Martin Loebl · Jaroslav Nešetřil
Robin Thomas *Editors*

A Journey Through Discrete Mathematics

A Tribute to Jiří Matoušek

Springer

A Journey Through Discrete Mathematics

Martin Loebl • Jaroslav Nešetřil • Robin Thomas
Editors

A Journey Through Discrete Mathematics

A Tribute to Jiří Matoušek

Springer

Editors

Martin Loebl
Department of Applied Mathematics
Charles University
Praha, Czech Republic

Jaroslav Nešetřil
Computer Science Institute of Charles University
Charles University
Praha, Czech Republic

Robin Thomas
School of Mathematics
Georgia Institute of Technology
Atlanta, GA, USA

ISBN 978-3-319-44478-9 ISBN 978-3-319-44479-6 (eBook)
DOI 10.1007/978-3-319-44479-6

Library of Congress Control Number: 2017949385

Mathematics Subject Classification (2010): 00B15, 05C10, 05C15, 05C35, 05C50, 05C55, 05C62, 05C80, 05C85, 05D05, 05D10, 05D15, 05D40, 52A35, 52C10, 55S91, 68R05, 68R10, 68W40

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Introduction

Professor Jiří Matoušek passed away in March 2015 on the eve of his 52nd birthday. He left behind a large body of work, including over 170 research articles, eight books, and numerous unpublished lecture notes. This book is a celebration of his mathematical, pedagogical, and personal legacy.

Jirka was an excellent researcher whose work transcended individual boundaries of particular areas of mathematics and theoretical computer science. He excelled in every subject in which he was active. The book's content demonstrates Jirka's broad interests and influence. We hope that the carefully selected papers of this volume show the beauty and relevance of his scientific contribution in shaping mathematics and theoretical computer science research of today.

Jirka was an exceptional teacher as many of his colleagues and students can confirm. Over the years, he taught and shaped most of the basic courses offered by his home Department of Applied Mathematics, Charles University, Prague, and later also at ETH Zurich. Several of Jirka's courses, spiced with his particular sense of humor, were the basis of his excellent textbooks and monographs, which became the standard of scientific writing around the world. We tried to reflect the qualities of Jirka's style in this book.

Jirka was a great humanist. He used to say that mathematicians are useful for the society through their wisdom. We hope that this book bears a testimony of the kind wisdom of Jirka Matoušek. We believe that his legacy will be here for many years to come.

Prague, Czech Republic — Martin Loebl
Prague, Czech Republic — Jaroslav Nešetřil
Atlanta, GA, USA — Robin Thomas

Photo courtesy of ETH Zurich, Giulia Marthaler

Curriculum Vitae of Jiří Matoušek

March 10, 1963–March 9, 2015 Prague, Czechoslovakia/Czech Republic.

Academic History

- Charles University, Prague, Faculty of Mathematics and Physics, 1981–1986, Diploma, RNDr. degree (analogue of master's degree).
- CSc. degree (analogue of Ph.D.): Charles University, Prague, 1991.
- Habilitation ("docent" at Charles University) 1995.
- Dr.Sc. (higher doctorate) degree in mathematics, 1996.
- Professor at Charles University, 2000.
- Professor (35% position) at ETH Zurich, 2012.

Employment Record

Charles University, Prague 1986–2015 (1986–1997 doctoral study, 1987–1995 assistant professor of mathematics, 1995–2000 associate professor (docent), 2000–2015 full professor).
ETH Zurich 2012–2015 full professor (35 %).

Visiting Positions

- 1991, visiting assistant professor, Georgia Institute of Technology, 6 months.
- 1992, Humboldt research fellow, Freie Universität Berlin, 12 months.
- 1996–2011, ETH Zurich, visiting position every year, 3–4 months each year (6 months in 2011).

Other Longer Visits

Freie Universität Berlin 1989 (1 month), 1990 (7 weeks), and 1994 (7 weeks); DIMACS Center, Rutgers University, and Princeton University, 1989 (6 weeks), 1990 (2 weeks), 1993 (3 weeks), and 1994 (3 weeks); Macquarie University, Sydney, 1994 (3 weeks); Tel Aviv University, 1995 (2 months); University College London, 2001 (1 month); Univ. Santiago de Chile, 2003 (5 weeks); and Japanese Advanced Institute of Science and Technology, 2005 (5 weeks).

Publications

Published eight books (four of them with co-authors), over 170 original research papers, and several surveys.

Honors and Awards

- Several prizes in high-school and university mathematical competitions.
- Prize of the Czechoslovak Academy of Sciences (1986) for a joint paper with M. Loebl.
- Prize of the Second European Mathematical Congress for young mathematicians (Budapest, 1996).
- Prize of the Czech Learned Society for Young Scientists (2000).
- Elected member of the Czech Learned Society in 2006.
- The book *Thirty-Three Miniatures* published by the AMS was included among the “Outstanding Academic Titles 2011” of the journal *Choice*.
- Best paper award of the ACM-SIAM Symposium on Discrete Algorithms (2012) for a joint paper with M. Čadek, M. Krčál, F. Sergeraert, L. Vokřínek, and U. Wagner.
- Computational Geometry: Theory and Application 2016 Test of Time Award for *Reporting Points in Halfspaces*, Comput. Geom. 2: 169–186 (1992).

Teaching Activity

- Teaching regularly since 1987, courses given in Prague, Zurich, and Atlanta; many subjects in mathematics and computer science including basic undergraduate courses (discrete mathematics, linear algebra), graduate courses (discrete and computational geometry, programming languages, program construction and verification, data structures, algorithms, probabilistic method), and advanced courses on more special topics (topological methods in combinatorics and geometry, metric embeddings, semidefinite programming and approximation algorithms, advanced topics in discrete geometry).

- With J. Nešetřil developed the introductory course of discrete mathematics in Prague and wrote a successful textbook based on the course (so far published in six languages). With E. Welzl developed a course "Algorithms, Probability, Computing" at the ETH Zurich. Wrote books based on several other courses taught in Prague and/or Zurich.
- Taught two intensive block courses for Ph.D. students ("Topological Methods in Combinatorics and Geometry," ETH Zurich, 2001; "Metric embeddings," Univ. Autonoma de Barcelona, 2009).
- Supervised ten Ph.D. students (seven of them have finished successfully) and numerous M.Sc. and B.Sc. theses.
- Led research seminars for undergraduate students (combinatorics) and a reading seminar for Ph.D. students for many years.

Lecture Activities

- Invited speaker at the International Congress of Mathematicians in Berlin (1998) and European Congress of Mathematics (Budapest, 1996).
- Invited tutorial at the 39th IEEE Symposium on Foundations of Computer Science (FOCS), 1998.
- Invited lectures at several other conferences (e.g., Random Structures & Algorithms, Eurocomb, Graph Drawing, MFCS).
- Erdős Memorial Lectures at the Hebrew University of Jerusalem (2000).
- Plenary lecture at the annual meeting of the London Mathematical Society (1999).
- Plenary lecture at the joint meeting of the German Mathematical Society and the German Society for Mathematical Education (Berlin, 2007).
- Colloquia and invited lectures at numerous universities and research centers over the world, e.g., Institute for Advanced Study in Princeton, Princeton Univ., Univ. of Tokyo, Univ. College London, Cambridge Univ., Oxford Univ., ETH Zurich, Courant Institute of Math. Sciences in New York, University of California at Berkeley, EPFL Lausanne, Rutgers Univ., Tel Aviv Univ., Techn. Univ. Berlin, Freie Universität Berlin, IBM Tokyo, and Xerox Parc in Palo Alto. Several invited lectures every year.
- For other conference contributions (IEEE Symposium on Foundations of Comput. Sci., ACM Sympos. Theor. Comput., ACM Symposium on Comput. Geom., and others), see publications list.

Organization, Program Committees, and Refereeing

- Member of the program committee of the Int. Congress of Mathematicians (Section Combinatorics) 2006 and member of program committees of over ten conferences in computer science (STOC, ESA, ICALP, SODA, Eurocomb, LATIN, MFCS).
- Member of editorial boards of several journals (*Order, Discrete Computational Geometry, Random Structures and Algorithms, Theory of Computing, Contributions to Discrete Mathematics, Comment. Math. Univ. Carolinae, Comput. Geom. Theor. Appl., SIAM J. Discrete Math.*).
- Co-organized several conferences, including three Oberwolfach seminars on "Discrete Geometry" (2005, 2008, 2011).
- Co-organized with J. Nešetřil three intensive 3-month international programs for Ph.D. students in Prague "DocCourse"(2004–2006).

Research Interests

Discrete geometry, algorithms, combinatorics, topological methods, metric embeddings, and combinatorial optimization.

Professional Orientation and Summary of Major Results

Main Areas of Interest: Combinatorics and combinatorial geometry, geometric discrepancy, computer science (design and analysis of algorithms mainly computational geometry), topology, and some aspects of metric space theory.

Most of the papers belong to the fields of discrete and computational geometry. Series of works built efficient tools for removing randomization from geometric algorithms. Other works give new solutions to simplex and halfspace range searching problems, which have been investigated by many researchers for more than 10 years, and investigate various bounds on complexity of geometric configurations, geometric discrepancy, linear programming algorithms, motion planning, etc.

Works on geometric discrepancy developed new techniques for proving upper bounds. Asymptotically tight bounds were obtained in some long open cases, such as for the discrepancy of point sets with respect to halfspaces.

In papers on embedding finite metric spaces into Banach spaces, new bounds on the required distortion and dimensions were obtained; in particular, a question of Johnson and Lindenstrauss was answered.

The joint work with J. Kratochvíl deals with intersection graphs of planar geometric objects and some aspects of planar drawings of graphs.

Other results concern problems in mathematical analysis, graph algorithms, undecidability of combinatorial statements, Euclidean Ramsey theory, generalized convexity, and numerical taxonomy (with microbiology applications).

Some experience in applied areas was gained while developing computer software (e.g., numerical simulation of daylight conditions in a room, Lisp and Prolog language interpreters, syntax-driven editor).

