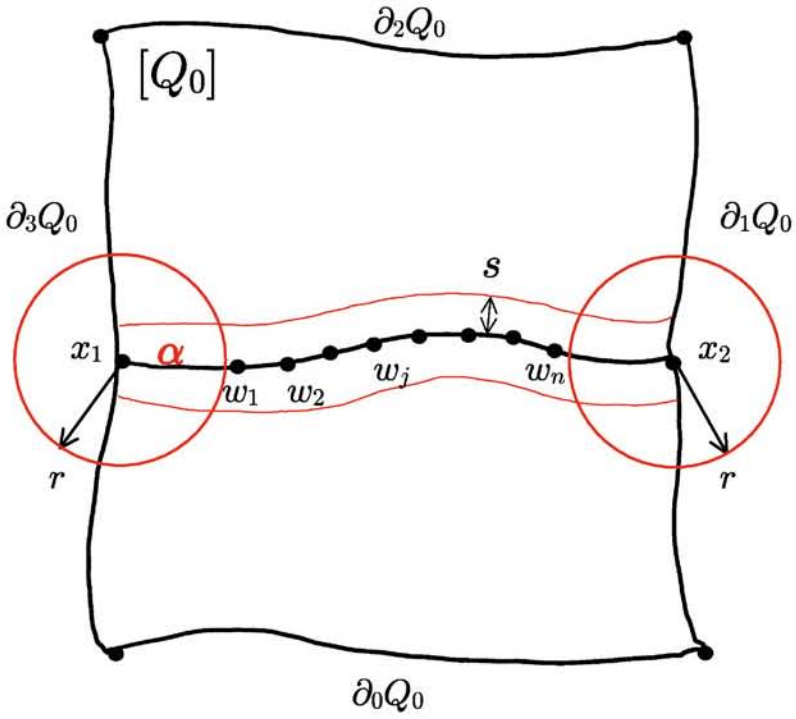


FIG 2.1. *Setup for the proof of the discrete gluing theorem in the simplest case.*

Assumption (1.3) we have

$$\mathbf{P}\left[\omega_{j-1} \in \boxminus_{Q_0}, \omega_j \notin \boxminus_{Q_0}\right] \leq P_\eta^4(z, 2s, d/2) \; < \; \left(\frac{2s}{d/2}\right) \cdot \Delta_4(2s, d/2) \, .$$

Summing over j, we conclude that

$$\begin{aligned}
\mathbf{P}\left[\omega_0 \in \boxminus_{Q_0}, \omega_n \notin \boxminus_{Q_0}\right] &\leq \sum_{j=1}^{n} \mathbf{P}\left[\omega_{j-1} \in \boxminus_{Q_0}, \omega_j \notin \boxminus_{Q_0}\right] \\
&\leq n \cdot \left(\frac{2s}{d/2}\right) \cdot \Delta_4(2s, d/2) \\
&\leq \frac{C(\alpha)}{s} \cdot \frac{4s}{d} \cdot \Delta_4(2s, d/2) \\
&\leq \frac{4C(\alpha)}{d} \cdot \Delta_4(2s, d/2) \, .
\end{aligned}$$

By Assumption 1.1 the right hand side tends to 0 with s, and so we can make sure that $\mathbf{P}\left[\omega_0 \in \boxminus_{Q_0}, \omega_n \notin \boxminus_{Q_0}\right] < \epsilon_0$ by choosing s sufficiently small. This is the only place in the paper, where we use the pivotal estimate (1.3).

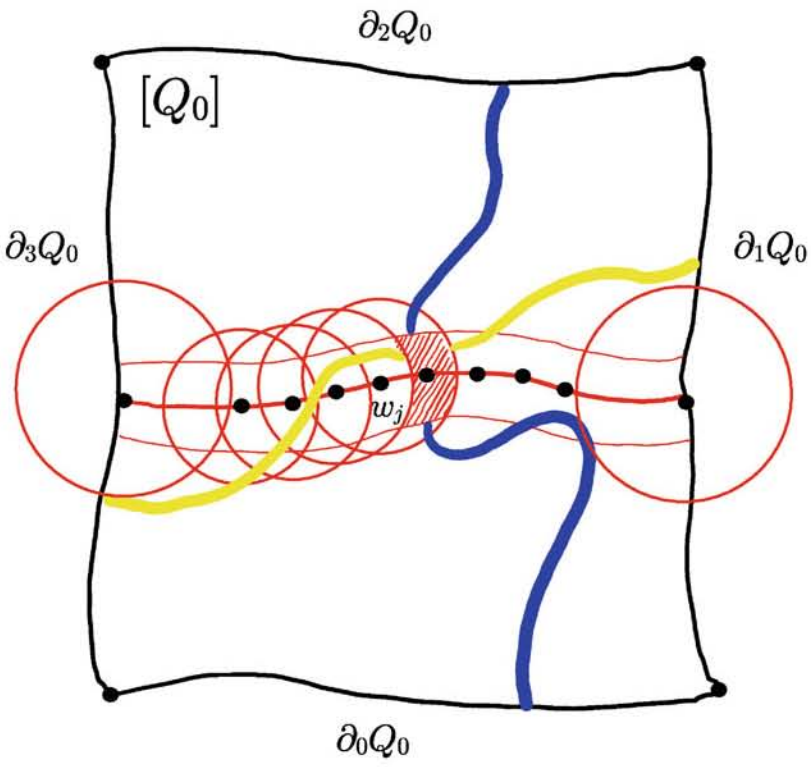


FIG 2.2. *If resampling a part of the ball B_j changes the crossing event, than there are two open crossings from B_j to $\partial_0 Q_0$ and $\partial_2 Q_0$, and two dual closed crossings from B_j to $\partial_0 Q_0$ and $\partial_2 Q_0$, — a four arm event.*

Now, recall that $\omega_0 = \omega$, and ω_n differs from ω' only near endpoints of curves in α, i.e. inside $\cup_i B(x_i, r)$. By our choice of r, we have the estimate $\mathbf{P}\left[\omega_n \in \boxminus_{Q_0}, \omega' \notin \boxminus_{Q_0}\right] < \epsilon_0$. Summing the above, we get the bound

$$\mathbf{P}\left[\omega \in \boxminus_{Q_0}, \omega' \notin \boxminus_{Q_0}\right] < 2\,\epsilon_0.$$

Now note that by first conditioning on $\mathcal{F}_s$ and then using the fact that ω and ω' are conditionally independent given $\mathcal{F}_s$, we obtain

$$2\,\epsilon_0 > \mathbf{P}\left[\omega \in \boxminus_{Q_0}, \omega' \notin \boxminus_{Q_0}\right] = \mathbf{E}\left[\mu_\eta(\boxminus_{Q_0} \mid \mathcal{F}_s)(1 - \mu_\eta(\boxminus_{Q_0} \mid \mathcal{F}_s))\right].$$

In particular, by Chebyshev inequality,

$$\mu_\eta\Big(\epsilon < \mu_\eta(\boxminus_{Q_0} \mid \mathcal{F}_s) < 1 - \epsilon\Big) < 2\,\epsilon_0/(\epsilon\,(1-\epsilon)).$$

Since ϵ_0 was an arbitrary positive number, this completes the proof. □

3. The space of lower sets.

The main goal of this Section is to provide an abstract setup for the percolation scaling limit and to prove Theorem 1.13. We discuss topology on the space $\mathcal{H}_X$ of closed lower subsets of an abstract partially ordered topological space X. In the percolation case, X corresponds to the space $\mathcal{Q}_D$ of quads, and the space $(\mathcal{H}_X, \mathcal{T}_\mathcal{H})$ to the space $(\mathcal{H}_D, \mathcal{T}_D)$ of percolation configurations in a domain $D \subset \mathbb{C}$.

With these conventions, Theorem 1.13 directly follows from Theorem 3.10 below. The latter applies to the space of percolation configurations (colorings of tilings), since $\mathcal{Q}_D$ is clearly second-countable, (3.1) holds by definition and (3.2) holds since a quad Q_0 can be easily approximated by smaller quads (e.g. by quads Q^q with $q = (-\epsilon, \epsilon, 1+\epsilon, 1-\epsilon)$ from the proof of Lemma 5.1).

We now start the (mostly classical) abstract construction, inspired by the Dedekind's sections. Let (X, τ) be a topological space, let $\mathfrak{F}_X$ denote the collection of topologically closed subsets of X. For a topologically open $U \subset X$ and a compact $K \subset X$ let

$$\begin{aligned} W_U &:= \{F \in \mathfrak{F}_X : F \cap U \neq \varnothing\}, \\ W^K &:= \{F \in \mathfrak{F}_X : F \cap K = \varnothing\}. \end{aligned}$$

Let $\hat{\mathcal{T}}$ be the topology on $\mathfrak{F}_X$ generated by all such sets W^K and W_U. It is a more convenient version of the Vietoris topology, called *Fell topology* or *topology of closed convergence.*

The following result is Lemma 1 in [Fel62]:

PROPOSITION 3.1. *$(\mathfrak{F}_X, \hat{\mathcal{T}})$ is compact.*

For completness we include a short proof, based on Alexander's subbase theorem (see, e.g., theorem 5.6 in [Kel75]), which states that if $\mathcal{B}$ is a subbase for the topology on a space X and every cover of X by elements of $\mathcal{B}$ contains a finite subcover, then X is compact.

PROOF. By Alexander's subbase theorem it suffices to show that every cover of $\mathfrak{F}_X$ by sets of the form W_U (where $U \in \tau$) and W^K (where $K \subset X$ is compact) has a finite subcover. Let

$$\{W_U : U \in I\} \cup \{W^K : K \in J\} \supset \mathfrak{F}_X$$

be such a cover. Observe that $\bigcup_{U\in I} W_U = W_{\tilde{U}}$, where $\tilde{U} = \bigcup_{U\in I} U$. Let $F := X \setminus \tilde{U}$, and note that F is topologically closed. Since $F \in \mathfrak{F}_X \subset$

$W_{\tilde{U}} \cup \bigcup_{K\in J} W^K$, we know that $F \in \bigcup_{K\in J} W^K$, namely, there is some $K_0 \in J$ with $F \in W^{K_0}$. In other words, $F \cap K_0 = \varnothing$; that is, $K_0 \subset \tilde{U}$. Since K_0 is compact and $\{U : U \in I\}$ covers K_0, there is some finite $I' \subset I$ such that $\{U : U \in I'\}$ covers K_0. Then $\{W_U : U \in I'\} \cup \{W^{K_0}\}$ is a finite cover of $\mathfrak{F}_X$. □

Now suppose that "$<$" denotes a partial order on X such that

$$\mathcal{R} := \{(x, y) \in X^2 : x < y\} \text{ is a topologically open subset of } X^2. \tag{3.1}$$

Define $\mathcal{R}^x := \{y \in X : y < x\}$ and $\mathcal{R}_x := \{y \in X : x < y\}$. Then for every $x \in X$ these sets are topologically open. A set $H \subset X$ is said to be *lower* if $x \in H$ implies $\mathcal{R}^x \subset H$. Let $\mathcal{H}_X$ denote the collection of all topologically closed lower sets.

PROPOSITION 3.2. *Assuming* (3.1), $\mathcal{H}_X \subset \mathfrak{F}_X$ *is closed with respect to the topology* $\hat{\mathcal{T}}$*; that is,* $\mathfrak{F}_X \setminus \mathcal{H}_X \in \hat{\mathcal{T}}$.

PROOF. Set $U_x := W^{\{x\}} \cap W_{\mathcal{R}_x}$ and

$$U := \bigcup_{x\in X} U_x = \{F : \exists x, y \text{ with } x < y,\ y \in F,\ x \notin F\}.$$

Then U is topologically open and $\mathcal{H}_X = \mathfrak{F}_X \setminus U$. □

Let $\mathcal{T}_{\mathcal{H}}$ be the topology of $\mathcal{H}_X$ as a subspace of $\mathfrak{F}_X$. It is the topology generated by the sets $V_U := W_U \cap \mathcal{H}_X$ and $V^K := W^K \cap \mathcal{H}_X$, for topologically open U and compact K subsets of X. As we will later see, in our setting it is sufficient to restrict ourselves to one-point compact sets, working with $V^{\{x\}}$, which we will often abbreviate V^x.