List of Publications

Jiří Matoušek

Books

1. *Invitation to Discrete Mathematics*
with Jaroslav Nešetřil
Oxford University Press, Oxford, 1998, 432 pp.; revised 2nd edition 2008
Czech version: *Kapitoly z diskrétní matematiky*
preliminary version KAM Series 95–299, 1995, 218 pp.;
1st edition Matfyzpress, Praha 1997; 2nd edition Nakladatelství Karolinum, Praha 2000, reprinted 2003; revised 3rd edition 2008; revised 4th edition 2009.
German translation: *Diskrete Mathematik (Eine Entdeckungsreise)*, Springer, Heidelberg, 2002; revised 2nd edition 2007.
Japanese translation: Springer, Tokyo, 2003.
French translation: Springer, Heidelberg, 2004.
Spanish translation: Editorial Reverte, 2008.
2. *Geometric Discrepancy. An Illustrated Guide*
Volume 18 of *Algorithms and Combinatorics*, 288 pp., Springer-Verlag, Berlin, etc., 1999.
Revised second printing 2010.
3. *Lectures on Discrete Geometry*
Graduate Texts in Mathematics 212, 481 pp., Springer, New York, April 2002.
4. *Using the Borsuk–Ulam theorem. Lectures on topological methods in combinatorics and geometry*
Universitext, Springer, Berlin, etc., 196 pp., 2003.
Revised second printing 2008.
5. *Understanding and using linear programming*
with Bernd Gärtner
Universitext, Springer, Berlin, etc., 222 pp., 2006.
A shorter Czech version in *Lineární programování a lineární algebra pro informatiky*, ITI Series 2006-311, Charles University, Prague 2006.

6. *Thirty-three miniatures (Mathematical and algorithmic applications of linear algebra)*
   Student Mathematical Library, Amer. Math. Soc., Providence, 182 pp., 2010.
   Japanese translation: Springer, Tokyo, 2014.
7. *Approximation Algorithms and Semidefinite Programming*
   with Bernd Gärtner
   Springer, Berlin, etc., 251 pp., 2012.
8. *Mathematics++ (Selected topics beyond the basic courses)*
   with Ida Kantor and Robert Šámal
   Student Mathematical Library, Amer. Math. Soc., Providence, 343 pp., 2014.

Research Papers in Journals

9. *Approximate symmetric derivative and monotonicity*
   Comment. Math. Univ. Carolinae 27,1(1986), 83–86.
10. *Few colored cuts or cycles in edge colored graphs*
    Comment. Math. Univ. Carolinae 29(1988), 227–232.
11. *Line arrangements and range search*
    Information Processing Letters 27(1988), 275–280.
12. *On polynomial-time decidability of induced minor-closed classes*
    with Jaroslav Nešetřil and Robin Thomas
    Comment. Math. Univ. Carolinae 29,4(1988), 703–710.
13. *A typical property of the symmetric differential quotient*
    Colloquium Math. 57,2(1989), 339–343.
14. *Selecting a small well-discriminating subset of tests*
    with Jiří Schindler
    Binary 1(1989), 19–28.
15. *On-line computation of convolutions*
    Information Processing Letters 32(1989), 15–16.
16. *NP-hardness results for intersection graphs*
    with Jan Kratochvíl
    Comment. Math. Univ. Carolinae 30,4(1989), 761–773.
17. *Construction of ε-nets*
    Discr. Comput. Geom. 5(1990), 427–448 (invited paper).
    Extended abstract: Proc. 5. ACM Symposium on Computational Geometry 1989, 1–9.
18. *Extension of Lipschitz mappings on metric trees*
    Comment. Math. Univ. Carolinae 31,1(1990), 99–104.
19. *Bi-Lipschitz embeddings into low-dimensional Euclidean spaces*
    Comment. Math. Univ. Carolinae 31,3(1990), 589–600.
20. *Algorithms finding tree-decomposition of graphs*
    with Robin Thomas
    J. Algorithms 12,1(1991), 1–22.

21. *Lower bound on the length of monotone paths in arrangements*
Discr. Comput. Geom. 6,2(1991) 129–134.
22. *Approximate halfplanar range counting*
KAM Series in Discrete Mathematics 87–59 (tech. report), Charles University, Prague 1987.
revised in 1989 as *Approximate levels in line arrangements*, SIAM J. Computing 20,2(1991), 222–227.
23. *Spanning trees with low crossing number*
Informatique théorique et applications 6,25(1991), 103–123.
24. *Computing dominances in E^n*
Information Processing Letters 38,5(1991), 277–288.
25. *String graphs requiring huge representations*
with Jan Kratochvíl
J. Combin. Theory ser. B 31,1(1991), 1–4.
26. *Cutting hyperplane arrangements*
Discr. Comput. Geom. 6,5(1991), 385–406 (invited paper).
Extended abstract: Proc. 6. ACM Symposium on Computational Geometry (1990), 1–9.
27. *Randomized optimal slope selection*
Information Processing Letters 39(1991), 183–187.
28. *Hercules versus Hidden Hydra Helper*
with Martin Loebl
Comment. Math. Univ. Carolinae 32,4(1992), 731–741.
29. *Good splitters for counting points in triangles*
with Emo Welzl
J. Algorithms 13(1992), 307–319.
Extended abstract: Proc. 5. ACM Symposium on Computational Geometry (1989), 124–130.
30. *Note on bi-Lipschitz embeddings into normed spaces*
Comment. Math. Univ. Carolinae 33,1(1992), 51–55.
31. *Relative neighborhood graphs in three dimensions*
with Pankaj K. Agarwal
Computational Geometry: Theory and applications 2,1(1992), 1–14.
Preliminary version: Proc. 3. ACM-SIAM Symposium on Discrete Algorithms (1992), 58–65.
32. *Efficient partition trees*
Discr. Comput. Geom., 8(1992), 315–334 (invited paper).
Extended abstract: Proc. 7. ACM Symposium on Computational Geometry (1991), 1–9.
33. *Reporting points in halfspaces*
Computational Geometry: Theory and Applications 2,3(1992) 169–186.
Extended abstract: Proc. 32. IEEE Symposium on Foundations of Computer Science (1991), 207–215.

34. *On the complexity of finding iso- and other morphisms for partial k-trees*
with Robin Thomas
Discrete Math. 108(1992), 343–364.
35. *Ramsey-like properties for bi-Lipschitz embeddings of finite metric spaces*
Comment. Math. Univ. Carolinae 33,3(1992), 451–463.
36. *Farthest neighbors, maximum spanning trees and related problems in higher dimensions*
with Pankaj K. Agarwal and Subhash Suri
Computational Geometry: Theory and Applications 1,4(1992) 189–201.
Extended abstract: Proc. 2. Workshop on Algorithms and Data Structures, Lecture Notes in Computer Science 519, 105–116, Springer-Verlag 1991.
37. *On vertical ray shooting in arrangements*
Computational Geometry: Theory and Applications 2(1993), 279–285.
38. *Linear optimization queries*
J. Algorithms 14(1993), 432–448
new version, with Otfried Schwarzkopf:
Proc. 8. ACM Symposium on Computational Geometry (1992), 16–25.
39. *Ray shooting and parametric search*
with Pankaj K. Agarwal
SIAM J. Computing 22,4(1993), 794–806. Extended abstract: Proc. 24. ACM Symposium on Theory of Computing (1992), 517–526.
40. *Range searching with efficient hierarchical cuttings*
Discr. Comput. Geom. 10,2(1993), 157–182.
Extended abstract: Proc. 8. ACM Symposium on Computational Geometry (1992), 276–285.
41. *On ray shooting in convex polytopes*
with Otfried Schwarzkopf
Discr. Comput. Geom. 10,2(1993), 215–232.
42. *Discrepancy and approximations for bounded VC-dimension*
with Emo Welzl and Lorenz Wernisch
Combinatorica, 13(1993), 455–466.
Extended abstract: Proc. 32. IEEE Symposium on Foundations of Computer Science (1991), 424–430.
43. *On the sum of squares of cell complexities in hyperplane arrangements*
with Boris Aronov and Micha Sharir
J. Combin. Theory Ser. A 65(1994), 311–321.
Extended abstract: Proc. 7. ACM Symposium on Computational Geometry (1991), 307–313.
44. *On range searching with semialgebraic sets*
with Pankaj K. Agarwal
Discr. Comput. Geom. 11(1994), 393–418.
Extended abstract: Proc. 17. Symposium "Mathematical Foundations of Computer Science" (1992), Lecture Notes in Computer Science 629, Springer-Verlag, 1–13.

45. *Fat triangles determine linearly many holes*
with János Pach, Micha Sharir, Shmuel Sifrony, and Emo Welzl
SIAM J. Comput. 23(1994), 154–169.
Extended abstract: Proc. 32. IEEE Symposium on Foundations of Computer Science (1991), 49–58.
46. *Lower bound for a subexponential optimization algorithm*
Random Structures & Algorithms 5,4(1994), 591–607.
47. *Ham-sandwich cuts in R^d*
with Chi-Yuan Lo and William Steiger
Discr. Comput. Geom. 11(1994), 433–452.
Extended abstract: Proc. 24. ACM Symposium on Theory of Computing (1992), 539–545.
48. *Intersection graphs of segments*
with Jan Kratochvíl
J. Combin. Theory Ser. B 35,2(1994), 317–339.
49. *Complexity of projected images of convex subdivisions*
with Tomio Hirata, Xue-Hou Tan, and Takeshi Tokuyama
Computational Geometry: Theory and Applications 4,6(1994), 293–308.
Extended abstract: Proc. 4. Canad. Conference on Comput. Geometry (1992).
50. *A Ramsey-type result for planar convex sets*
with David Larman, János Pach, and Jenő Törőcsik
Bull. London Math. Soc. 26(1994), 132–136.
51. *Derandomizing an output-sensitive convex hull algorithm in three dimensions*
with Bernard Chazelle
Comput. Geom.: Theor. Appl. 5,1(1995), 27–32.
52. *On enclosing k points by a circle*
Information Processing Letters 53(1995), 217–221.
53. *Dynamic half-space range reporting and its applications*
with Pankaj K. Agarwal
Algorithmica 13(1995), 325–345.
Extended abstract, including also results of D. Eppstein:
Proc. 33. IEEE Symposium on Foundations of Computer Science (1992), 51–60.
54. *Approximations and optimal geometric divide-and-conquer*
J. of Computer and System Sciences 50,2(1995), 203–208 (invited paper).
Extended abstract: Proc. 23. ACM Symposium on Theory of Computing (1991), 506–511.
55. *On Ramsey sets in spheres*
with Vojtěch Rödl
J. Combin. Theory Ser. A 70,1(1995), 30–44.
56. *Tight upper bounds for the discrepancy of half-spaces*
Discr. Comput. Geom. (L. Fejes Tóth Festschrift) 13(1995), 593–601.
57. *An elementary approach to lower bounds in geometric discrepancy*
with Bernard Chazelle and Micha Sharir
Discr. Comput. Geom. (L. Fejes Tóth Festschrift) 13(1995), 363–381.

58. *Piecewise linear paths among convex obstacles*
with Mark de Berg and Otfried Schwarzkopf
Discr. Comput. Geom. 14(1995), 9–29.
Extended abstract: Proc. 25. ACM Symposium on Theory of Computing (1993), 505–514.
59. *On stabbing triangles by lines in 3-space*
with Boris Aronov
Comment. Math. Univ. Carolinae 36,1(1995), 109–113.
60. *On vertical decomposition of arrangements of hyperplanes in four dimensions*
with Leonidas J. Guibas, Dan Halperin, and Micha Sharir
Discr. Comput. Geom. 14(1995), 113–122.
Extended abstract: Proc. 5th Canadian Conference on Computational Geometry (1993), 127–132.
61. *Note on the colored Tverberg theorem*
J. Comb. Theory Ser. B 66(1996), 146–151.
62. *On geometric optimization with few violated constraints*
Discr. Comput. Geom. (invited paper) 14(1995), 365–384.
Extended abstract: Proc. 10. ACM Symposium on Comput. Geom. (1994), 312–321.
63. *Discrepancy in arithmetic progressions*
with Joel Spencer
J. Amer. Math. Soc. 9,1(1996), 195–204.
64. *On the distortion required for embedding finite metric spaces into normed spaces*
Israel J. Math. 93(1996), 333–344.
65. *A deterministic algorithm for the three-dimensional diameter problem*
with Otfried Schwarzkopf
Comput. Geom. Theor. Appl. 6(1996), 253–262.
Extended abstract: Proc. 25. ACM Symposium on Theory of Computing (1993), 478–484.
66. *On linear-time deterministic algorithms for optimization problems in fixed dimension*
with Bernard Chazelle
J. Algorithms 21(1996), 116–132.
Extended abstract: Proc. 4. SIAM-ACM Symposium on Discrete Algorithms (1993), 281–290.
67. *Improved upper bounds for approximation by zonotopes*
Acta Mathematica 177(1996), 55–73.
68. *A subexponential bound for linear programming*
with Micha Sharir and Emo Welzl
Algorithmica 16(1996), 498–516.
Extended abstract: Proc. 8. ACM Symposium on Computational Geometry (1992), 1–8.
69. *On discrepancy bounds via dual shatter function*
Mathematika 44(1997), 42–49.

70. *A Helly-type theorem for unions of convex sets*
Discr. Comput. Geom. 18(1997), 1–12.
Extended abstract: Proc. 11. ACM Symposium on Comput. Geom. (1995), 138–145.
71. *On embedding expanders into ℓ_p spaces*
Israel J. Math. 102(1997), 189–197.
72. *On functional separately convex hulls*
with Petr Plecháč
Discr. Comput. Geom. 19(1998), 105–130.
73. *An L_p-version of the Beck–Fiala conjecture*
European J. Combinatorics 19(1998), 175–182.
74. *Guarding galleries where every point sees a large area*
with Gil Kalai
Israel J. Math. 101(1997), 125–140.
75. *Computing many faces in arrangements of lines and segments*
with Pankaj K. Agarwal and Otfried Schwarzkopf
SIAM J. Comput. 27,2(1998), 491–505.
Extended abstract: 10. ACM Symposium on Comput. Geom. (1994), 76–84.
76. *Constructing levels in arrangements and higher order Voronoi diagrams*
with Pankaj K. Agarwal, Mark de Berg, and Otfried Schwarzkopf
SIAM J. Comput. 27,3(1998), 654–667.
Extended abstract: 10. ACM Symposium on Comput. Geom. (1994), 67–75.
77. *An $O(n \log n)$ randomized algorithm for the repeated median line estimator*
with David M. Mount and Nathan S. Netanyahu
Algorithmica 20,2(1998), 136–150.
Extended abstract: Proc. 4. SIAM-ACM Symposium on Discrete Algorithms (1993), 74–82.
78. *On the L_2-discrepancy for anchored boxes*
J. of Compexity 14(1998), 527–556.
79. *The exponent of discrepancy is at least 1.0669*
J. of Compexity 14(1998), 448–453.
80. *On constants for cuttings in the plane*
Discr. Comput. Geom. 20(1998), 427–448.
81. *On the discrepancy for boxes and polytopes*
Monatsh. Math. 127(1999), 325–336.
82. *Almost-tiling the plane with ellipses*
with Krystyna Kuperberg, Wlodzimierz Kuperberg, and Pavel Valtr
Discr. Comput. Geom. 22(1999), 367–375.
83. *A highly non-smooth norm on Hilbert space*
with Eva Matoušková
Israel J. Math. 112(1999), 1–27.
84. *Visibility and covering by convex sets*
with Pavel Valtr
Israel J. Math. 113(1999), 341–379.

85. *Product range spaces, sensitive sampling and derandomization*
with Hervé Brönnimann and Bernard Chazelle
SIAM J. Comput. 28,5(1999), 1552–1575.
Extended abstract: Proc. 34. IEEE Symposium on Foundations of Computer Science (1993), 400–409.
86. *On the signed domination in graphs*
Combinatorica 20,1(2000), 103–108.
87. *On embedding trees into uniformly convex Banach spaces*
Israel J. Math. 114(1999), 221–237.
88. *On the linear and hereditary discrepancies*
European J. Combin. 21(2000), 519–521.
89. *Discrepancy of point sequences on fractal sets*
with Hansjörg Albrecher and Robert Tichy
Publicationes Mathematicae Debrecen (spec. volume dedicated to K. Győry) 56(2000), 233–249.
90. *On approximate geometric k-clustering*
Discr. Comput. Geom. 24(2000), 61–84.
91. *On the discrepancy for Cartesian products*
J. London Math. Soc. 61(2000), 737–747.
92. *Simultaneous partitions of measures by k-fans*
with Imre Bárány
Discr. Comput. Geom. 25,3(2001), 317–334.
93. *On directional convexity*
Discr. Comput. Geom. 25,3(2001), 389–405.
94. *Lower bound on the minus-domination number*
Discr. Math. 233(2001), 361–370.
95. *On dominated ℓ_1 metrics*
with Yuri Rabinovich
Israel J. Math. 123(2001), 285–301.
96. *A lower bound for families of Natarajan dimension d*
with Paul Fischer
J. Combin. Theory Ser. A 95(2001), 198–195.
97. *Lower bounds on the transversal numbers of d-intervals*
Discr. Comput. Geom. 26(2001), 283–287.
98. *Random lifts of graphs III: independence and chromatic number*
with Alon Amit and Nathan Linial
Rand. Struct. Algo. 20(2002), 1–22.
Extended abstract appeared as a part of "A. Amit, N. Linial, J. Matoušek, E. Rozenman: *Random lifts of graphs*, Proc. 12th annual ACM-SIAM Symposium on Discrete Algorithms, 883–894, 2001."
99. *Separating an object from its cast*
with Hee-Kap Ahn, Mark de Berg, Prosenjit Bose, Siu-Wing Cheng, Dan Halperin, and Otfried Schwarzkopf
Computer-Aided Design 34(2002), 547–559.

Extended abstract: Proc. 12th ACM Symposium on Computational Geometry, 1997.

100. *Equipartition of two measures by a 4-fan*
with Imre Bárány
Discr. Comput. Geom. 27(2002), 293–302.
101. *On the chromatic number of Kneser hypergraphs*
Proc. Amer. Math. Soc. 130 (2002), 2509–2514.
102. *Transversal numbers for hypergraphs arising in geometry*
with Noga Alon, Gil Kalai, and Roy Meshulam
Adv. Appl. Math. 130,9(2002), 2509–2514.
103. *A lower bound for weak epsilon-nets in high dimension*
Discr. Comput. Geom. 28(2002), 45–48.
104. *A lower bound on the size of Lipschitz subsets in dimension 3*
Combin. Probab. Comput. 12(2003), 427–430.
105. *A fractional Helly theorem for convex lattice sets*
with Imre Bárány
Adv. Math. 174(2003), 227–235.
106. *On restricted min-wise independence of permutations*
with Miloš Stojaković
Random Structures and Algorithms 23,4(2003), 397–408.
Extended abstract: Proc. of Eurocomb 2003.
107. *Low-distortion embeddings of trees*
with Robert Babilon, Jana Maxová, and Pavel Valtr
J. Graph Algorithms Appl. 7(2003), 399–409.
Extended abstract: Proc. Graph Drawing, 2001 (Lecture Notes in Computer Science 2265), 2002, 343–351.
108. *The one-round Voronoi game*
with Otfried Cheong, Sariel Har-Peled, and Nathan Linial
Discrete Comput. Geom. 31(2004), 125–138.
Extended abstract: Proc. 18th ACM Sympos. Comput. Geom. 2002.
109. *No Helly theorem for stabbing translates by lines in R^3*
with Andreas Holmsen
Discr. Comput. Geom. 31,3(2004), 405–410.
110. *Bounded VC-dimension implies a fractional Helly theorem*
Discr. Comput. Geom., 31,2(2004), 251–255.
111. *Topological lower bounds for the chromatic number: A hierarchy*
with Günter M. Ziegler
Jahresbericht der Deutschen Mathematiker-Vereinigung 106(2004), 71–90.
112. *Crossing number, pair-crossing number, and expansion*
with Petr Kolman
J. Comb. Theory Ser. B 92(2004), 99–113.
113. *New constructions of weak epsilon-nets*
with Uli Wagner
Discr. Comput. Geom. 32,2(2004), 195–206.
Extended abstract: Proc. 19th ACM Sympos. Comput. Geom. 2003.