For simplicity we will also abbreviate $\mathcal{H} := \mathcal{H}_X$ and $\mathcal{T} := \mathcal{T}_{\mathcal{H}}$. Propositions 3.1 and 3.2 imply that $(\mathcal{H}, \mathcal{T})$ is compact.

In general, $\mathcal{H}$ does not have to be a Hausdorff space, even if X is Hausdorff. Indeed, the sufficient condition would be for X to be Hausdorff and compact, but for the application we have in mind, X is not compact. However, the following result gives another sufficient condition for $\mathcal{H}$ to be Hausdorff:

PROPOSITION 3.3. *Suppose, in addition to* (3.1), *that*

$$\forall x \in X,\ x \in \overline{\mathcal{R}^x}. \tag{3.2}$$

Then $(\mathcal{H}, \mathcal{T})$ *is a Hausdorff space.*

PROOF. Let $H_1, H_2 \in \mathcal{H}$ be different, then one of the two differences $H_1 \setminus H_2$ and $H_2 \setminus H_1$ is non-empty, without loss of generality the first one. Take some $x \in H_1 \setminus H_2$. Since $X \setminus H_2$ is topologically open in X and contains $x \in \overline{\mathcal{R}^x}$, we can find some $y \in \mathcal{R}^x \cap (X \setminus H_2) = \mathcal{R}^x \setminus H_2$. Then $V_{\mathcal{R}_y}$ is disjoint from V^y. Moreover, $H_1 \in V_{\mathcal{R}_y}$ and $H_2 \in V^y$. So there are disjoint topologically open sets in $\mathcal{T}$ containing H_1 and H_2, respectively, and the space is Hausdorff. □

REMARK 3.4. *Let $\mathcal{T}'$ be the topology on $\mathfrak{F}_X$ generated by the sets W_U, $U \in \mathcal{T}_X$ and $W^{\{x\}}$, $x \in X$. Since this topology is coarser than $\hat{\mathcal{T}}$, Proposition 3.1 implies that $(\mathfrak{F}_X, \mathcal{T}')$ is compact. The proof of Proposition 3.2 shows that $\mathcal{H}$ is a topologically closed subset of $\mathfrak{F}_X$, also with respect to $\mathcal{T}'$. Moreover, assuming (3.2), the proof of Proposition 3.3 shows that $\mathcal{H}$ is Hausdorff with respect to $\mathcal{T}'$. It is well known (and not hard to verify, see theorem 5.8 in [Kel75]) that if a compact topology on a space is finer than a Hausdorff topology on the same space, then they must be equal. We conclude that the topology on $\mathcal{H}$ induced by $\hat{\mathcal{T}}$ is the same as that induced by $\mathcal{T}'$ whenever (3.2) and (3.1) hold.*

LEMMA 3.5. *Suppose that (3.1) and (3.2) hold, and that X_0 is a dense subset of X. If $H_1, H_2 \in \mathcal{H}$ and $H_1 \neq H_2$, then $H_1 \cap X_0 \neq H_2 \cap X_0$.*

PROOF. Take $x \in H_1 \setminus H_2$, if the latter is non-empty. Then (3.2) implies that $\mathcal{R}^x \setminus H_2 \neq \varnothing$. Since X_0 is dense and $\mathcal{R}^x \setminus H_2$ is topologically open and nonempty, we have $X_0 \cap \mathcal{R}^x \setminus H_2 \neq \varnothing$. Because $\mathcal{R}^x \subset H_1$, this implies that $H_1 \cap X_0 \neq H_2 \cap X_0$. A symmetric argument applies if $H_2 \setminus H_1 \neq \varnothing$. □

In the next Lemma we characterize convergence of nets in $\mathcal{H}_X$. As we shall see, $(\mathcal{H}_X, \mathcal{T}_{\mathcal{H}})$ is first countable, thus convergence of sequences actually leads to the same properties, but we decided to work with the more general notion for now, since it does not lead to additional difficulties.

Suppose that $\mathcal{D}$ is a directed set (i.e. partially ordered by "$\preceq$" so that every two elements have a common upper bound), and $\{Y_n\}_{n \in \mathcal{D}}$ is a net (i.e. a sequence indexed by $\mathcal{D}$) of subsets of X. We write $\limsup_n Y_n$ for the set of all $x \in X$ with the property that for every topologically open $U \subset X$ containing x and for every $m \in \mathcal{D}$ there is some $n \succeq m$ in $\mathcal{D}$ satisfying $Y_n \cap U \neq \varnothing$. Similarly, $\liminf_n Y_n$ is the set of all $x \in X$ with the property that for every topologically open $U \subset X$ containing x there is some $m \in \mathcal{D}$ such that $Y_n \cap U \neq \varnothing$ for every $n \succeq m$. We write $Y = \lim_n Y_n$ if $Y = \limsup_n Y_n = \liminf_n Y_n$.

LEMMA 3.6. *Assume that (3.1) and (3.2) hold, and that S_n, $n \in \mathcal{D}$, is a net with $S_n \in \mathcal{H}$ for every n.*

1. *If $\lim_n S_n \subset X$ exists, then $\lim_n S_n \in \mathcal{H}$.*
2. *If the net S_n converges to S with respect to $\mathcal{T}$, then $S = \lim_n S_n$.*
3. *Conversely, if $S = \lim_n S_n$, then the net S_n converges to S with respect to $\mathcal{T}$.*

PROOF. To prove (1), suppose that $S = \lim_n S_n$, $x \in S$ and $y < x$. Then $x \in \mathcal{R}_y$. Therefore, there is some $m \in \mathcal{D}$ such that $S_n \cap \mathcal{R}_y \neq \varnothing$ for every $n \succeq m$. Since $S_n \in \mathcal{H}$, this implies that $y \in S_n$ for every $n \succeq m$. Thus, $y \in S$, which proves that S is a lower set. The definition of $\lim_n S_n$ makes it clear that S is topologically closed. This proves (1).

To prove (2), suppose that S_n converges to S with respect to $\mathcal{T}$. Let $x \in S$. If $U \subset X$ is topologically open and contains x, then $S \in V_U$. Therefore, there is some $m \in \mathcal{D}$ such that $S_n \in V_U$ for every $n \succeq m$. But $S_n \in V_U$ is equivalent to $S_n \cap U \neq \varnothing$. Therefore, $x \in \liminf_n S_n$; that is, $S \subset \liminf_n S_n$. Now suppose that $y \notin S$. By (3.2) there is some $z \in \mathcal{R}^y \setminus S$. Then $S \in V^z$. Consequently, there is some $m \in \mathcal{D}$ such that $S_n \in V^z$ for all $n \succeq m$. Hence $S_n \cap \mathcal{R}_z = \varnothing$ for all $n \succeq m$. Since $y \in \mathcal{R}_z$ and $\mathcal{R}_z$ is topologically open, this implies that $y \notin \limsup_n S_n$. Thus, $S \supset \limsup_n S_n$. Summing it up, we have $\liminf_n S_n \supset S \supset \limsup_n S_n$. Since $\liminf_n S_n \subset \limsup_n S_n$, this proves (2).

To prove (3), suppose that $S = \lim_n S_n$. Then we know from (1) that $S \in \mathcal{H}$. Suppose that S_{n_j} is a subnet of S_n that converges to some $S' \in \mathcal{H}$. By (2), we have $S' = \lim_j S_{n_j}$. Therefore, $S' = S$. Thus, every convergent subnet of S_n converges to S. Since $\mathcal{H}$ is compact, this proves that S_n converges to S, and completes the proof of the Lemma. □

LEMMA 3.7. *Assuming (3.1) and (3.2), if X is second-countable (i.e. with a countable base), then $\mathcal{H}$ is second-countable and hence metrizable.*

PROOF. Assume that X is second-countable. In particular, there is a countable dense set $X_0 \subset X$. Let $\mathcal{B}$ be a countable basis for $\mathcal{T}_X$. We claim that $\mathcal{B}' := \{V_U : U \in \mathcal{B}\} \cup \{V^x : x \in X_0\}$ is a subbase for $\mathcal{T}$.

Indeed, if $U' \in \mathcal{T}_X$, then there is a subset $J \subset \mathcal{B}$ such that $U' = \bigcup_{U \in J} U$. Then $V_{U'} = \bigcup_{U \in J} V_U$; thus, $V_{U'}$ is in the minimal topology containing $\mathcal{B}'$.

Now suppose that $x \in X$, then

$$V^x = \bigcup_{y \in X_0 \cap \mathcal{R}^x} V^y. \tag{3.3}$$

Indeed, if $S \in V^x$, then $x \notin S$ and by (3.2) the topologically open set $\mathcal{R}^x \setminus S$ is nonempty. Take some $y \in X_0 \cap \mathcal{R}^x \setminus S$, then $S \in V^y$. This proves that $V^x \subset \bigcup_{y \in X_0 \cap \mathcal{R}^x} V^y$. Since $y < x$, we also have $V^y \subset V^x$ for $y \in \mathcal{R}^x$, concluding that $V^x = \bigcup_{y \in X_0 \cap \mathcal{R}^x} V^y$. Thus V^x is also in the minimal topology containing $\mathcal{B}'$.

Now Remark 3.4 shows that $\mathcal{B}'$ is a subbase for $\mathcal{T}$. Since $\mathcal{B}'$ is countable, $\mathcal{T}$ also has a countable basis (of all finite intersections of elements of $\mathcal{B}'$).

By Proposition 3.3, topology $\mathcal{T}$ is Hausdorff. Since a second-countable compact Hausdorff space is metrizable (which follows from the Urysohn's theorem 4.16 in [Kel75]), this completes the proof. □

LEMMA 3.8. *Suppose that* (3.1) *and* (3.2) *hold, and that* X_0 *is a dense subset of* X. *Then the Borel* σ*-field of* $(\mathcal{H}, \mathcal{T})$ *is generated by* V^y, $y \in X_0$.

PROOF. Observe that the Borel σ-field of $(\mathcal{H}, \mathcal{T})$ is generated by topologically open sets, which in turn are generated by the sets of the form V_U and V^x for topologically open $U \subset X$ and $x \in X$. So we must prove that those can be obtained from V^y, $y \in X_0$ by the operations of countable union, intersection and taking complements.

For $x \in X$, the set V^x can be obtained as $V^x = \bigcup_{y \in X_0 \cap \mathcal{R}^x} V^y$, by (3.3).

For a topologically open $U \subset X$ note that by the definitions $V_U = \cup_{x \in U} \neg V^x$. Therefore $V_U \supset \cup_{y \in U \cap X_0} \neg V^y$. On the other hand, take $x \in U$. By (3.2) we can find $y \in X_0$ inside the topologically open set $U \cap \mathcal{R}^x$, then $y < x$ and $\neg V^x \subset \neg V^y$. Thus $V_U \subset \cup_{y \in U \cap X_0} \neg V^y$. We conclude that $V_U = \cup_{y \in U \cap X_0} \neg V^y$, and so the sets V_U can also be obtained. □

Finally observe that a topologically open $X' \subset X$ inherits the properties (3.1) and (3.2) from X, and the following continuity under inclusions holds:

LEMMA 3.9. *Let* $X' \subset X$ *be topologically open, and for* $S \in \mathcal{H}_X$ *let* $\Phi(S) := S \cap X'$. *Then* Φ *is a continuous map from* $\mathcal{H}_X$ *to* $\mathcal{H}_{X'}$.