114. *A combinatorial proof of Kneser's conjecture*
Combinatorica 24,1(2004), 163–170.
115. *The randomized integer convex hull*
with Imre Bárány
Discr. Comput. Geom. 33,1(2005), 3–25.
116. *Triangles in random graphs*
with Martin Loebl and Ondřej Pangrác
Discrete Math. 289(2004), 181–185.
117. *Expected length of the longest common subsequence for large alphabets*
with Marcos Kiwi and Martin Loebl
Adv. Math. 197(2005), 480–498.
Extended abstract: Proc. LATIN 2004: Theoretical Informatics: 6th Latin American Symposium, Lecture Notes in Computer Science, Springer Berlin Heidelberg, (2004), 302–311.
118. *Bounded-degree graphs have arbitrarily large geometric thickness*
with János Barát and David Wood
The Electronic Journal of Combinatorics 13,1(2006).
119. *Discrepancy after adding a single set*
with Jeong Han Kim and Van H. Vu
Combinatorica 25(2005), 499–501.
120. *The number of unique-sink orientations of the hypercube*
Combinatorica 26(2006), 91–99.
121. *On k-sets in four dimensions*
with Micha Sharir, Shakhar Smorodinsky, and Uli Wagner
Discr. Comput. Geom. 35,2(2006), 177–191.
122. *RANDOM EDGE can be exponential on abstract cubes*
with Tibor Szabó
Advances in Mathematics 204(2006), 262–277.
Extended abstract: in Proc. 45th IEEE Symposium on Foundations of Computer Science (FOCS), 2004.
123. *On-line conflict-free colorings for intervals*
with Ke Chen, Amos Fiat, Haim Kaplan, Meital Levy, Elchanan Mossel, János Pach, Micha Sharir, Shakhar Smorodinsky, Uli Wagner, and Emo Welzl
SIAM J. Computing 36(2006), 1342–1359.
Extended abstract, not involving Ke Chen as author: Proc. ACM-SIAM Symposium on Discrete Algorithms, 2005, 545–554.
124. *Segmenting object space by geometric reference structures*
with Pankaj K. Agarwal and David Brady
ACM Transactions on Sensor Networks 2,4(2006), 455–465.
125. *Minimum independence number of a Hasse diagram*
with Aleš Přívětivý
Combin. Probab. Comput., 15,3(2006), 473–475.
126. *Berge's theorem, fractional Helly, and art galleries*
with Imre Bárány

Discr. Math. (special volume in memory of Claude Berge) 35,2(2006), 177–191.

127. *Quadratically many colorful simplices*
with Imre Bárány
SIAM J. Discrete Math. 21,1(2007), 191–198.

128. *The distance trisector curve*
with Tetsuo Asano and Takeshi Tokuyama
Advances in Mathematics 212(2007), 338–360.
Extended abstract: Proc. 38th ACM Symposium on Theory of Computing, 2006, 336–343.

129. *Packing cones and their negatives in space*
with Imre Bárány
Discr. Comput. Geom (L. Fejes Toth special volume) 38(2007), 177–187.

130. *Removing degeneracy may require a large dimension increase*
with Petr Škovroň
Theory of Computing 3/8(2007), 159–177.
Extended abstract: Proc. Eurocomb 2007, Electronic Notes in Discrete Mathematics 29C(2007), 107–113.

131. *Zone diagrams: existence, uniqueness, and algorithmic challenge*
with Tetsuo Asano and Takeshi Tokuyama
SIAM J. Computing 37,4(2007), 1182–1198.
Extended abstract: Proc. ACM-SIAM Symposium on Discrete Algorithms, 2007, 756–765.

132. *Induced trees in triangle-free graphs*
with Robert Šámal
Electr. J. Combin., 15,1(2008), R41.
Extended abstract: Proc. Eurocomb 2007, Electronic Notes in Discrete Mathematics 29C(2007), 307–313.

133. *Large monochromatic components in two-colored grids*
with Aleš Přívětivý
SIAM J. Discr. Math. 22(2008), 295–311.
Extended abstract: Proc. Eurocomb 2007, Electronic Notes in Discrete Mathematics 29C(2007), 3–9.

134. *Removing degeneracy in LP-type problems revisited*
Discr. Comput. Geom. 42,4(2009), 517–526.

135. *Dimension gaps between representability and collapsibility*
with Martin Tancer
Discr. Comput. Geom. 42,4(2009), 631–639.

136. *Violator spaces: structure and algorithms*
with Bernd Gärtner, Leo Rüst, and Petr Škovroň
Discr. Appl. Math. 156(2008), 2124–2141.
Extended abstract: Proc. European Symposium on Algorithms, Springer, 2006, 387–398.

137. *Graph coloring with no large monochromatic components*
with Nathan Linial, Or Sheffet, and Gábor Tardos

Combin. Probab. Comput. 17,4(2008), 577–589.
Extended abstract: Proc. Eurocomb 2007, Electronic Notes in Discrete Mathematics 29C(2007), 115–122.

138. *On variants of the Johnson–Lindenstrauss lemma*
Random Structures & Algorithms 33,2(2008), 142–156.

139. *Computing D-convex hulls in the plane*
with Vojtěch Franěk
Comput. Geom. Theor. Appl. 42(2009), 81–89.

140. *LC reductions yield isomorphic simplicial complexes*
Contributions to Discrete Mathematics (electronic) 3,2(2008).

141. *How many points can be reconstructed from k projections?*
with Aleš Přívětivý and Petr Škovroň
SIAM J. Discr. Math. 22,4(2008), 1605–1623.
Extended abstract: Proc. Eurocomb 2007, Electronic Notes in Discrete Mathematics 29C(2007), 427–434.

142. *Blocking visibility for points in general position*
Discr. Comput. Geom. 42,2(2009), 19–22 (special issue dedicated to Victor Klee).

143. *Stabbing simplices by points and flats*
with Boris Bukh and Gabriel Nivasch
Discr. Comput. Geom. 42,2(2010), 321–338.

144. *Distance k-sectors exist*
with Keiko Imai, Akitoshi Kawamura, Daniel Reem, and Takeshi Tokuyama
Comput. Geom. Theor. Appl. 43,9(2010), 713–720.
Extended abstract: Proc. 26th ACM Symposium Comput. Geom., 2010, 210–215.

145. *Inapproximability for metric embeddings into R^d*
with Anastasios Sidiropoulos
Trans. Amer. Math. Soc. 362(2010), 6341–6365.
Extended abstract: Proc. IEEE Symposium on Foundations of Computer Science (FOCS), 2008, 405–413.

146. *The number of unit distances is almost linear for most norms*
Adv. Math. 226(2011), 2618–2628.

147. *Hardness of embedding simplicial complexes in* R^d
with Uli Wagner and Martin Tancer
J. European Math. Soc. 13(2011), 259–295.
Extended abstract: Proc. ACM-SIAM Symposium on Discrete Algorithms, 2009, 855–864.

148. *Lower bounds for weak epsilon-nets and stair-convexity*
with Boris Bukh and Gabriel Nivasch
Israel J. Math 182(2011), 199–228.
Extended abstract: Proc. 25th Sympos. Comput. Geom., Aarhus, Denmark, 2009, 1–10.

149. *The t-pebbling number is eventually linear in t*
with Michael Hoffmann, Yoshio Okamoto, and Philipp Zumstein
Electronic J. Combin. 18,1(2011), P153.
150. *On the nonexistence of k-reptile tetrahedra*
with Zuzana Safernová
Discr. Comput. Geom. 46,3(2011), 599–609.
151. *Reachability by paths of bounded curvature in convex polygons*
with Hee-kap Ahn, Otfried Cheong, and Antoine Vigneron
Comput. Geom. Theor. Appl. 45,1–2(2012), 21–32.
Extended abstract: Proc. 16th ACM Sympos. Comput. Geom. 2000, 251–259.
152. *A geometric proof of the colored Tverberg theorem*
with Martin Tancer and Uli Wagner
Discr. Comput. Geom. 47,2(2012), 245–265.
153. *A doubly exponentially crumbled cake*
with Tobias Christ, Andrea Francke, Heidi Gebauer, and Takeaki Uno
Electronic Notes in Discrete Mathematics 38(2011), 265–271.
154. *Simple proofs of classical theorems in discrete geometry via the Guth–Katz polynomial partitioning technique*
with Haim Kaplan and Micha Sharir
Discr. Comput. Geom. 48,3(2012), 499–517.
Preprint: arXiv:1102.5391.
155. *Unit distances in three dimensions*
with Haim Kaplan, Zuzana Safernová, and Micha Sharir
Combinatorics, Probability, Computing 21(2012), 597–610.
Preprint: arXiv:1107.1077.
156. *Minimum and maximum against k lies*
with Michael Hoffmann, Yoshio Okamomoto, and Phillip Zumstein
Chicago J. Theor. Comput. Sci. 2012, Article 2.
Extended abstract: Proc. 12th Scandinavian Symposium and Workshops on Algorithm Theory, Bergen (Lecture Notes in Computer Science 6139), Springer, 2010, 139–149.
157. *Vectors in a box*
with Kevin Buchin, Robin A. Moser, and Dömötör Pálvölgyi
Math. Programming Ser. A 135,1–2(2012), 323–335.
158. *Zone diagrams in Euclidean spaces and in other normed spaces*
with Akitoshi Kawamura and Takeshi Tokuyama
Mathematische Annalen 354,4(2012), 1201–1221.
Extended abstract: Proc. 26th ACM Symposium Comput. Geom., 2010, 216–221.
159. *The determinant bound for discrepancy is almost tight*
Proc. Amer. Math. Soc. 141(2013), 451–460.
160. *Higher-order Erdős–Szekeres theorems*
with Marek Eliáš
Advances in Mathematics 244(2013), 1–15.

Extended abstract: Proc. 28th Annu. ACM Symposium on Comput. Geom., Chapel Hill, NC, 2012, 81–90.

161. *Polynomial-time homology for simplicial Eilenberg–MacLane spaces*
with Marek Krčál and Francis Sergeraert
Journal of Foundations of Computational Mathematics, 13,6(2013), 935–963.
Preprint: arXiv:1201.6222.

162. *On range searching with semialgebraic sets II*
with Pankaj K. Agarwal and Micha Sharir
SIAM J. Computing 42,6(2013), 2039–2062.
Extended abstract: Proc. 53rd Annual IEEE Symposium on Foundations of Computer Science (FOCS 2012), New Brunswick, NJ, 2012, 420–429.

163. *Near-optimal separators in string graphs*
Combinatorics, Probability and Computing 23,1(2014), 135–139.
Preprint: arXiv:1302.6482.

164. *Extendability of continuous maps is undecidable*
with Martin Čadek, Marek Krčál, Lukáš Vokřínek, and Uli Wagner
Discr. Comput. Geom. 51,1(2014), 24–66.
Preprint: arXiv:1302.2370.