PROOF. We have to check that preimages of topologically open sets are also topologically open. This follows easily since a topologically open subset $U \subset X'$ is topologically open inside X and $\Phi^{-1}(V^U) = V^U$, while for $x \in X'$ we have $\Phi^{-1}(V^x) = V^x$. □

Summing it up, we arrive at the following

THEOREM 3.10. *Let* (X, τ) *be a partially ordered second-countable topological space, such that the ordering satisfies* (3.1) *and* (3.2). *Then*

1. *$(\mathcal{H}_X, \mathcal{T}_\mathcal{H})$ is a compact metrizable Hausdorff space.*
2. *Let X_0 be a dense subset of X. If $S_1, S_2 \in \mathcal{H}_X$ satisfy $S_1 \cap X_0 = S_2 \cap X_0$, then $S_1 = S_2$. Moreover, the σ-field generated by V^y, $y \in X_0$ is the Borel σ-field of $(\mathcal{H}_X, \mathcal{T}_\mathcal{H})$.*
3. *If $\{S_n\}$ is a sequence in $\mathcal{H}_X$ and $S \in \mathcal{H}_X$, then $S_n \to S$ in $\mathcal{T}_\mathcal{H}$ is equivalent to $S = \limsup_n S_n = \liminf_n S_n$.*
4. *Let $X' \subset X$ be a topologically open subset, and for $S \in \mathcal{H}_X$ let $\Phi(S) := S \cap X'$. Then Φ is a continuous map from $\mathcal{H}_X$ to $\mathcal{H}_{X'}$.*

PROOF. Part 1 follows from Propositions 3.1, 3.2 and Lemma 3.7; part 2 follows from Lemmas 3.5 and 3.8; part 3 follows from Lemma 3.6; and part 4 follows from Lemma 3.9. □

4. Uniformity in the mesh size.

In this Section, we prove a mesh-independent version of the gluing theorem:

PROPOSITION 4.1 (MESH-INDEPENDENT GLUING). *Let $D \subset \mathbb{C}$ be some domain, and let $Q_0 \in \mathcal{Q}_D$ be a piecewise smooth quad in D. Suppose that $\alpha \subset \mathbb{C}$ is a finite union of finite length paths, with finitely many double points. Consider a collection of percolation models indexed by a set $\{\eta\}$. Assume that μ_η converges to (a subsequential) scaling limit μ_0 as mesh $|\eta| \to 0$. Then for every $\epsilon > 0$ there is a finite set of piecewise smooth quads $\mathcal{Q}_\epsilon \subset \mathcal{Q}_{D\setminus\alpha}$ and a subset $\mathcal{W}_\epsilon \subset \mathcal{H}$, which is measurable with respect to the finite σ-field $\sigma(\boxminus_Q : Q \in \mathcal{Q}_\epsilon)$, such that*

$$\lim_{|\eta|\to 0} \mu_\eta(\mathcal{W}_\epsilon \Delta \boxminus_{Q_0}) < \epsilon\,. \tag{4.1}$$

PROOF. This is the most technically difficult part of our paper. The rough plan is as follows: we set up a thin strip around α, and cut it into "bays," forming a neighborhood of α, and a part away from α, bounded by "beaches." Partition depends on the percolation configuration, and has the properties that (i) percolation in the neighborhood of α is decoupled from the rest, and (ii) the crossings outside this neighborhood can be encoded by a finite information, stable under small perturbations of the percolation configuration. We use a complicated construction of the neighborhood for the sake of the property (ii). To effectively bound the needed crossing information away from α (whose amount grows as we pass to the scaling limit), we further coarse-grain this procedure to a fixed small scale, when it is described by a finite σ-field. This will allow to write the required estimates.

We now start the proof by fixing a countable subsequence $\{\eta_j\}$ from our set, such that $|\eta_j| \to 0$ (and μ_{η_j} converges to μ_0). Recall that by Remark 1.17 in our setting it is sufficient to work with sequences, rather than nets.

Setup for the neighborhood of α. Reasoning like in the proof of Theorem 1.5 (the first paragraph of Section 2), we can approximate the quad Q_0, by quads whose boundaries intersect α on a finite set. Thus we can assume that α has finitely many double points, intersects ∂Q_0 on a finite set. Prolonging some of the curves in α if necessary (to cut the non-simply-connected components into smaller pieces), we can further assume that each component of $[Q_0] \setminus \alpha$ is simply connected.

Fix small $\epsilon_0 > 0$. Define two random variables: one given by the indicator function of the crossing event $Y_0 := 1_{\boxminus_{Q_0}}$; and another by the conditional probability of a crossing, given the percolation configuration s-away from α:

$$Y_s := \mu_\eta(\boxminus_{Q_0} \mid \mathcal{F}_s).$$

Here $\mathcal{F}_s$, $s > 0$, denotes the σ-field generated by the restriction of ω to the complement of the s-neighborhood of α, as in Theorem 1.5. By that theorem, for all sufficiently small $s > 0$,

$$\sup_{|\eta| \in (0,s)} \mu_\eta(\epsilon_0 < Y_s < 1 - \epsilon_0) < \epsilon_0 \,. \tag{4.2}$$

Denote by x_i, $i = 1, \ldots, n$, the finitely many points where α intersects ∂Q_0, as well as the endpoints and the double points of curves in α. Let d be the minimum over i of the distances from x_i to other x_j's and the farthest of $\partial_0 Q_0$ and $\partial_2 Q_0$. We now fix $s > 0$ sufficiently small so that (4.2) holds, and sufficiently small so that for $i = 1, \ldots, n$, the balls $B(x_i, 2s)$ are disjoint, and

$$\sup_{|\eta| \in (0,s)} \mu_\eta(\exists \text{ a crossing from } B(x_i, 2\,s) \text{ to } \partial B(x_i, d/3)) < \epsilon_0/n \,. \tag{4.3}$$

The latter follows for small s from the RSW estimate (1.2). We also require that the intersection of α with $\cup_i \partial B(x_i, s)$ is finite (which holds for almost every value of s, since the total length of α is finite).

For the following construction, the reader is advised to refer to Figure 4.1. Let β and β' be finite unions of disjoint smooth simple paths in $[Q_0] \setminus \cup_i B(x_i, s)$, and let K be the closure of the union of connected components of

$$[Q_0] \setminus (\cup_i B(x_i, s) \cup \beta \cup \beta')$$

whose boundary intersects both β and β'. We can choose these unions of paths so that

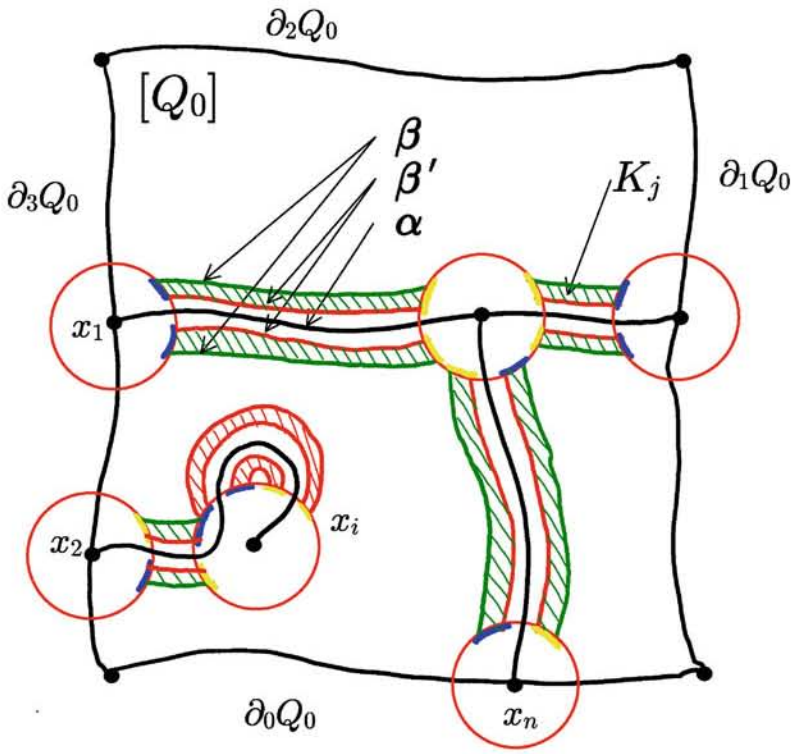


FIG 4.1. *Curves in α are between β' which are in turn between β. We modify configuration on the small discs so that long quads bounded by β and β' on long sides, have open tiles on one short side and closed on another.*

- all their endpoints belong to $\cup_i \partial B(x_i, s)$,
- β' separates β from α,
- the connected component of $[Q_0] \setminus \beta$ that contains α is contained in the $(s/2)$-neighborhood of α,
- each component of connectivity K_j of K is a quad with two "long" sides on β and β' and two "short" sides on two of the circles $\partial B(x_i, s)$.

It is immediate to verify that there exist such unions of paths.

Bays and beaches. The following construction is illustrated in Figure 4.2. Let ω be a sample from μ_η for some small $\eta > 0$. For somewhat technical reasons it will be convenient to consider in place of ω the configuration $\tilde{\omega}$ which is modified on the tiles intersecting disks $B(x_j, s)$. We alter ω so that for every component K_i one of two short sides is covered by open tiles and another by closed. Clearly, (4.3) implies that if $\eta < s$,

$$\mu_\eta(\{\omega \in \boxminus_{Q_0}\} \Delta \{\tilde{\omega} \in \boxminus_{Q_0}\}) \le \epsilon_0 \,. \tag{4.4}$$

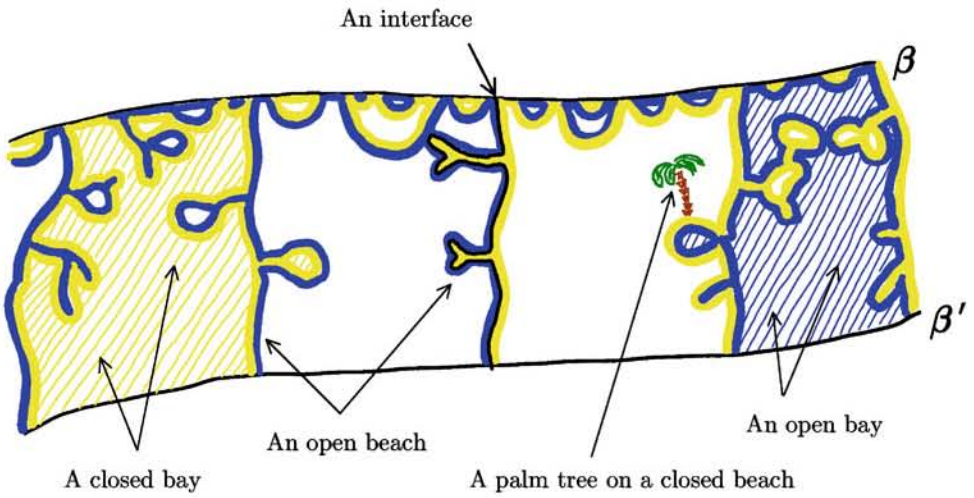


FIG 4.2. *Interfaces between open and closed clusters, starting from β, cut the strips K_i into components.* Bays *are the components touching β', and are alternatively open and closed.* Beaches *are the inner boundaries of the bays, when entered through β'.*

For this reason, studying $\tilde{\omega}$ will still be useful.

In what follows, we start on the curve β and explore all the potential crossings of the strip between β and β'. To be more precise, denote by $\partial\tilde{\omega}$ the set of percolation interfaces — curves, separating open tiles in $\tilde{\omega}$ from the closed ones. Let I be the set of connected components of the intersection of $\partial\tilde{\omega}$ with the interior of K (i.e., percolation interface in the interior of K), and let $\tilde{I}$ denote the set of elements of I that have at least one endpoint on β. The definition of $\tilde{\omega}$ guarantees that in each component K_i there is at least one interface of $\tilde{\omega}$ with one endpoint on β and the other on β'.

Let Γ denote the union of the elements of $\tilde{I}$; that is, $\Gamma = \bigcup \tilde{I} = \bigcup_{\gamma\in\tilde{I}} \gamma$. Each connected component of $K \setminus \Gamma$ which meets β' will be called a *bay*. Clearly, each point in β' is in the closure of some bay, and for each bay its intersection with β' is connected. Let B be some bay. The union of the tiles of H_η that intersect B as well as $\partial B \setminus \beta'$ will be called the *beach* of B. It is easy to see that the beach is connected, and all the tiles in the beach are either open (i.e. in $\tilde{\omega}$) or closed (not in $\tilde{\omega}$). If they are all open [respectively, closed], then B itself will be referred to as an open [respectively, closed] bay.

Let $M' = M'(\omega)$ be the union of the component of $[Q_0] \setminus \beta'$ containing α, and of all the bays without beaches. Denote by $M = M(\omega)$ the complement in $[Q_0]$ of M', as well as the corresponding set of tiles. The following property of M is essential: given M and the restriction of ω to it, the conditional law of the restriction of ω to M' (i.e. the complement of M) is unbiased. In

other words, if we take ω_1 independent from ω and of the same law, and we define ω_2 to agree with ω on M and to agree with ω_1 elsewhere, then ω_2 also has the law μ_η. This follows directly from the fact that $M(\omega_2) = M(\omega)$. Below we will denote the restriction of a percolation configuration ω to M by $\omega \llcorner M$.

Fix some small $s' \in (0, s)$, and let $S = S_{s'}$ denote the event that there is a small bay B, i.e. with $\operatorname{diam}(B \cap \beta') < s'$. Observe that for every bay B the endpoints of $B \cap \beta'$ are endpoints of interfaces of ω which meet both β and β'. Thus three crossings of K (which are alternating, e.g. closed, open and closed) land on the arc $B \cap \beta'$ and Lemma A.2 applied to all components of K implies that

$$\lim_{s' \to 0} \sup_{|\eta| \in (0, s')} \mu_\eta(S) \;=\; 0\,.$$

By choosing s' sufficiently small, we therefore assume, with no loss of generality, that $\mu_\eta(S) < \epsilon_0$ holds for every $\eta \in (0, s')$ and that $s' < s$. Observe that on $\neg S$ the number of bays is bounded by a constant depending only on s', β, β' and Q_0.

We will find the structure of the bays to be a convenient component in the estimation of the conditional probability (given the bays and some additional information) of $\tilde{\omega} \in \boxminus_{Q_0}$. However, as $|\eta| \to 0$, the number of possible bay configurations increases without bound, and therefore their structure cannot be completely captured by a finite σ-field. For this purpose, we need to introduce a discretization, and any sort of coarse-graining, e.g. by the square lattice, would do.

Discretization to a finite σ-field by coarse-graining. We need some tessellation of K, and the following one is somewhat ad hoc, but suitable. Let $\delta \in (0, s')$ be small, to be chosen later. Fix some finite set of points $W \subset K$ such that the distance between any two points in W is at least $\delta/4$ and every point in K is within distance $\delta/2$ of some point in W. Let T be the collection of Voronoi tiles in K with respect to W; that is, each T_w, $w \in W$, is the set of points $x \in K$ whose distance to W is equal to $|x - w|$. Then the diameters of the tiles in the tessellation T are all less than δ (but the lattice H_η tiles are still much smaller). Let $\hat{T}$ denote the tessellation obtained by adjoining to T the closures of the (finite in number) connected components of $[Q_0] \setminus K$ that do not contain α. We think of tessellations T and $\hat{T}$ as cell complexes. The vertices of T are defined as the vertices of the Voronoi tiles together with the points in $(\beta \cup \beta') \cap \partial Q_0$. The vertices of $\hat{T}$ are defined as those of T plus the four corners of Q_0. Note that some of the edges of $\hat{T}$ are not necessarily straight. Denote by $[\hat{T}]$ the union of the tiles in $\hat{T}$.

We say that a quad $Q \in \mathcal{Q}_D$ is *compatible* with $\hat{T}$ if $[Q]$ is a union of tiles of $\hat{T}$ and each of its four corners $Q(0,0)$, $Q(0,1)$, $Q(1,0)$, $Q(1,1)$ is a vertex of $\hat{T}$. Since $\hat{T}$ is finite, it is clear that up to reparameterization preserving corners there are only finitely many compatible Q. Let $\mathcal{Q}_{\hat{T}}$ denote a set of representatives of equivalence classes of compatible quads up to reparameterization preserving corners. Then $\mathcal{Q}_{\hat{T}}$ is a finite collection of piecewise smooth quads in $D \setminus \alpha$. Let $\mathcal{F}_T$ denote the σ-field generated by the events $\boxminus_Q$, $Q \in \mathcal{Q}_{\hat{T}}$.

Without loss of generality we assume that along our chosen sequence $\{\eta_j\}$, for each edge e of the tessellation T that does not touch ∂Q_0,

- each edge of H_η intersects the tessellation edge e in at most finitely many points,
- the tessellation edge e does not contain any vertices of H_η,
- the endpoints of the tessellation edge e are not on edges of H_η.

There is no loss of generality in this assumption, because we may slightly perturb β, β' and T (and we don't even need T to be a Voronoi tessellation).

Let B be some bay. Define $T(B)$ to be its T-discretization, namely the union of all edges of T inside $B \cap \beta'$ and all tiles of T inside B, which can be connected to β' by a path of tiles of T also contained inside B. By construction, each point in $\partial T(B)$ is within distance δ from ∂B. Note that there is a finite number of possibilities for the choices of $T(B)$, and also observe that whenever S does not hold, all T-discretizations are non-empty and so $T(B) = T(B')$ implies $B = B'$.

Connectivity in the coarse-grained model. The next objective is to show that on the complement of S the collections of discretized bays

$$\begin{aligned} Z_o &:= \{T(B) : B \text{ an open bay}\}, \text{ and} \\ Z_c &:= \{T(B) : B \text{ a closed bay}\}, \end{aligned}$$

can be determined from $\mathcal{F}_T$. That is, there are $\mathcal{F}_T$-measurable random variables Z_o' and Z_c' such that $Z_o = Z_o'$ and $Z_c = Z_c'$ on the complement of S. Here and in the following, we ignore events that have zero μ_η-measure for each $\eta \in \{\eta_j\}$.

Note that on the event $\neg S$ the cardinality of Z_o and Z_c is bounded by a constant depending on Q_0, β' and s' only.

Let e be an edge of T that lies on the component $\beta_j' = \beta' \cap K_j$, denote also $\beta_j := \beta \cap K_j$. Let Q_e and Q_e' be the quads in $\mathcal{Q}_{\hat{T}}$ that satisfy $\partial_0 Q_e = \partial_1 Q_e' = e$, $\partial_2 Q_e = \partial_3 Q_e' = \beta_j$ and $[Q_e] = [Q_e'] = K_j$. Then the event that there is an interface $\gamma \in \tilde{I}$ that meets e is the same as the event $\boxminus_{Q_e} \setminus \boxminus_{Q_e'}$.

If E is the set of all edges of T on β'_j which meet interfaces in $\tilde{I}$, then on the event $\neg S$, the connected components of $\beta'_j \setminus \bigcup E$ are the intersections of the elements of $Z_o \cup Z_c$ with β'_j, and since the bays alternate in color, it follows that the sets $\{\tau \cap \beta'_j : \tau \in Z_o\}$ and $\{\tau \cap \beta'_j : \tau \in Z_c\}$ can each be determined from $\mathcal{F}_T$, up to an event contained in S.

Next, suppose that $\gamma \subset \beta'_j$ is of the form $\tau_0 \cap \beta'_j$ for some $\tau_0 \in Z_c$ (assuming $\neg S$, such intersection is always non-empty). Let T_0 be some tile of T. Then $T_0 \subset \tau_0$ if and only if there is a simple path γ' of edges of T with the endpoints of γ' on γ such that γ' separates T_0 from β_j in K_j and there is no open crossing in $K_j \cap \tilde{\omega}$ from γ' to β_j. This, along with the dual argument, implies the existence of Z'_o and Z'_c, as claimed.

Describing the crossing structure by a finite graph. We now introduce a graph that describes the connectivity (by open crossings) structure of the various open beaches and the two boundary edges $\partial_0 Q_0$ and $\partial_2 Q_0$.

We start by defining a random graph G which would describe the connectivity away from α. The vertices of G are $Z_o \cup \{\partial_0 Q_0, \partial_2 Q_0\}$, where an edge is placed between vertices $v, v' \in Z_o$ if the beaches of the corresponding bays are connected by an open crossing in the restriction of $\tilde{\omega}$ to M, denoted by $\tilde{\omega} \llcorner M$; while if v and/or v' are in $\{\partial_0 Q_0, \partial_2 Q_0\}$, the connectivity to the beach is simply replaced by the connectivity to the corresponding boundary edge itself. Note that there is a finite number of possibilities for the choice of G.

Suppose that $\tau_0, \tau_1 \in Z_o$. Let e_j, $j = 0, 1$, be an edge of T that is contained in β' and has an endpoint in τ_j, but e_j itself is not in τ_j. Then e_j meets an interface on the boundary of the corresponding bay. On $\neg S$, the two edges e_j are connected by an open crossing in $\tilde{\omega} \llcorner [\hat{T}]$ if and only if $[\tau_0, \tau_1]$ is an edge of G. Thus, on the event $\neg S$ the subgraph of G induced by Z_o is determined from $\mathcal{F}_T$. A similar argument applies also to the edges with endpoints in $\{\partial_0 Q_0, \partial_2 Q_0\}$. Thus, on the complement of S the graph G can be determined from $\mathcal{F}_T$.

Now we define a random graph G^* so that it describes the connectivity near α. Consider $L = (\partial_0 Q_0 \cup \partial_2 Q_0) \setminus [\hat{T}]$, which contains finitely many arcs on ∂Q_0. Denote by G^* the graph whose vertices are Z_o and the connected components of L, where $v, v' \in Z_o$ are connected by an edge in G^* if the two corresponding beaches of $\tilde{\omega}$ are connected by an open crossing in $\tilde{\omega} \llcorner ([Q_0] \setminus M)$, and if v and/or v' are components of L, then the connectivity to the beaches is replaced by connectivity to v and/or v'.

Clearly, $\tilde{\omega} \in \boxminus_{Q_0}$ if and only if there is a path from $\partial_0 Q_0$ (or a subarc thereof) to $\partial_2 Q_2$ in $G \cup G^*$. Thus, on the event $\neg S$, the graphs G and G^*

determine whether $\tilde{\omega} \in \boxminus_{Q_0}$.

Denoting by $\omega \llcorner M$ the restriction of ω to M, we set

$$\tilde{Y} = \tilde{Y}(\omega) := \mu_\eta(\tilde{\omega} \in \boxminus_{Q_0} \mid \omega \llcorner M) .$$

We now show that on $\neg S$, the knowledge of G, Z_c and Z_o can be used to approximate $\tilde{Y}$.

Let

$$T(M) := [\hat{T}] \setminus \Big(\bigcup Z_c \cup \bigcup Z_o \Big),$$

that is, the complement in $[\hat{T}]$ of the union of $T(B)$, where B runs over all the bays.

Let $G^\dagger$ be the graph on the same vertex set as G^*, where an edge appears between $v, v' \in Z_o$ if the common boundary of v and $T(M)$ is connected in $\tilde{\omega} \llcorner ([Q_0] \setminus T(M))$ with the common boundary of v' and $T(M)$, while if v and/or v' are components of L, then the connectivity is instead to v and/or v', but still within $\tilde{\omega} \llcorner ([Q_0] \setminus T(M))$. We now show that if δ is sufficiently small, $\eta < \delta$ and $\eta \in \{\eta_n\}$, then

$$\mu_\eta(G^\dagger \neq G^*, \neg S \mid \omega \llcorner M) < \epsilon_0 . \tag{4.5}$$

As we have mentioned, on the event $\neg S$ there is a finite upper bound on the size of $Z_c \cup Z_o$, depending on Q_0, β' and s' only. Therefore, it is enough to have a good estimate for the probability that a particular pair of vertices v and v' of G^* are connected by an edge in one of the two graphs but not in the other, and we are free to impose additional conditions on δ. Connectivity of any pair of vertices is reduced to δ-perturbations of quads of size at least s', and so is easily obtained by several applications of Lemma A.1 (some of them applied with the corners of the quads appropriately permuted), proving (4.5).