165. *On Gromov's method of selecting heavily covered points*
with Uli Wagner
Discr. Comput. Geom. 52,1(2014), 1–33.
Preprint: arXiv:1102.3515.

166. *Lower bounds on geometric Ramsey functions*
with Marek Eliáš, Edgardo Roldán-Pensado, and Zuzana Safernová, SIAM J. Discr. Math. 28,4(2014), 1960–1970.
Extended abstract: Proc. 30th Annual Symposium on Computational Geometry, 2014, 558–564.
Preprint: arXiv:1307.5157.

167. *Erdős–Szekeres-type statements: Ramsey function and decidability in dimension 1*
with Boris Bukh
Duke Math. J. 163,12(2014), 2243–2270.
Preprint: arXiv:1207.0705.

168. *Computing all maps into a sphere*
with Martin Čadek, Marek Krčál, Francis Sergeraert, Lukáš Vokřínek, and Uli Wagner
J. ACM 61,3(2014), Article No. 17.
Short abstract: Report No. 08/2011, Mathematisches Forschungsinstitut Oberwolfach, 65–68.
Extended abstract: Proc. ACM–SIAM Symposium on Discrete Algorithms, 2012.
Preprint: arXiv:1105.6257.

169. *Polynomial-time computation of homotopy groups and Postnikov systems in fixed dimension*
with Martin Čadek, Marek Krčál, Lukáš Vokřínek, and Uli Wagner

SIAM J. Computing 43,5(2014), 1728–1780.
Preprint: arXiv:1211.3093.

170. *Curves in R^d intersecting every hyperplane at most $d+1$ times*
with Imre Bárány and Attila Pór
J. European Math. Soc. 18,11(2016), 2469–2482.
Extended abstract: Proc. 30th Annual Symposium on Computational Geometry, 2014, 565–574.
Preprint: arXiv:1309.1147.

171. *Untangling two systems of noncrossing curves*
with Eric Sedgwick, Martin Tancer, and Uli Wagner
Israel J. Math. 212,1(2016), 37–79.
Extended abstract: Proc. 21st International Symposium on Graph Drawing (2013), Lecture Notes in Computer Science 8242, Springer, Berlin 2013, 472–483.
Preprint: arXiv:1302.6475.

172. *Simplifying inclusion-exclusion formulas*
with Xavier Goaoc, Pavel Paták, Zuzana Safernová, and Martin Tancer
Combin. Probab. Comput. 24,2(2015), 438–456.
Extended abstract: Eurocomb 2013.

173. *Three-monotone interpolation*
with Josef Cibulka and Pavel Paták
Discr. Comput. Geom. 54,1(2015), 3–21.
Preprint: arXiv:1404.4731.

174. *Computing higher homotopy groups is $W[1]$-hard*
Preprint: arXiv:1304.7705.

175. *Multilevel polynomial partitions and simplified range searching*
with Zuzana Safernová
Discr. Comput. Geom. 54,1(2015), 22–41.

Surveys and Expository Notes

176. *Epsilon-nets and computational geometry*
In Algorithms and Combinatorics, vol. 10: "New Trends in Discrete and Computational Geometry" (J. Pach ed.), Springer-Verlag 1993, 69–89.

177. *Derandomization in computational geometry*
J. Algorithms 20(1996), 545–580.
Extended and updated version: in "Handbook of Computational Geometry" (J.-R. Sack and J. Urrutia, eds.), North Holland, Amsterdam, 2000, 559–596.

178. *Geometric range searching*
ACM Comput. Surveys 26(1995), 421–461.

179. *Geometric set systems*
European Congress of Mathematics (Budapest, July 22–26, 1996), vol. II, 2–27, Birkhäuser, Basel, 1998.

180. *Mathematical snapshots from the computational geometry landscape*
Documenta Mathematica J. DMV, Extra volume ICM 1998, vol. III, 1998.
181. *Geometric computation and the art of sampling (tutorial)*
Abstract of an invited lecture, Proc. 39. IEEE Symposium on Foundations of Computer Science, 1998.
182. *Low-distortion embeddings of discrete metric spaces*
with Piotr Indyk
Chapter 8 of CRC Handbook of Discrete and Computational Geometry (J. E. Goodman and J. O'Rourke, eds.), 2nd edition, CRC Press, LLC, Boca Raton, FL, 177–196, 2004.
183. *The dawn of an algebraic era in discrete geometry?*
Proc. 27th European Workshop on Computational Geometry (EuroCG) 2011, 5–10.
184. *String graphs and separators*
in Geometry, Structure and Randomness in Combinatorics (J. Matoušek, J. Nešetřil, and M. Pellegrini, eds.), Scuola Normale Superiore, Pisa, 2014, 61–97.
Preprint: arXiv:1311.5048.

Conference Contributions Not Published in Journals and Papers in Special Volumes

185. *On undecidability of the weakened Kruskal theorem*
with Martin Loebl
Contemporary Mathematics vol. 65 (Logics and Combinatorics), Am. Math. Soc. 1987, 275–279.
186. *On perfect codes in a random graph*
with Jan Kratochvíl and Jan Malý
in Random Graphs '87 (M. Karoński, J. Jaworski, and A. Ruciński, eds.), J. Wiley & Sons 1990, 141–149.
187. *Computing the center of planar point sets*
In Computational Geometry: papers from the DIMACS special year (J. E. Goodman, R. Pollack, and W. Steiger, eds.), AMS-ACM DIMACS series, Amer. Math. Soc. 1991, 221–230.
188. *How to net a lot with a little: Small ε-nets for disks and halfspaces*
with Raimund Seidel and Emo Welzl
Proc. 6. ACM Symposium on Computational Geometry (1990), 16–22.
189. *On Lipschitz mappings onto a square*
In The Mathematics of Paul Erdős II (R. Graham and J. Nešetřil, eds.), Springer-Verlag 1997, 303–309.
190. *The complexity of the lower envelope of segments with h endpoints*
with Pavel Valtr

Bolyai Society Math. Studies 6, Intuitive Geometry, Budapest (Hungary), 1995, J. Bolyai Society, Budapest 1997, 407–411.

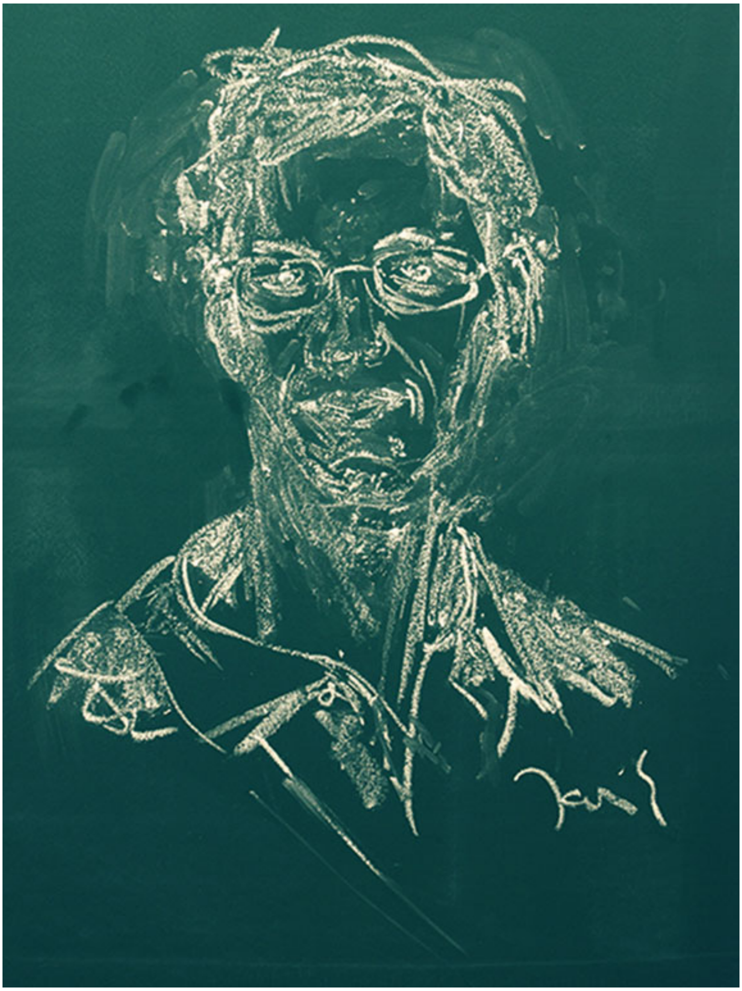
191. *Integer points in rotating convex bodies*
with Imre Bárány
In Discrete and Computational Geometry: The Goodman–Polack Festschrift (B. Aronov, S. Basu, J. Pach, and M. Sharir, eds.), in the series: Algorithms and Combinatorics 25, Springer-Verlag, Berlin 2003, 177–201.
192. *Towards asymptotic optimality in probabilistic packet marking*
with Micah Adler and Jeff Edmonds
Proc. 37th ACM Symposium on Theory of Computing, 2005, 450–459.
193. *Nonexistence of* 2*-reptile simplices*
in Discrete and Computational Geometry: Japanese Conference, JCDCG 2004, Lecture Notes in Computer Science 3742, Springer, Berlin, etc., 2005, 151–160.
194. *Extending continuous maps: polynomiality and undecidability*
with Martin Čadek, Marek Krčál, Lukáš Vokřínek, and Uli Wagner
Proc. 45th ACM Symposium on Theory of Computing, 2013, 595–604.
195. *Embeddability in the 3-sphere is decidable*
with Eric Sedgwick, Martin Tancer, and Uli Wagner
Proc. 30th Annual Symposium on Computational Geometry, 2014, 78–84, best paper award.

Technical Reports

196. *More on cutting arrangements and spanning trees with low crossing number*
Tech. Report B–90–2, FU Berlin, FB Mathematik, February 1990.
197. *A simple proof of the weak zone theorem*
KAM Series (Tech. Report) 90–178, Charles University, Prague 1990.

Book Manuscripts and Lecture Notes

198. *The probabilistic method*
with Jan Vondrák
Lecture notes, ca. 60 pp., KAM Series.
199. *Kombinatorika a grafy I* (in Czech)
with Tomáš Valla (first author)
Lecture notes, ITI Series 2005-260, 51 pages.
200. *Metric embeddings*
available online, 125 pages.