Final estimates. Define $\tilde{Y}_0 = \tilde{Y}_0(\omega) := 1_{\tilde{\omega} \in \boxminus_{Q_0}}$ and

$$\tilde{Y}_T := \mu_\eta(\partial_0 Q_0 \text{ and } \partial_2 Q_0 \text{ are connected by a path in } G^\dagger \cup G \mid \omega \llcorner M).$$

Then $\tilde{Y}_T$ is a function of the triple (Z_c, Z_o, G), and is therefore $\mathcal{F}_T$ measurable. By the above observation that on $\neg S$ the event $\boxminus_{Q_0}$ is described by connectivity in $G \cup G^*$, we have

$$|\tilde{Y}_T - \tilde{Y}| \, 1_{\neg S} \leq \mu_\eta(G^\dagger \neq G^*, \neg S \mid \omega \llcorner M).$$

By taking expectations, applying (4.5) and recalling that s' was chosen to guarantee $\mu(S) < \epsilon_0$ one gets

$$\|\tilde{Y}_T - \tilde{Y}\|_2 < \sqrt{2\epsilon_0}\,, \tag{4.6}$$

where the norm refers to the measure μ_η.

By (4.2) and the definition of Y_s, we can write

$$\|Y_0 - Y_s\|_2^2 < 2\,\epsilon_0 + \epsilon_0^2 < 3\,\epsilon_0\,,$$

and by (4.4), we have

$$\|Y_0 - \tilde{Y}_0\|_2 < \sqrt{\epsilon_0}\,. \tag{4.7}$$

Hence $\|\tilde{Y}_0 - Y_s\|_2 < \sqrt{3\,\epsilon_0} + \sqrt{\epsilon_0} < 3\,\sqrt{\epsilon_0}$. Since Y_s is $\mathcal{F}_s$-measurable, it is also $\omega \llcorner M$-measurable. By its definition $\tilde{Y}$ minimizes $\|\tilde{Y}_0 - X\|_2^2$ among $\omega \llcorner M$-measurable random variables X. Therefore comparison to Y_s (recall that the set M includes the complement of the s-neighborhood of α) yields

$$\|\tilde{Y}_0 - \tilde{Y}\|_2 \le \|\tilde{Y}_0 - Y_s\|_2 < 3\,\sqrt{\epsilon_0}\,.$$

Combining this with (4.6) and (4.7), we conclude that

$$\|Y_0 - \tilde{Y}_T\|_2 < 6\,\sqrt{\epsilon_0}\,.$$

Because $\tilde{Y}_T$ is $\mathcal{F}_T$-measurable and ϵ_0 may be chosen arbitrarily small, this proves that there is an $\mathcal{F}_T$-measurable event $\mathcal{W}$ (which may depend on η), such that $\mu_\eta(\mathcal{W} \Delta \boxminus_{Q_0}) < \epsilon/2$. However, since $\mathcal{F}_T$ is finite, there are finitely many possibilities for the event $\mathcal{W}$, and one of those works for a sub-subsequence η_{j_k}. Since we work with a (subsequential) scaling limit, the limit in (4.1) exists along the original subsequence and is equal to $\mu_0\,(\mathcal{W} \Delta \boxminus_{Q_0})$. Since along a sub-subsequence the quantity in question is bounded by ϵ, so is the limit and we deduce (4.1). □

5. Factorization.

In this Section, we prove the Factorization Theorem 1.19.

LEMMA 5.1. *Let μ_0 be some subsequential scaling limit, then the boundary (in topology $\mathcal{T}$ on $\mathcal{H}$) of a crossing event has probability zero. Namely,*

$$\mu_0(\partial \boxminus_{Q_0}) = 0$$

holds for every $Q_0 \in \mathcal{Q}_D$.

PROOF. Fix $Q_0 \in \mathcal{Q}_D$ and let $\epsilon > 0$. It is easy to see (e.g., using the Riemann map onto $\hat{\mathbb{C}} \setminus [Q_0]$) that there is a continuous injective $\hat{Q}_0 : [-1,2]^2 \to \mathbb{C}$, whose restriction to $[0,1]^2$ is Q_0. Let $M := [-1, 1/3]^2 \times [2/3, 2]^2$. For every quadruple $q = (x_0, y_0, x_1, y_1) \in M$ define the quad $Q^q : [0,1]^2 \to \mathbb{C}$ by

$$Q^q(x,y) := \hat{Q}_0(x_0 + (x_1 - x_0)\, x, y_0 + (y_1 - y_0)\, y).$$

It is a perturbation of Q_0, obtained as an image of $[x_0, x_1] \times [y_0, y_1]$ by $\hat{Q}_0$. Lemma A.1 implies that there is some positive $\delta_0 > 0$ (depending on $\hat{Q}_0$), such that if q and q' are in M and differ in exactly one coordinate, and the difference in that coordinate is at most δ_0, then $\limsup_{|\eta|\to 0} \mu_\eta(\boxminus_{Q^q} \Delta \boxminus_{Q^{q'}}) < \epsilon$. Consequently, we have

$$\limsup_{|\eta|\to 0} \mu_\eta(\boxminus_{Q^q} \Delta \boxminus_{Q^{q'}}) < 4\,\epsilon, \quad \text{provided} \quad \|q - q'\|_\infty \le \delta_0.$$

Set $q_s := (-s, s, 1+s, 1-s)$, $Q' := Q^{q_{-\delta_0/2}}$ and $Q'' := Q^{q_{\delta_0/2}}$. Then $Q' < Q_0 < Q''$ and

$$\limsup_{|\eta|\to 0} \mu_\eta(\boxminus_{Q'} \Delta \boxminus_{Q''}) < 4\,\epsilon\,. \tag{5.1}$$

Also, $Q', Q'' \in \mathcal{Q}_D$, if δ_0 is chosen sufficiently small.

Recall that $V^{Q_0} = \mathcal{H}_D \setminus \boxminus_{Q_0}$ is open in $\mathcal{H}_D$. Hence $\boxminus_{Q_0}$ is closed. Let U' be the set of quads $Q \in \mathcal{Q}_D$ satisfying $Q_0 < Q < Q''$. Then

$$\boxminus_{Q_0} \supset V_{U'} \supset \boxminus_{Q''}.$$

By passing to the complements, we see that there is a closed subset (the complement of $V_{U'}$) that contains $V^{Q_0} = \neg\boxminus_{Q_0}$ and is contained in $V^{Q''}$. Hence, the closure of V^{Q_0} is contained in $V^{Q''}$, which gives

$$\partial\boxminus_{Q_0} = \partial V^{Q_0} \subset \overline{V^{Q_0}} \subset V^{Q''}.$$

On the other hand, let U be the set of quads $Q \in \mathcal{Q}_D$ satisfying $Q' < Q < Q''$, then

$$\partial\boxminus_{Q_0} \subset \boxminus_{Q_0} \subset V_U.$$

Now observe that $V_U \subset \boxminus_{Q'}$ while $V^{Q''} = \neg\boxminus_{Q''}$. We have shown that the open set $V_U \cap V^{Q''}$ contains $\partial\boxminus_{Q_0}$. The portmanteau theorem (see, e.g., [Dud89, Thm. 11.1.1]) therefore gives

$$\begin{aligned}\mu_0(\partial\boxminus_{Q_0}) \le \mu_0(V_U \cap V^{Q''}) &\le \liminf_{|\eta|\to 0} \mu_\eta(V_U \cap V^{Q''}) \\ &\le \liminf_{|\eta|\to 0} \mu_\eta(\boxminus_{Q'} \Delta \boxminus_{Q''})\,.\end{aligned}$$

Since ϵ was arbitrary, the result now follows from (5.1) □

COROLLARY 5.2. *If $\mathcal{M} \subset \mathcal{H}_D$ is in the Boolean algebra generated by finitely many of the events $\boxminus_Q$, $Q \in \mathcal{Q}_D$ (i.e., it can be expressed using finitely many $\boxminus_Q$ and the operations of union, intersection and taking complements), then $\mu_0(\mathcal{M}) = \lim_{|\eta|\to 0} \mu_\eta(\mathcal{M})$.*

PROOF. Lemma 5.1 implies that $\mu_0(\partial\mathcal{M}) = 0$. Thus, the corollary follows from the portmanteau theorem [Dud89, Thm. 11.1.1] and the weak convergence of μ_η to μ_0. □

PROOF OF THEOREM 1.19. Clearly, $\mathcal{F}_{D\setminus\alpha} = \vee_j \mathcal{F}_{D_j}$. It therefore remains to prove the left hand equality stated in the theorem. Note that we work up to sets of μ_0-measure zero.

Take a smooth (given by a diffeomorphism) quad $Q_0 \in \mathcal{Q}_D$. By Proposition 4.1 and Corollary 5.2, we have

$$\boxminus_{Q_0} \in \mathcal{F}_{D\setminus\alpha} \,. \tag{5.2}$$

Since such collection of quads is dense in $\mathcal{Q}_D$, Theorem 1.19 follows from Proposition 1.13.2. □

APPENDIX A: CONTINUITY OF CROSSING EVENTS

In this Section, we apply Russo-Seymour-Welsh techniques to deduce estimates of various crossing probabilities. We start by showing that crossing events are stable under small perturbations of quads.

LEMMA A.1. *There exist a positive function $\Delta_c(\delta, d)$, such that*

$$\lim_{\delta\to 0} \Delta_c(\delta, d) = 0 \text{ for any fixed } d \,,$$

and the following estimates hold. Let Q be a quad. Let d_j, $j = 0, 1$, be the infimum diameter of any path in $[Q]$ connecting $\partial_j Q$ and $\partial_{j+2} Q$, and define $d := \max\{d_0, d_1\}$. Fix some $\delta < d/2$. Let Q' be another quad, satisfying at least one of the following conditions, see Figures A.1, A.2 for a graphical interpretation.

(1). $[Q'] = [Q]$, $\partial_0 Q' = \partial_0 Q$, $\partial_1 Q' = \partial_1 Q$, and there is a path $\gamma \subset [Q]$ of diameter at most δ that separates $\{Q(1,1), Q'(1,1)\}$ from $\partial_0 Q \cup \partial_1 Q$ inside $[Q]$.

(2). $[Q'] \subset [Q]$, $\partial_0 Q' = \partial_0 Q$, $\partial_1 Q' \subset \partial_1 Q$, $\partial_3 Q' \subset \partial_3 Q$, and each point on $\partial_2 Q'$ can be connected to $\partial_2 Q$ by a path $\alpha \subset [Q]$ with $\operatorname{diam}(\alpha) \le \delta$.

(3). $[Q'] \subset [Q]$, $\partial_0 Q' \subset \partial_0 Q$, $\partial_1 Q' = \partial_1 Q$, $\partial_2 Q' \subset \partial_2 Q$, and each point on $\partial_3 Q'$ can be connected to $\partial_3 Q$ by a path $\alpha \subset [Q]$ with $\operatorname{diam}(\alpha) \le \delta$.

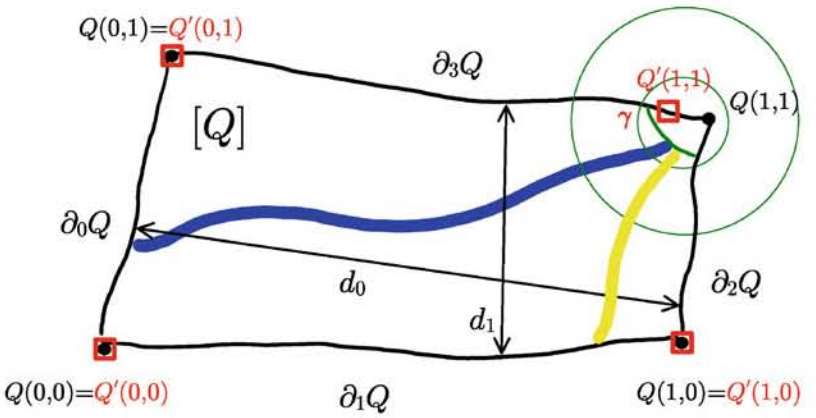


FIG A.1. *Case (1): quads Q and Q' differ by only one vertex, separated from the other vertices by a short cut γ. Configurations with only one of two quads crossed have an open crossing from γ to $\partial_0 Q$ and a dual closed crossing from γ to $p_1 Q$, making their probability small.*

Then for every $|\eta| < \delta$ we have

$$\mu_\eta(\boxminus_Q \Delta \boxminus_{Q'}) \;<\; \Delta_c(\delta, d)\ .$$

For site percolation on the triangular lattice, there is a short proof based on Cardy's formula. The proof below is a simple application of the RSW estimate (1.2) and the "lowest crossing" concept.