Drawing by Jarik N., chalk on blackboard, 2016

Simplex Range Searching and Its Variants: A Review

Pankaj K. Agarwal

Abstract A central problem in computational geometry, range searching arises in many applications, and numerous geometric problems can be formulated in terms of range searching. A typical range-searching problem has the following form. Let S be a set of n points in $\mathbb{R}^d$, and let $\mathcal{R}$ be a family of subsets of $\mathbb{R}^d$; elements of $\mathcal{R}$ are called *ranges*. Preprocess S into a data structure so that for a query range $\gamma \in \mathcal{R}$, the points in $S \cap \gamma$ can be reported or counted efficiently. Notwithstanding extensive work on range searching over the last four decades, it remains an active research area. A series of papers by Jirka Matoušek and others in the late 1980s and the early 1990s had a profound impact not only on range searching but also on computational geometry as a whole. This chapter reviews the known results and techniques, including recent developments, for simplex range searching and its variants.

1 Introduction

In the mid 1980s, the range-searching problem, especially simplex range searching, was wide open: neither efficient algorithms nor nontrivial lower bounds were known. A series of papers in the late 1980s and the early 1990s [42, 43, 65, 76, 77, 93] not only marked the beginning of a new chapter in range searching but also revitalized computational geometry as a whole. The impact of techniques developed for range searching—ε-nets, $(1/r)$-cuttings, partition trees, simplicial partitions, multi-level data structures, to name a few—is evident throughout computational geometry. The papers by Jirka Matoušek [74–79] were at the center of this range-searching revolution. This book honoring his work provides an excellent opportunity

Work on this chapter is supported by NSF under grants CCF-11-61359, IIS-14-08846, and CCF-15-13816, by an ARO grant W911NF-15-1-0408, and by Grant 2012/229 from the U.S.-Israel Binational Science Foundation.

P.K. Agarwal (✉)
Department of Computer Science, Duke University, 27708-0129 Durham, NC, USA
e-mail: pankaj@cs.duke.edu

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_1

to review the current status of geometric range searching and to summarize the recent progress in this area.

A typical range-searching problem has the following form: Let S be a set of n points in $\mathbb{R}^d$, and let $\mathcal{R}$ be a family of subsets of $\mathbb{R}^d$; elements of $\mathcal{R}$ are called *ranges*. The goal is to preprocess S into a data structure so that for a query range $\gamma \in \mathcal{R}$, the points in $S \cap \gamma$ can be reported or counted efficiently. Typical examples of ranges include rectangles, halfspaces, simplices, and balls. A single query can be answered in linear time using linear space, by simply checking for each point of S whether it lies in the query range. Most applications, however, call for querying the same point set S several times, in which case it is desirable to answer a query faster by preprocessing S into a data structure.

Range counting and range reporting are just two instances of range-searching queries. Other examples include *range-emptiness queries*: determine whether $S \cap \gamma = \emptyset$; and *range-min/max queries*: each point has a weight and one must return the point in the query range with the minimum/maximum weight. Many different types of range queries can be encompassed in the following general formulation of range searching.

Let $(\mathbf{S}, +)$ be a commutative semigroup. Each point $p \in S$ is assigned a weight $w(p) \in \mathbf{S}$. For any subset $S' \subseteq S$, let $w(S') = \sum_{p \in S'} w(p)$, where addition is taken over the semigroup.[1] For a query range $\gamma \in \mathcal{R}$, the goal is to compute $w(S \cap \gamma)$. For example, counting queries can be answered by choosing the semigroup to be $(\mathbb{N}, +)$, where $+$ denotes standard integer addition, and setting $w(p) = 1$ for every $p \in S$; emptiness queries by choosing the semigroup to be $(\{0, 1\}, \vee)$ and setting $w(p) = 1$; reporting queries by choosing the semigroup to be $(2^S, \cup)$ and setting $w(p) = \{p\}$; and range-max queries by choosing the semigroup to be $(\mathbb{R}, \max)$.

The performance of a data structure is measured by the time spent in answering a query, called the *query time*; by the *size* of the data structure; and by the time spent in constructing the data structure, called the *preprocessing time*. Since the data structure is constructed only once, its query time and size are generally more important than its preprocessing time. We should remark that the query time of a range-reporting query on any reasonable machine depends on the output size, so the query time for a range-reporting query consists of two parts—*search time*, which depends only on n and d; and *reporting time*, which depends on n, d, and the output size. Throughout this chapter we will use k to denote the output size.

We assume that d is a small fixed constant, and that big-O and big-Ω notation hide constants depending on d. The size of any range-searching data structure is at least linear, since it has to store each point (or its weight) at least once, and the query time in any reasonable model of computation such as pointer machine, RAM, or algebraic decision tree is $\Omega(\log n)$ even for $d = 1$, assuming the coordinates of input points are real values. Ideally, one would like to develop a linear-size data structure with logarithmic query time. If such a data structure is not feasible, then one seeks a tradeoff between the query time and the size of the data structure—

[1] Since $\mathbf{S}$ need not have an additive identity, we assign a special value *nil* to the empty sum.

How fast can a query be answered using $O(n \operatorname{polylog} n)$ space, how much space is required to answer a query in $O(\operatorname{polylog} n)$ time, and what kind of tradeoff between the size and the query time can be achieved?

The early work on range searching focused on *orthogonal range searching*, where ranges are axis-parallel boxes [22, 23]. Even after four decades of extensive work on orthogonal range searching, some basic questions still remain open; see the survey papers [4, 5]. Geometry plays almost no role in the known data structures for orthogonal range searching. The most basic and most studied truly geometric instance of range searching is with *halfspaces*, or more generally *simplices*, as ranges. We therefore focus on simplex range searching and its variants. We refer to [4] for a recent survey (more comprehensive but less detailed) on range searching, and to [5, 62, 80, 86] for earlier surveys on this topic.

The chapter is organized as follows. We describe, in Sect. 2, different models of computation that have been used to prove upper and lower bounds on the performance of range-searching data structures. Next, Sect. 3 surveys known techniques and data structures for simplex range searching. Section 4 focuses on the special case of halfspace range reporting for which faster data structures are known. Section 5 reviews semialgebraic range searching, where there has been some recent progress using algebraic techniques. Section 6 discusses a few variants and extensions of range searching. We conclude in Sect. 7 by making a few final remarks.

2 Models of Computation

Most algorithms and data structures in computational geometry are implicitly described in the familiar *random access machine* (RAM) model or the *real RAM* model. In the traditional RAM model, memory cells can contain arbitrary $(\log n)$-bit integers, which can be added, multiplied, subtracted, divided (computing $\lfloor x/y \rfloor$), compared, and used as pointers to other memory cells in constant time. In a real RAM, memory cells can store arbitrary real numbers (such as coordinates of points), and basic arithmetic and relational operations between real numbers can be performed in constant time. In the case of range searching over a semigroup other than the integers, memory cells are allowed to contain arbitrary values from the semigroup, but these values can only be added (using the semigroup's addition operator, of course).

All range-searching data structures discussed in this chapter can be described in the more restrictive *pointer machine* model. The main difference between the pointer-machine and the RAM models is that on a pointer machine, a memory cell can be accessed only through a series of pointers, while in the RAM model, any memory cell can be accessed in constant time. In the basic pointer-machine model, a data structure is a directed graph with out-degree 2; each node is associated with a label, which is an integer between 0 and n. Nonzero labels are indices of the points in S, and the nodes with label 0 store auxiliary information. The query algorithm traverses a portion of the graph and visits at least one node with label i for each point p_i in the query range. Chazelle [31] defines generalizations of the pointer-machine model that are more appropriate for answering counting and semigroup queries. In these models, nodes are labeled with arbitrary $O(\log n)$-bit integers, and the query algorithm is allowed to perform arithmetic operations on these integers.

Most lower bounds, and a few upper bounds, are described in the so-called *semigroup arithmetic model*, which was originally introduced by Fredman [60] and refined by Yao [96]. In the semigroup arithmetic model, a data structure can be *informally* regarded as a set of precomputed partial sums in the underlying semigroup. The size of the data structure is the number of sums stored, and the query time is the minimum number of semigroup operations required (on the precomputed sums) to compute the answer to a query. The query time ignores the cost of various auxiliary operations, including the cost of determining which of the precomputed sums should be added to answer a query. Unlike the pointer-machine model, the semigroup model allows immediate access, at no cost, to any precomputed sum. The informal model we have just described is much too powerful. For example, in this informal model, the optimal data structure for counting queries consists of the $n + 1$ integers $0, 1, \ldots, n$. To answer a counting query, we simply return the correct answer; since no additions are required, we can answer queries in zero “time”, using a “data structure” of only linear size! A more formal definition, using the notion of a *faithful semigroup*, that avoids this problem can be found in [32].

A weakness of the semigroup model is that it does not allow subtractions even if the weights of points belong to a group (e.g. range counting). Therefore, we will also consider the *group model*, in which each point is assigned a weight from a commutative group and the goal is to compute the group sum of the weights of points lying in a query range. The data structure consists of a collection of group elements and auxiliary data, and it answers a query by adding and subtracting a subset of the precomputed group elements to yield the answer to the query. The query time is the number of group operations performed. The lower-bound proofs in the semigroup model have a strong geometric flavor because subtractions are not allowed: the query algorithm can use a precomputed sum that involves the weight of a point p only if p lies in the query range. A typical proof basically reduces to arguing that not all query ranges can be “covered” with a small number of subsets of input objects [35]. Unfortunately, no such property holds for the group model, which makes proving lower bounds in the group model much harder. The known lower bounds for range searching in the group model are much weaker than those under the semigroup model.

Many geometric range-searching data structures are constructed by subdividing space into several regions with nice properties and recursively constructing a data structure for each region. Queries are answered with such a data structure by performing a depth-first search through the resulting recursive space partition. The *partition-graph* model, introduced by Erickson [57, 58], formalizes this divide-and-conquer approach. This model can be used to study the complexity of emptiness queries, which are trivial in semigroup and pointer-machine models.

We conclude this section by noting that most of the range-searching data structures discussed in this paper (halfspace range-reporting data structures being a notable exception) are based on the following general scheme. Given a point set S, they precompute a family $\mathcal{F} = \mathcal{F}(S)$ of *canonical subsets* of S and store the weight $w(C) = \sum_{p \in C} w(p)$ of each canonical subset $C \in \mathcal{F}$. For a query range γ, they determine a partition $\mathcal{C}_\gamma = \mathcal{C}(S, \gamma) \subseteq \mathcal{F}$ of $S \cap \gamma$ and add the weights of the subsets in $\mathcal{C}_\gamma$ to compute $w(S \cap \gamma)$. Borrowing terminology from [79], we refer to such a data structure as a *decomposition scheme*. There is a close connection between decomposition schemes and partial sums stored in the semigroup arithmetic model described earlier—$w(C)$, the weight of each canonical subset C, corresponds to a precomputed partial sum.

How exactly the weights of canonical subsets are stored and how $\mathcal{C}_\gamma$ is computed depends on the model of computation and on the specific range-searching problem. In the semigroup (or group) arithmetic model, the query time depends only on the number of canonical subsets in $\mathcal{C}_\gamma$, regardless of how they are computed, so the weights of canonical subsets can be stored in an arbitrary manner. In more realistic models of computation, however, some additional structure must be imposed on the decomposition scheme in order to efficiently compute $\mathcal{C}_\gamma$. In a *hierarchical* decomposition scheme, the weights are stored in a tree T. Each node v of T is associated with a canonical subset $C_v \in \mathcal{F}$, and the children of v are associated with subsets of C_v. Besides the weight of C_v, some auxiliary information is also stored at v, which is used to determine whether $C_v \in \mathcal{C}_\gamma$ for a query range γ. Typically, this auxiliary information consists of some geometric object, which plays the same role as a query region in the partition graph model.

If the weight of each canonical subset can be stored in $O(1)$ memory cells, then the total size of the data structure is just $O(|\mathcal{F}|)$. If the underlying searching problem is a range-reporting problem, however, then the "weight" of a canonical subset is the set itself, and thus it is not realistic to assume that each "weight" requires only constant space. In this case, the size of the data structure is $O(\sum_{C \in \mathcal{F}} |C|)$ if each subset is stored explicitly at each node of the tree. However, the size can be reduced to $O(|\mathcal{F}|)$ by storing the subsets implicitly (e.g., storing points only at leaves).

Finally, let $r \geq 2$ be a parameter, and set $\mathcal{F}_i = \{C \in \mathcal{F} \mid r^{i-1} \leq |C| \leq r^i\}$. A hierarchical decomposition scheme is called *r-convergent* if there exist constants $\alpha \geq 1$ and $\beta \geq 0$ so that the degree of every node in T is $O(r^\alpha)$ and for all $i \geq 1$, $|\mathcal{F}_i| = O((n/r^i)^\alpha)$ and, for all query ranges γ, $|\mathcal{C}_\gamma \cap \mathcal{F}_i| = O((n/r^i)^\beta)$, i.e., the number of canonical subsets in the data structure and in any query output decreases exponentially with their size. The size of the decomposition scheme is $O(n^\alpha)$, provided the weight of each canonical subset can be stored in $O(1)$ space.

To compute $\sum_{p_i \in \gamma} w(p_i)$ for a query range γ using a hierarchical decomposition scheme T, a query procedure performs a depth-first search of T, starting from the root. At each node v, using the auxiliary information stored at v, the procedure determines whether the query range γ contains C_v, intersects C_v, or is disjoint from C_v. If γ contains C_v, then C_v is added to $\mathcal{C}_\gamma$ (rather, the weight of C_v is added to a running counter). Otherwise, if γ intersects C_v, the query procedure identifies a subset of children of v, say $\{w_1, \ldots, w_a\}$, so that the canonical subsets $C_{w_i} \cap \gamma$, for $1 \leq i \leq a$, form a partition of $C_v \cap \gamma$. Then the procedure searches each w_i recursively. If the decomposition scheme is r-convergent, then its query time, under the semigroup model, is $O(n^\beta)$ if $\beta > 0$ and $O(\log n)$ if $\beta = 0$. A decomposition scheme is called *efficient* if for any query range γ, each $\mathcal{C}_\gamma \cap \mathcal{F}_i$ can be computed in time $O\left((n/r^i)^\beta\right)$.

We will see below in Sect. 6 that r-convergent hierarchical decomposition schemes can be cascaded together to construct multi-level structures that answer complex geometric queries.

3 Simplex Range Searching

In this section we focus on simplex range searching, the case in which the query ranges are simplices. No data structure is known that can answer a simplex range query in polylogarithmic time using near-linear storage. In fact, the lower bounds stated below indicate that there is little hope of obtaining such a data structure, since the query time of a linear-size data structure, under the semigroup model, is roughly at least $n^{1-1/d}$ (thus saving only a factor of $n^{1/d}$ over the naïve approach). Since the size and query time of any data structure have to be at least linear and logarithmic, respectively, we consider these two ends of the spectrum: (i) How large should the size of a data structure be in order to answer a query in logarithmic time, and (ii) how fast can a simplex range query be answered using a linear-size data structure. By combining these two extreme cases, as we describe below, a tradeoff between space and query time can be obtained.

3.1 Data Structures with Logarithmic Query Time

The *locus* approach s often used to answer a range query in $O(\log n)$ time, as follows: Let S be a set of weighted points in $\mathbb{R}^d$ and $\mathcal{R}$ a family of ranges (e.g. the set of all halfspaces, or the set of all simplices). If each range $\gamma \in \mathcal{R}$ is specified by b real numbers, then γ can be mapped to a point γ^* in a b-dimensional space, which we denote by $\mathcal{R}^*$. Each input point $p \in S$ is mapped to a region $p^* \subseteq \mathcal{R}^*$ such that $p \in \gamma$ if and only if $\gamma^* \in p^*$. Set $S^* = \{p^* \mid p \in S\}$. $\mathcal{R}^*$ is partitioned into connected "cells" so that all points within each cell τ lie in the same subset S^*_τ of S^*. We store $w_\tau = \sum_{p^* \in S^*_\tau} w(p)$ for each cell τ of the subdivision. A range query with $\gamma \in \mathcal{R}$ then reduces to locating the point γ^* in this subdivision of $\mathcal{R}^*$, identifying the cell τ that contains γ^*, and returning w_τ.

For the sake of simplicity, we first illustrate this approach for the halfspace range-counting problem. We assume that the query halfspace always lies above its bounding hyperplane. We need a few definitions and concepts before we describe the data structures.

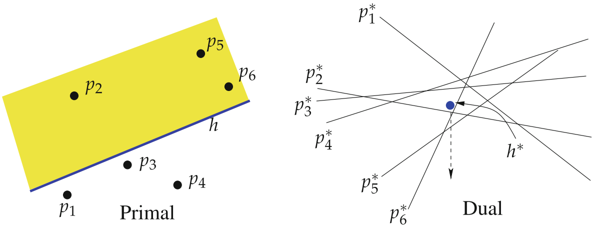


Fig. 1 A halfplange range-counting query in primal and dual

The *dual* of a point $p = (a_1, \ldots, a_d) \in \mathbb{R}^d$ is the hyperplane $p^* : x_d = a_1x_1 + \cdots + a_{d-1}x_{d-1} - a_d$, and the dual of a hyperplane $h : x_d = b_1x_1 + \cdots + b_{d-1}x_{d-1} + b_d$ is the point $h^* = (b_1, \ldots, b_{d-1}, -b_d)$. A nice property of duality is that a point p is above (resp. on below) a hyperplane h if and only if the dual point h^* is above (resp. on below) the dual hyperplane p^*. In the dual setting, the halfspace range-counting problem thus can be formulated as follows: Given a set H of n hyperplanes in $\mathbb{R}^d$, return the number of hyperplanes of H that lie below a query point ξ. See Fig. 1.

The *arrangement* of a set H of hyperplanes in $\mathbb{R}^d$, denoted by $\mathcal{A}(H)$, is the subdivision of $\mathbb{R}^d$ into cells of dimensions k, for $0 \leq k \leq d$, each cell being a maximal connected set contained in the intersection of a fixed subset of H and not intersecting any other hyperplane of H. Since the same subset of hyperplanes lies below all points in a single cell of $\mathcal{A}(H)$, the number of hyperplanes of H lying below a query point ξ can be computed by locating the cell of $\mathcal{A}(H)$ that contains ξ. For $d = 2$, by computing the planar subdivision $\mathcal{A}(H)$ and preprocessing it for planar-point location queries into a data structure of size $O(n^2)$, the cell of $\mathcal{A}(H)$ containing ξ can be computed in $O(\log n)$ time [51]. Point-location queries are, however, more difficult in higher dimensions, and one needs geometric cuttings, defined below.

For a parameter $r \in [1, n]$, a $(1/r)$-*cutting* of H is a set Ξ of (relatively open) disjoint simplices covering $\mathbb{R}^d$ so that at most n/r hyperplanes of H cross (i.e., intersect but do not contain) each simplex of Ξ. Clarkson [45] and Haussler and Welzl [65] were the first to show the existence of a $(1/r)$-cutting of H of size $O(r^d \log^d r)$. Chazelle and Friedman [36] improved the size bound to $O(r^d)$, which is optimal in the worst case. Matoušek [75] was the first to develop an efficient algorithm for constructing a $(1/r)$-cutting. The best algorithm known for computing a $(1/r)$-cutting was discovered by Chazelle [33]; his result is summarized in the following theorem.

Theorem 3.1 (Chazelle [33]) *Let H be a set of n hyperplanes in $\mathbb{R}^d$, let $r \leq n$ be a parameter, and let $b > 1$ be a constant. Set $s = \lceil \log_b r \rceil$. There exist s cuttings $\Xi_1, \ldots, \Xi_s$ so that Ξ_i is a $(1/b^i)$-cutting of size $O(b^{id})$, each simplex of Ξ_i is contained in a simplex of Ξ_{i-1}, and each simplex of Ξ_{i-1} contains a constant number of simplices of Ξ_i. Moreover, $\Xi_1, \ldots, \Xi_s$ can be computed in time $O(nr^{d-1})$.*

The key idea of Chazelle is that Ξ_i is constructed by refining each simplex $\triangle \in \Xi_{i-1}$. Let $H_\triangle \subseteq H$ be the set of hyperplanes that cross $\triangle$, let $n_\triangle = |H_\triangle|$, and let $\chi_\triangle$ be the number of vertices of $\mathcal{A}(H)$ that lie in the interior of $\triangle$. Chazelle's algorithm computes a $(1/b)$-cutting of $H_\triangle$ within $\triangle$ whose size depends on $\chi_\triangle$. More precisely, it partitions $\triangle$, in time $O(n_\triangle b^{O(1)})$, into a set $\Xi_\triangle$ of simplices so that each of them is crossed by at most $n_\triangle/b \le n/b^i$ hyperplanes of H, and more importantly $|\Xi_\triangle| = O(\chi_\triangle(\frac{b^i}{n})^d + b^{d-1})$.