PROOF. First we deal with Case (1) in the statement and prove the estimate with Δ_c equal to Δ_1 from Assumption (1.2). Suppose first that $d = d_0$. The event $\boxminus_Q \Delta \boxminus_{Q'}$ is contained in the event that there is a percolation cluster meeting γ and $\partial_0 Q$. The latter has an open crossing of the annulus $A(Q(1,1), \delta, d_0)$, which by the RSW estimate (1.2) has probability at most $\Delta_1(\delta, d)$, and we are done with Case (1). If $d = d_1$, a similar argument shows the symmetric difference between the event of a closed crossing from $\partial_1 Q$ to $\partial_3 Q$ in $[Q]$ and the corresponding event in Q' is bounded by $\Delta_1(\delta, d_1)$. Duality shows that the latter symmetric difference is the same as $\boxminus_Q \Delta \boxminus_{Q'}$, which proves Case (1).

We now deal with Case (2), when clearly $\boxminus_Q \Delta \boxminus_{Q'} = \boxminus_{Q'} \setminus \boxminus_Q$. We start by estimating the crossing probability for Q'. Recall our convention that $\Delta(r, R) = 1$ for $r \geq R$ (i.e. when annulus in question is empty). Let x be some point on the path τ of diameter d_1, connecting $\partial_1 Q$ to $\partial_3 Q$. Any crossing of Q' has diameter at least $d_0 - \delta$ and passes within distance δ of τ, so in particular it crosses an annulus $A(x, d_1 + \delta, (d_0 - \delta)/2)$ if well defined,

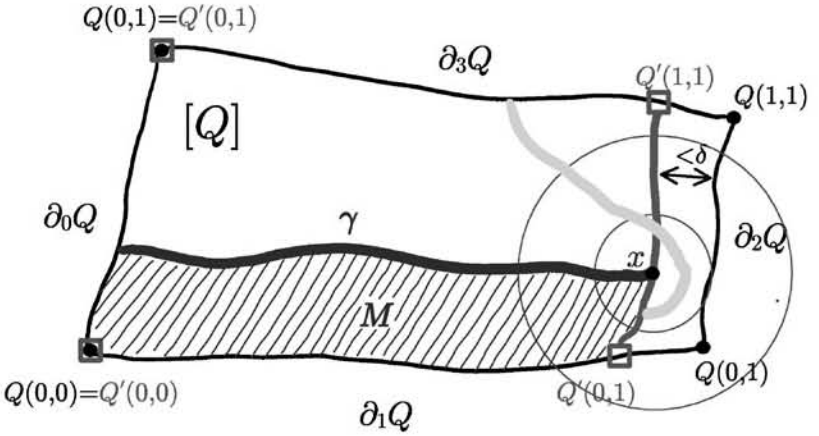


FIG A.2. *Case (2): side $\partial_2 Q'$ is close to the side $\partial_2 Q$. We take the lowest crossing γ of Q', landing at x. A dual closed crossing prevents it from landing on $\partial_2 Q$, making the probability small. Case (3) is symmetric.*

and by the RSW estimate (1.2) we have

$$\mu_\eta(\boxminus_{Q'}) \;\le\; \Delta_1\left(d_1+\delta, \frac{d_0-\delta}{2}\right), \tag{A.1}$$

Next we cut $\partial_2 Q'$ by a point z into two parts, σ_1 and σ_3, so that σ_j cannot be connected to $\partial_j Q$ by a path of diameter less than $d_1/2$ inside Q. If $\boxminus_{Q'}$ occurs, there is a crossing to at least one of σ_1 and σ_3. We will work with configurations with a crossing landing on σ_3, the other case being symmetric, with the lowest crossing replaced by the uppermost, so the total estimate will be double of what we obtain.

Consider a percolation configuration ω. On the event $\boxminus_{Q'}$, let γ be the "lowest" ω-crossing of Q' (in the sense that it is the closest to $\partial_1 Q'$, i.e. separates $\partial_1 Q'$ from all other ω-crossings of Q' within $[Q']$), and let x be the endpoint of γ on $\partial_2 Q'$. Our assumption implies that $x \in \sigma_3$, or there would be a lower crossing.

We now estimate $\mu_\eta(\neg\boxminus_Q \mid \boxminus_{Q'}, \gamma, x \in \sigma_3)$. Let M be the connected component of $[Q']\setminus\gamma$ which has $\partial_1 Q'$ on its boundary. The event $\neg\boxminus_Q$ would imply that γ cannot be connected to $\partial_2 Q$ by a crossing inside $[Q]\setminus M$. Hence there is a dual closed crossing from $\partial_3 Q$ to $(\sigma_3 \cap \bar{M}) \cup \partial_1 Q$, in particular crossing the annulus $A(x, \delta, d_1/2)$ inside $[Q] \setminus M$.

The lowest crossing γ depends only on the configuration inside M, so the restriction of ω to $[Q] \setminus M$ is unbiased. Therefore, the conditional probability of the mentioned annulus crossing is can be bound by the RSW

estimate (1.2) and we conclude that

$$\mu_\eta(\neg \boxminus_Q \mid \boxminus_{Q'}, \gamma, x \in \sigma_3) \;\leq\; \Delta_1\,(\delta, d_1/2)\;. \tag{A.2}$$

Now we are ready to prove the estimate, working out separately three possibilities:

(i) $d = d_1$,
(ii) $d = d_0 > \delta \geq d_1$.
(iii) $d = d_0 > d_1 > \delta$,

When (i) occurs, we majorate the probability of the event in question by the maximum of its conditional probability, estimated by (A.2), to arrive at

$$\mu_\eta(\boxminus_{Q'} \setminus \boxminus_Q) \;\leq\; 2\,\Delta_1\left(\delta, \frac{d_1}{2}\right) \;=\; 2\,\Delta_1\left(\delta, \frac{d}{2}\right)\;. \tag{A.3}$$

When (ii) occurs, we use the estimate (A.1):

$$\mu_\eta(\boxminus_{Q'} \setminus \boxminus_Q) \;\leq\; \mu_\eta(\boxminus_{Q'}) \;\leq\; \Delta_1\left(d_1 + \delta, \frac{d_0 - \delta}{2}\right) \;\leq\; \Delta_1\left(2\delta, \frac{d}{4}\right)\;. \tag{A.4}$$

Finally, when (iii) occurs, we apply the total expectation law, using both estimates (A.2) and (A.1) along the way:

$$\begin{aligned}
\mu_\eta(\boxminus_{Q'} \setminus \boxminus_Q) \leq\; & \mathbf{E}\left[\mu_\eta(\neg \boxminus_Q \mid \boxminus_{Q'}, \gamma, x \in \sigma_3)\right] \\
& + \mathbf{E}\left[\mu_\eta(\neg \boxminus_Q \mid \boxminus_{Q'}, \gamma, x \in \sigma_1)\right] \\
\leq\; & 2\;\mu_\eta(\boxminus_{Q'})\;\Delta_1\left(\delta, \frac{d_1}{2}\right) \\
\leq\; & 2\;\Delta_1\left(d_1 + \delta, \frac{d_0 - \delta}{2}\right)\;\Delta_1\left(\delta, \frac{d_1}{2}\right) \\
\leq\; & 2\;\Delta_1\left(2d_1, \frac{d}{4}\right)\;\Delta_1\left(\delta, \frac{d_1}{2}\right). \qquad \text{(A.5)}
\end{aligned}$$

To prove Lemma in Case (2), we have to show that for fixed d and any $\epsilon > 0$, the right hand side of the estimates (A.3), (A.4) and (A.5) is smaller than ϵ, if δ is small enough. For estimates (A.3) and (A.4) this follows directly from Assumption 1.2. For the remaining (A.5), let $\rho > 0$ be such that for $r < \rho$ we have $\Delta_1\left(2r, \frac{d}{4}\right) < \epsilon/2$. Then if $d_1 \leq \rho$, then the right hand side of (A.5) can be bounded by

$$2\;\Delta_1\left(2d_1, \frac{d}{4}\right)\;\Delta_1\left(\delta, \frac{d_1}{2}\right) \;\leq\; 2\;\Delta_1\left(2d_1, \frac{d}{4}\right) \;<\; \epsilon\;,$$

we are done. Otherwise $d_1 > \rho$ and the right hand side of (A.5) can be bounded by

$$2\ \Delta_1\left(2d_1, \frac{d}{4}\right)\ \Delta_1\left(\delta, \frac{d_1}{2}\right)\ \leq\ 2\ \Delta_1\left(\delta, \frac{\rho}{2}\right)\ ,$$

which tends to zero with δ by Assumption 1.2, and we finish the proof in Case (2).

The proof for Case (3) is easily obtained by considering the dual closed crossing from ∂_1 to ∂_3 and applying Case (2). The details are left to the reader. □

The following Lemma shows that it is unlikely to see three crossings approaching the same boundary point. This would make the crossing event unstable under boundary perturbation, so in principle the Lemma can be deduced from the previous one, but instead we give a self-contained proof. The Lemma would follow from considering the three arm event in *half-annuli*, for whose probability Aizenman proposed an argument to be comparable to $(r/R)^2$. We will use a version of his reasoning along with RSW techniques to show a weaker estimate, sufficient for our purposes.

LEMMA A.2. *Let Q be a quad with smooth sides, two opposite being labeled β and β', and with the tiles on each of the two other sides ν and ν' possibly declared all open or all closed. Then there is a function $\Delta_Q(\delta)$, tending to zero with δ and such that the following estimate holds.*

Denote by $S(\theta)$ the event that there are three crossings with some prescribed alternating order (say closed, open, and closed) inside $[Q]$ between the set $\theta \subset \beta'$ and β. *Then for $|\eta| < \delta$ we have*

$$\mu_\eta\left\{\exists\theta \subset \beta' \,:\, S(\theta),\ \operatorname{diam}(\theta) < \delta\right\}\ \leq\ \Delta_Q(\delta)\ .$$

PROOF. Fix $\epsilon > 0$. We shall prove that for sufficiently small δ the probability of the event in question is less than ϵ.

Since crossings can in principle use tiles on the sides ν and ν' (which are declared open or closed), we have to treat θ's near the corners of Q separately. Let z_1 and z_2 be the endpoints of β'. Since tiles on each of the arcs ν and ν' can be used by at most one of the three crossings, event $S(\theta)$ implies existence of at least one crossing from θ to β inside $[Q]$, not touching the sides. By Assumption (1.2), there is $r \in (0, \delta)$ such that probability to have a crossing between $\cup_i B(z_i, 2r)$ and β is smaller than $\epsilon/2$:

$$\mu_\eta\left\{\exists\theta \subset \cup_i B(z_i, 2r) \,:\, S(\theta)\right\}\ \leq\ \frac{\epsilon}{2}\ .$$

Set $\beta'' := \beta' \setminus \cup_i B(z_i, r)$, then to prove Lemma it remains to show that

$$\mu_\eta \{\exists \theta \subset \beta'' : S(\theta), \ \operatorname{diam}(\theta) < \delta\} \ \leq \ \frac{\epsilon}{2} . \tag{A.6}$$

Using Assumption (1.2), choose $\rho \in (0, r/4)$ such that $\Delta_1(2\rho, r/2) < 1/2$.

Cover β'' by a finite number $n = n(\rho, \beta'')$ of overlapping arcs θ_j of diameter at most ρ. Let $\delta < \rho$ be so small that $\Delta_1(4\delta, r) < \epsilon/(12n)$. Since β' is smooth, and decreasing δ if necessary, we can cover each arc θ_j by arcs θ_j^l of length 3δ, overlapping at most thrice, and such that any set $\theta \subset \beta''$ of diameter less than δ is entirely contained within one of the arcs θ_j^l.

To deduce (A.6) it is sufficient to show that, for a fixed j,

$$\sum_l \mu_\eta \left[S\left(\theta_j^l\right) \right] \ \leq \ \frac{\epsilon}{2n} . \tag{A.7}$$

Fix j and let $B(z, \rho)$ be a radius ρ ball, containing θ_j. The point z_1 splits $\partial Q \setminus \theta_j$ into two arcs: $\lambda' \subset \beta'$ and a $\lambda \supset \beta$. Both start at endpoints of θ_j inside $B(z, \rho)$ and end at least r-away.

Observe that event $S\left(\theta_j^l\right)$ is contained in the event $S'\left(\theta_j^l\right)$ that *there are three alternating crossings inside* $[Q]$ *between* θ *and* $\beta \cup \nu \cup \nu'$.

Condition on the event $S'\left(\theta_j^l\right)$, and, starting from z_2, denote the three such crossings, closest to z_2, by γ_1, γ_2, γ_3. Namely, let γ_1 be the closed crossing between θ and $\beta \cup \nu \cup \nu'$ inside $[Q]$, closest to z_2. Let M_1 be the component of connectivity of $[Q] \setminus \gamma_1$, not containing z_2. Then γ_1 depends only on the percolation configuration in $[Q] \setminus M_1$.

Now, take γ_2 be the open crossing between θ and $\beta \cup \nu \cup \nu'$ inside M_1, closest to γ_1. Let M_2 be the component of connectivity of $M_2 \setminus \gamma_2$, not containing γ_1. Then γ_1 and γ_2 depend only on the percolation configuration in $[Q] \setminus M_2$.

We define γ_3 and M_3 similarly, and observe that γ_1, γ_2, γ_3 depend only on the percolation configuration in $[Q] \setminus M_3$. Therefore, given M_3 and the restriction $\omega_\llcorner([Q] \setminus M_3)$ (i.e. the restriction of percolation configuration ω to $[Q] \setminus M_3$), the conditional law of the restriction of ω to M_3 is unbiased.

Let $S^* = S^*(M_3)$ be the event that *there is a closed crossing* γ_4 *between* λ *and* γ_3 *inside* M_3. Using that the restriction of ω to M_3 is unbiased, we can then write

$$\mu_\eta \left[S^* \,|\, M_3, \omega_\llcorner[Q] \setminus M_3 \right] \ = \ \mu_\eta \left[S^* \right] \ > \ 1 - P_\eta^1(z, \rho, r) \ > \ \frac{1}{2} . \tag{A.8}$$

Above we use that (in an unbiased percolation configuration), if there is no open crossing of the annulus $A(z, \rho, r)$, then by duality there is a closed

circuit (a path going around the annulus), whose intersection with M_3 would then contain a crossing required for S^*.

Note that when $S'\left(\theta_j^l\right)$ and S^* occur, the crossing γ_3 from $(S\left(\theta_j^l\right))$ and the crossing γ_4 (from S^*) together give a closed crossing between θ and λ, and so the following event $S''\left(\theta_j^l\right)$ occurs: *there is a closed crossing γ_3' from θ to λ, an open crossing from γ_2' from θ to $\beta \cup \nu \cup \nu'$, and a closed crossing from γ_1' from θ to $\beta \cup \nu \cup \nu'$*. Then the estimate (A.8) can be rephrased as

$$\mu_\eta \left[S''\left(\theta_j^l\right) \mid S'\left(\theta_j^l\right) \right] \;\geq\; \frac{1}{2}\,,$$

and therefore

$$\mu_\eta \left[S\left(\theta_j^l\right) \right] \;\leq\; \mu_\eta \left[S'\left(\theta_j^l\right) \right] \;\leq\; 2\,\mu_\eta \left[S''\left(\theta_j^l\right) \right]\,. \tag{A.9}$$

Consider some percolation configuration ω in $[Q]$, and assume that $S''\left(\theta_j^l\right)$ holds for some $\theta_j^l \subset \theta_j$.

Choose inside Q a closed crossing $\gamma_3'(\omega)$ between λ' and θ_j; and an open crossing $\gamma_2'(\omega)$ between λ which are the closest to each other (and the point z_1 separating λ' from λ). Denote by $L = L(\omega)$ the union of tiles in γ_2', γ_3', and in the part of $[Q]$ between them.

Then γ_2' and γ_3' depend only on the restriction of ω to L, and the rest is unbiased. Moreover, if γ_2' ends at a tile on θ_j, then γ_3' ends at a neighboring tile on θ_j — otherwise using duality reasoning we can choose two closer crossings. Denote by x the only common vertex of these two tiles lying on the boundary of N. Note that each configuration ω has a unique such point $x = x(\omega)$ and it depends only on the restriction of ω to N.

Now for $S\left(\theta_j^l\right)$ to occur, besides γ_2' and γ_3' we must also have a closed crossing γ_1' from θ to $\beta \cup \nu \cup \nu'$. The crossing γ_1' necessarily lies in $[Q] \setminus N$, which is unbiased on $\omega \llcorner N$. Note also that γ_1' contains a crossing of the annulus $A(x, 2\delta, r)$, whose center x is a random point depending on $\omega \llcorner L$ only. Summing it up, we can estimate

$$\mu_\eta \left\{ \omega :\, S''\left(\theta_j^l\right) \mid x(\omega) = x \right\} \;\leq\; P_\eta^1\left(x, 4\delta, r\right) \leq\; \Delta_1(4\delta, r) < \frac{\epsilon}{12n}\,.$$

Since every point x is covered by at most three arcs θ_j^l, we recall (A.9) and

conclude that

$$\begin{aligned}\sum_l \mu_\eta\left[S\left(\theta_j^l\right)\right] &\le \sum_l 2\mu_\eta\left[S''\left(\theta_j^l\right)\right] \\ &= \sum_x \sum_l 2\mu_\eta\left\{\omega : S\left(\theta_j^l\right) \mid x(\omega) = x\right\} \mu_\eta\left\{\omega : x(\omega) = x\right\} \\ &\le \sum_x \left(2 \cdot 3 \cdot \frac{\epsilon}{12n}\right) \mu_\eta\left\{\omega : x(\omega) = x\right\} \\ &\le \frac{\epsilon}{2n} \sum_x \mu_\eta\left\{\omega : x(\omega) = x\right\} \le \frac{\epsilon}{2n},\end{aligned}$$

so estimate (A.7) and the Lemma follow. □

APPENDIX B: MULTI-SCALE BOUND ON THE FOUR-ARM EVENT ON $\mathbb{Z}^2$ (BY CHRISTOPHE GARBAN)

In this appendix, we give a proof of Assumption (1.3) for critical bond percolation on the square lattice. Namely:

LEMMA B.1. *For critical bond percolation on the square lattice with mesh $|\eta|$, there is a positive ϵ such that the probability of a four arm event satisfies*

$$P_\eta^4(z, r, R) \le \text{const} \left(\frac{r}{R}\right)^{1+\epsilon}, \tag{B.1}$$

whenever $|\eta| < r$.

The case $r = |\eta|$ can be extracted from [Kes87] as well as [BKS99] or [SS10]. Note that the above multi-scale generalization is not a consequence of the so-called "quasi-multiplicativity" property. If one wanted to use this property, the problem would more or less boil down to the **existence** of the *four-arm* critical exponent. Since its existence is still open, we believe that Lemma (B.1) cannot be extracted directly from the above mentioned papers.

In the papers [BKS99, SS10], the general idea which enables one to prove that points are unlikely to be pivotal is the observation that a macroscopic crossing event for critical percolation is asymptotically uncorrelated with the Majority Boolean function defined on the same set of bits. This asymptotic uncorrelation can be understood using an exploration path which will compute the percolation event while giving little information on the Majority event. A careful analysis then shows that this decorrelation is possible only if points are unlikely to be pivotal for the percolation event. This heuristical program has been carried out in a very convenient manner in [OS07].

The present proof has the same flavor except that one now looks at the correlation of a percolation event in a domain of diameter $O(R)$ with a *two-layers majority function*: namely the bits of this Majority function are now indexed by the $O(R^2/r^2)$ r-squares included in the domain and for each such r-square, its corresponding bit ± 1 depends on how "connected" the percolation configuration is within this square. This setup allows to "interpolate" the above program from the macroscopic scale R to the mesoscopic scale r.

PROOF. We leave out a few details, which are easy to fill in for readers familiar with the applications of the RSW theory, similar to [Kes87, SW01, GPS10a].

After rescaling and, if needed, changing the radii by bounded factors, we can assume without loss of generality that the lattice mesh is 1 and r, R are positive integers.

Denote by Q an $R \times R$ square, and cut the concentric $\frac{1}{3}Q$ square into $r \times r$ squares, denoted by Q_j, with $j = 1, \dots, \left(\frac{R}{3r}\right)^2$. Denote by $X = 2 \cdot 1_{\boxminus_Q} - 1$ the "crossing random variable" equal to 1 when Q is crossed and -1 otherwise. Given a percolation configuration ω, we say that Q_j is pivotal for X, if altering ω so that all the bonds in Q_j are open, and so that all the bonds in Q_j are closed yields two different values of X. Note that Q_j is pivotal for X if and only if then there are two open arms connecting ∂Q_j to $\partial_0 Q$ and $\partial_2 Q$, and two dual closed arms connecting ∂Q_j to $\partial_1 Q$ and $\partial_3 Q$. By arm separation properties similar to ones used in [SW01],

$$P^4(r, R) \asymp \mathbf{P}\left[Q_j \text{ pivotal for } X\right], \tag{B.2}$$

and so it is sufficient to estimate the latter probability, or its sum over all Q_j's.

Let S_j be an annulus $(1-\delta)Q_j \setminus (1-2\delta)Q_j$ where αQ_j denotes a copy shrank of Q_j by a factor of α (if r is not exactly divisible by α, the copy shrank is understood modulo integer parts). The fixed parameter $\delta > 0$ will be chosen later. Let S_j^* denote the same annulus but on the dual lattice: i.e. $S_j^* = S_j + (1/2, 1/2)$.

Define a random variable C_j equal to

- 1, if there is an open circuit in S_j and no dual-closed circuit in S_j^*,
- -1, if there is a dual-closed circuit in S_j^* and no open circuit in S_j,
- 0, otherwise.

Note that by symmetry and by the RSW theory, up to the constant depending on δ,

$$\mathbf{E}\left[C_j\right] = 0\,, \quad \mathbf{E}\left[C_j^2\right] \asymp 1\,. \tag{B.3}$$

Now, let us argue that if $\delta > 0$ is chosen small enough, one has

$$\mathbf{E}[XC_j] \asymp \mathbf{P}[Q_j \text{ pivotal for } X] \quad (\asymp P^4(r,R)). \tag{B.4}$$

Indeed, one has

$$\mathbf{E}[XC_j] = \mathbf{P}[Q_j \text{ pivotal for } X]\, \mathbf{E}[XC_j \mid Q_j \text{ pivotal for } X], \tag{B.5}$$

since, conditioned on the event that Q_j is not pivotal for X, C_j is independent of X and is such that its (conditional) expectation is still 0. Therefore it remains to bound from below $\mathbf{E}[XC_j \mid Q_j \text{ pivotal for } X]$. Let $\rho = \rho(\delta) > 0$ be such that $\mathbf{P}[C_j = 1] = \mathbf{P}[C_j = -1] > \rho$.

$$\begin{aligned}\mathbf{E}[XC_j \mid Q_j \text{ pivotal for } X] =& \mathbf{P}[C_j = 1]\, \mathbf{E}[X \mid C_j = 1 \,\&\, Q_j \text{ pivotal for } X] \\ &- \mathbf{P}[C_j = 1]\, \mathbf{E}[X \mid C_j = -1 \,\&\, Q_j \text{ pivotal for } X].\end{aligned}$$

Now, again by arm separation properties similar to ones used in [SW01, SS10], and using the important fact that the event $\{C_j = 1\}$ is increasing which enables to use FKG for the given conditional law inside Q_j, one can see that, if δ is chosen small enough,

$$\mathbf{E}[X \mid C_j = 1 \text{ and } Q_j \text{ pivotal for } X] > \frac{1}{4}.$$

One has the opposite bound for the term conditioned on $\{C_j = -1\}$. All together this gives

$$\mathbf{E}[XC_j \mid Q_j \text{ pivotal for } X] > \frac{\rho}{2},$$

which implies the desired estimate (B.4) if δ is chosen to be small enough.

Consider Q with open boundary conditions on the side $\partial_0 Q$ and dual closed on the complementary three sides. Let γ be the interface running between the two ends of the side $\partial_0 Q$ and separating the open cluster rooted on it from the dual closed cluster rooted on the three other sides.

Denote by Y_j the event that γ intersects Q_j, as well as its indicator function. Note that, by the RSW theory,

$$\mathbf{E}[Y_j] \lesssim \left(\frac{r}{R}\right)^{2\epsilon}, \tag{B.6}$$

for some positive ϵ. Note that the interface γ drawn from $Q(0,0)$ until some stopping time depends only on the percolation configuration in the immediate neighborhood of the drawn part. Drawing γ until the first time it hits

∂Q_j, we conclude that Y_j is independent from the inside of Q_j and hence from C_j.

Note also that, if Q_j is pivotal for X, then there are four alternating arms connecting ∂Q_j to four sides of Q which forces the interface γ to intersect Q_j, and Y_j to occur. Therefore, similarly as for the estimate (B.5) above, one can check that

$$\mathbf{E}\,[XC_j] \;=\; \mathbf{E}\,[XC_jY_j]\,, \tag{B.7}$$

which combined with (B.4) gives

$$\mathbf{P}\,[Q_j \text{ pivotal for } X] \;\asymp\; \mathbf{E}\,[XC_jY_j]\;. \tag{B.8}$$

Summing over all Q_j's, we use Cauchy-Schwarz inequality to write

$$\sum_j \mathbf{E}\,[XC_jY_j] \;=\; \mathbf{E}\left[X\textstyle\sum_j C_jY_j\right] \;\le\; \sqrt{\mathbf{E}\,[X^2]\cdot\mathbf{E}\left[\left(\sum_j C_jY_j\right)^2\right]}$$

$$= \sqrt{\mathbf{E}\,[1]\cdot\mathbf{E}\left[\textstyle\sum_{i,j} C_iY_iC_jY_j\right]} \;=\; \sqrt{\textstyle\sum_{i,j}\mathbf{E}\,[C_iY_iC_jY_j]} \tag{B.9}$$

In the last sum, non-diagonal terms vanish. Indeed, let $i \neq j$ and stop the interface γ the first time it touched both squares Q_i and Q_j, or when it ends.

Denote by Q_k the first square of Q_i, Q_j to be hit (or Q_i if none was) and by Q_l the other one, so that (k,l) is a superposition of (i,j). Let W be the percolation configuration in the immediate neighborhood of the interface so far, and in Q_k.

Then W determines Y_i, Y_j, C_k, while being independent of C_l. Thus by the identity in (B.3),

$$\mathbf{E}\,[C_iY_iC_jY_j|W] \;=\; \mathbf{E}\,[Y_iY_jC_k|W]\cdot\mathbf{E}\,[C_l] \;=\; 0.$$

By the total expectation formula we conclude that, for $i \neq j$,

$$\mathbf{E}\,[C_iY_iC_jY_j] \;=\; 0\;.$$

Thus we can continue (B.9), leaving only the diagonal terms, and use (B.6) and (B.3) to write

$$\ldots = \sqrt{\textstyle\sum_j \mathbf{E}\left[C_j^2Y_j^2\right]} \;=\; \sqrt{\textstyle\sum_j \mathbf{E}\left[C_j^2\right]\cdot\mathbf{E}\left[Y_j^2\right]}$$

$$\lesssim \sqrt{\textstyle\sum_j 1\cdot\left(\frac{r}{R}\right)^{2\epsilon}} \;=\; \left(\frac{r}{R}\right)^{\epsilon-1}\;.$$

Combining this with (B.2), (B.8), (B.9) we conclude

$$\begin{aligned} P^4(r,R) &\lesssim \left(\frac{3r}{R}\right)^2 \sum_j \mathbf{P}\left[Q_j \text{ pivotal for } X\right] \\ &\lesssim \left(\frac{r}{R}\right)^2 \left(\frac{r}{R}\right)^{\epsilon-1} = \left(\frac{r}{R}\right)^{1+\epsilon}, \end{aligned}$$

proving the lemma. □

REFERENCES

[AB99] M. Aizenman and A. Burchard. Hölder regularity and dimension bounds for random curves. *Duke Math. J.*, 99(3):419–453, 1999.

[APS10] Kari Astala, István Prause, and Stanislav Smirnov. Holomorphic motions and harmonic measure. Preprint, 2010.

[BKS99] Itai Benjamini, Gil Kalai, and Oded Schramm. Noise sensitivity of Boolean functions and applications to percolation. *Inst. Hautes Études Sci. Publ. Math.*, 90:5–43 (2001), 1999.

[BR06] Béla Bollobás and Oliver Riordan. *Percolation.* Cambridge University Press, 2006.

[BR10] Bela Bollobas and Oliver Riordan. Percolation on self-dual polygon configurations. *Preprint, arXiv:1001.4674*, 2010.

[Car92] John L. Cardy. Critical percolation in finite geometries. *J. Phys. A*, 25(4):L201–L206, 1992.

[CFN06] Federico Camia, Luiz Renato G. Fontes, and Charles M. Newman. Two-dimensional scaling limits via marked nonsimple loops. *Bull. Braz. Math. Soc. (N.S.)*, 37(4):537–559, 2006.

[CN06] Federico Camia and Charles M. Newman. Two-dimensional critical percolation: the full scaling limit. *Comm. Math. Phys.*, 268(1):1–38, 2006.

[Dud89] Richard M. Dudley. *Real analysis and probability.* The Wadsworth & Brooks/Cole Mathematics Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1989.

[Fed69] Herbert Federer. *Geometric measure theory.* Die Grundlehren der mathematischen Wissenschaften, Band 153. Springer-Verlag New York Inc., New York, 1969.

[Fel62] J. M. G. Fell. A Hausdorff topology for the closed subsets of a locally compact non-Hausdorff space. *Proc. Amer. Math. Soc.*, 13:472–476, 1962.

[GPS10a] Christophe Garban, Gábor Pete, and Oded Schramm. The Fourier spectrum of critical percolation. *Acta Math.*, 205(1):19–104, 2010.

[GPS10b] Christophe Garban, Gábor Pete, and Oded Schramm. The scaling limit of the Minimal Spanning Tree — a preliminary report. In P. Exner, editor, *XVIth International Congress on Mathematical Physics, Prague, 2009*, pages 475–480. World Sci. Publ., Singapore, 2010.

[GPS10c] Christophe Garban, Gábor Pete, and Oded Schramm. Pivotal, cluster and interface measures for critical planar percolation. Preprint, ArXiv:1008.1378, 2010.

[Gri99] Geoffrey Grimmett. *Percolation*, volume 321 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, second edition, 1999.

[GRS08] C. Garban, S. Rohde, and O. Schramm. Continuity of the SLE trace in simply connected domains. Israel J. Math., to appear. Preprint, ArXiv:0810.4327, 2008.

[Kel75] John L. Kelley. *General topology.* Springer-Verlag, New York, 1975. Reprint of the 1955 edition [Van Nostrand, Toronto, Ont.], Graduate Texts in Mathematics, No. 27.

[Ken09] Richard Kenyon. Lectures on dimers. In *Statistical mechanics*, volume 16 of *IAS/Park City Math. Ser.*, pages 191–230. Amer. Math. Soc., Providence, RI, 2009.

[Kes80] Harry Kesten. The critical probability of bond percolation on the square lattice equals 1/2. *Comm. Math. Phys.*, 74(1):41–59, 1980.

[Kes82] Harry Kesten. *Percolation theory for mathematicians*, volume 2 of *Progress in Probability and Statistics.* Birkhäuser Boston, Mass., 1982.

[Kes87] Harry Kesten. Scaling relations for 2D-percolation. *Comm. Math. Phys.*, 109(1):109–156, 1987.

[KS10] Antti Kemppainen and Stanislav Smirnov. Conformal invariance in random cluster models. III. Full scaling limit. In preparation, 2010.

[Lan05] R. P. Langlands. The renormalization fixed point as a mathematical object. In *Twenty years of Bialowieza: a mathematical anthology*, volume 8 of *World Sci. Monogr. Ser. Math.*, pages 185–216. World Sci. Publ., Hackensack, NJ, 2005.

[LL94] R. P. Langlands and M.-A. Lafortune. Finite models for percolation. In *Representation theory and analysis on homogeneous spaces (New Brunswick, NJ, 1993)*, volume 177 of *Contemp. Math.*, pages 227–246. Amer. Math. Soc., Providence, RI, 1994.

[LPSA94] Robert Langlands, Philippe Pouliot, and Yvan Saint-Aubin. Conformal invariance in two-dimensional percolation. *Bull. Amer. Math. Soc. (N.S.)*, 30(1):1–61, 1994.

[LSW02] Gregory F. Lawler, Oded Schramm, and Wendelin Werner. One-arm exponent for critical 2D percolation. *Electron. J. Probab.*, 7:no. 2, 13 pp. (electronic), 2002.

[Mat95] Pertti Mattila. *Geometry of sets and measures in Euclidean spaces*, volume 44 of *Cambridge Studies in Advanced Mathematics.* Cambridge University Press, Cambridge, 1995. Fractals and rectifiability.

[Nie84] B. Nienhuis. Critical behavior of two-dimensional spin models and charge asymmetry in the Coulomb gas. *Journal of Statistical Physics*, 34(5):731–761, 1984.

[Nol08] Pierre Nolin. Near-critical percolation in two dimensions. *Electron. J. Probab.*, 13:no. 55, 1562–1623, 2008.

[OS07] Ryan O'Donnell and Rocco A. Servedio. Learning monotone decision trees in polynomial time. *SIAM J. Comput.*, 37(3):827–844 (electronic), 2007.

[Rud73] Walter Rudin. *Functional analysis.* McGraw-Hill Book Co., New York, 1973. McGraw-Hill Series in Higher Mathematics.

[She09] S. Sheffield. Exploration trees and conformal loop ensembles. *Duke Math. J*, 147(1):79–129, 2009.

[Smi01a] Stanislav Smirnov. Critical percolation in the plane. Preprint, arXiv:0909.4499, 2001.

[Smi01b] Stanislav Smirnov. Critical percolation in the plane: Conformal invariance, Cardy's formula, scaling limits. *C. R. Math. Acad. Sci. Paris*, 333(3):239–244, 2001.

[SS10] Oded Schramm and Jeffrey E. Steif. Quantitative noise sensitivity and exceptional times for percolation. *Ann. of Math. (2)*, 171(2):619–672, 2010.

[Ste09] Jeffrey E. Steif. A survey of dynamical percolation. In *Fractal geometry and stochastics IV. Proceedings of the 4th conference, Greifswald, Germany, September 8–12, 2008*, volume 61 of *Progress in Probability*, pages 145–174. Birkhäuser, Basel, 2009.

[SW01] Stanislav Smirnov and Wendelin Werner. Critical exponents for two-dimensional percolation. *Math. Res. Lett.*, 8(5-6):729–744, 2001.

[Tsi04a] Boris Tsirelson. Nonclassical stochastic flows and continuous products. *Probab. Surv.*, 1:173–298 (electronic), 2004.

[Tsi04b] Boris Tsirelson. Scaling limit, noise, stability. In *Lectures on probability theory and statistics*, volume 1840 of *Lecture Notes in Math.*, pages 1–106. Springer, Berlin, 2004.

[Tsi05] B. Tsirelson. Percolation, boundary, noise: an experiment. Preprint, Arxiv:math/0506269, 2005.

[TV98] B. S. Tsirelson and A. M. Vershik. Examples of nonlinear continuous tensor products of measure spaces and non-Fock factorizations. *Reviews in Mathematical Physics*, 10:81–145, 1998.

[Wer09] Wendelin Werner. Lectures on two-dimensional critical percolation. In *Statistical mechanics*, volume 16 of *IAS/Park City Math. Ser.*, pages 297–360. Amer. Math. Soc., Providence, RI, 2009.

[Wie81] John C. Wierman. Bond percolation on honeycomb and triangular lattices. *Adv. in Appl. Probab.*, 13(2):298–313, 1981.

Stanislav Smirnov
Section de Mathématiques, Université de Genève
2-4 rue du Lièvre, CP 64
1211 Genève 4, Switzerland
E-mail: stanislav.smirnov@unige.ch
URL: http://www.unige.ch/~smirnov