Theorem 3.1 can be used in a straightforward manner to obtain a data structure of size $O(n^d/\log^{d-1} n)$ that can return, in $O(\log n)$ time, the number of hyperplanes of H lying below a query point ξ as follows: Choose $r = \lceil \frac{n}{\log_2 n} \rceil$. Construct the cuttings $\Xi_1, \dots, \Xi_s$, for $s = \lceil \log_2 r \rceil$; for each simplex $\triangle \in \Xi_i$, for $i < s$, store pointers to the simplices of Ξ_{i+1} that are contained in $\triangle$; and for each simplex $\triangle \in \Xi_s$, store $H_\triangle \subseteq H$, the set of hyperplanes that intersect $\triangle$, and $k_\triangle$, the number of hyperplanes of H that lie below $\triangle$. Since $|\Xi_s| = O(r^d)$, the total size of the data structure is $O(nr^{d-1}) = O(n^d/\log^{d-1} n)$. For a query point $\xi \in \mathbb{R}^d$, by traversing the pointers, find the simplex $\triangle \in \Xi_s$ that contains ξ, count in $O(\log n)$ time the number of hyperplanes of $H_\triangle$ that lie below ξ by checking each hyperplane of $H_\triangle$ explicitly, and return $k_\triangle$ plus this quantity. The total query time is $O(\log n)$. Using a linear-size partition tree described in the next subsection for counting the number of hyperplanes of $H_\triangle$ lying below ξ, the size of the data structure can be reduced to $O(n^d/\log^d n)$ while keeping the query time to be $O(\log n)$ [79] (see Sect. 3.3).

The above locus approach for halfspace range counting can be extended to the simplex range-counting problem as well. That is, we map a d-simplex $\triangle$ to a point $\triangle^* \in \mathbb{R}^{d(d+1)}$ and each point $p \in S$ to a region $p^* \subset \mathbb{R}^{d(d+1)}$ such that $p \in \triangle$ if and only if $\triangle^* \in p^*$. Then a simplex range-counting query reduces to point location in the arrangement $\mathcal{A}(S^*)$. Since $\mathcal{A}(S^*)$ has $\Omega(n^{d(d+1)})$ cells, such an approach will require $\Omega(n^{d(d+1)})$ storage; see [49, 54]. More efficient data structures have been developed using the multi-level decomposition scheme mentioned in Sect. 2 and described in Sect. 6.1.

Cole and Yap [49] were the first to present a near-quadratic size data structure that could answer a triangle range-counting query in the plane in $O(\log n)$ time. They present two data structures: the first one answers a query in time $O(\log n)$ using $O(n^{2+\varepsilon})$ space, and the other in time $O(\log n \log\log n)$ using $O(n^2/\log n)$ space. For $d = 3$, their approach gives a data structure of size $O(n^{7+\varepsilon})$ that can answer a tetrahedron range-counting query in time $O(\log n)$. Chazelle et al. [42] describe a multi-level data structure of size $O(n^{d+\varepsilon})$ that can answer a simplex range-counting query in time $O(\log n)$. The space bound can be reduced to $O(n^d)$ by increasing the query time to $O(\log^{d+1} n)$ [79]. These data structures can answer simplex range-reporting queries by spending an additional $O(k)$ time.

3.2 Linear-Size Data Structures

Most of the linear-size data structures for simplex range searching are based on *partition trees*, originally introduced by Willard [94]. Roughly speaking, partition trees are based on the following idea: Given a set S of points in $\mathbb{R}^d$, partition the space into a few, say, a constant number of, regions, each containing roughly equal number of points, so that for any hyperplane h, the number of points lying in the regions that intersect h is much less than the total number of points. Then recursively construct a similar partition for the subset of points lying in each region.

Willard's original partition tree for a set S of n points in the plane is a 4-way tree, constructed as follows. Let us assume that n is of the form 4^k for some integer k, and that the points of S are in general position. If $k = 0$, the tree consists of a single node that stores the coordinates of the only point in S. Otherwise, using the ham-sandwich theorem [81], find two lines ℓ_1, ℓ_2 so that each quadrant Q_i, for $1 \leq i \leq 4$, induced by ℓ_1, ℓ_2 contains exactly $n/4$ points. The root stores the equations of ℓ_1, ℓ_2 and the value of n. For each quadrant, recursively construct a partition tree for $S \cap Q_i$ and attach it as the ith subtree of the root. The total size of the data structure is linear, and it can be constructed in $O(n \log n)$ time. A halfplane range-counting query can be answered by refining the generic procedure described in Sect. 2, as follows: Let h be a query halfplane. Traverse the tree, starting from the root, and maintain a global count. At each node v, perform the following step: If the line ∂h intersects the quadrant Q_v associated with v, recursively visit the children of v. If $Q_v \cap h = \emptyset$, do nothing. Otherwise, since $Q_v \subseteq h$, add the number of points of S lying in the subtree rooted at S_v to the global count. The quadrants associated with the four children of any interior node are induced by two lines, so ∂h intersects at most three of them, which implies that the query procedure does not explore the subtree of one of the children. Hence, the query time of this procedure is $O(n^\alpha)$, where $\alpha = \log_4 3 \leq 0.7925$. A similar procedure can answer a triangle range-counting query within the same time bound, and a triangle range-reporting query in time $O(n^\alpha + k)$. Edelsbrunner and Welzl [56] described a simple variant of Willard's partition tree that improved the exponent in the query-search time to $\log_2(1 + \sqrt{5}) - 1 \approx 0.695$.

A partition tree for points in $\mathbb{R}^3$ was first proposed by Yao [95], which can answer a counting query in time $O(n^{0.98})$. Using the Borsuk-Ulam theorem (see the monograph by Matoušek [81]), Yao et al. [97] showed that, given a set S of n points in $\mathbb{R}^3$, one can find three planes so that each of the eight (open) octants determined by them contains at most $\lfloor n/8 \rfloor$ points of S. This approach leads to a linear-size data structure with query time $O(n^{0.899})$. Avis [20] proved that such a partition of $\mathbb{R}^d$ by d hyperplanes is not always possible for $d \geq 5$; the problem is still open for $d = 4$. Weaker partitioning schemes for $d \geq 4$ were proposed in [48, 98].

After the initial improvements and extensions on Willard's partition tree, a major breakthrough was made by Haussler and Welzl [65]. They formulated range searching in an abstract setting and, using elegant probabilistic methods, gave a randomized algorithm to construct a linear-size partition tree with $O(n^\alpha)$ query time, where $\alpha = 1 - \frac{1}{d(d-1)+1} + \varepsilon$ for any $\varepsilon > 0$. Besides an improved range searching data structure, the major contribution of their paper is the abstract framework and the notion of ε-nets. This and related abstract frameworks have popularized randomized algorithms in computational geometry [51].

The first linear-size data structure with near-optimal query time for simplex range queries in the plane was developed by Welzl [93]. His algorithm is based on the following idea. A *spanning path* of a set S of points is a polygonal chain whose vertices are the points of S. The *crossing number* of a polygonal path is the maximum number of its edges that can be crossed by a hyperplane (i.e., the endpoints of the edge lie in opposite open halfspaces bounded by the hyperplane). Welzl constructs a spanning path $\Pi = \Pi(S)$ of any set S of n points in $\mathbb{R}^d$ whose crossing number is $O(n^{1-1/d} \log n)$. The bound on the crossing number was improved by Chazelle and Welzl [43] to $O(n^{1-1/d})$, which is tight in the worst case.[2] Let $p_1, p_2, \ldots, p_n$ be the vertices of Π. If we know the edges of Π that cross h, then the weight of points lying in one of the halfspaces bounded by h can be computed by answering $O(n^{1-1/d})$ partial-sum queries on the sequence $W = \langle w(p_1), \ldots, w(p_n) \rangle$. Hence, by processing W for partial-sum queries, we obtain a linear-size data structure for simplex range searching, with $O(n^{1-1/d}\alpha(n))$ query time, in the semigroup arithmetic model, where $\alpha(n)$ is the inverse Ackermann function. (Recall that the time spent in finding the edges of Π crossed by h is not counted in the semigroup model.) In any realistic model of computation such as pointer machines or RAMs, however, we also need an efficient linear-size data structure for computing the edges of Π crossed by a hyperplane. Chazelle and Welzl [43] produced such a data structure for $d \leq 3$, but no such structure is known for higher dimensions.

The first data structure with roughly $n^{1-1/d}$ query time and near-linear space, for $d > 3$, was obtained by Chazelle et al. [42]. Given a set S of n points in $\mathbb{R}^d$ and a parameter $r > 1$, they construct a family $\mathcal{F} = \{\Xi_1, \ldots, \Xi_k\}$ of triangulations of $\mathbb{R}^d$, each of size $O(r^d)$. For any hyperplane h, there is at least one Ξ_i so that only $O(n/r)$ points lie in the simplices of Ξ_i that intersect h. Applying this construction recursively, they obtain a tree structure of size $O(n^{1+\varepsilon})$ that can answer a halfspace range-counting query in time $O(n^{1-1/d+\varepsilon})$. The extra n^{ε} factor in the space is due to the fact that they maintain a family of partitions instead of a single partition. Another consequence of maintaining a family of partitions is that, unlike partition trees, this data structure cannot be used directly to answer simplex range queries. Instead, they construct a multi-level data structure (Sect. 6.1) to answer simplex range queries.

Matoušek [79] developed a simpler, slightly faster data structure for simplex range queries, by returning to the theme of constructing a single partition, as in the earlier partition-tree papers (though unlike earlier papers, the regions associated with the children of a node in his partition tree could overlap). His algorithm is based on the following partition theorem, which can be regarded as an extension of the results by Welzl [93] and Chazelle and Welzl [43].

[2] Given a point set, the problem of computing a spanning path with the minimum corssing number is NP-Complete [59].

Theorem 3.2 (Matoušek [76]) *Let S be a set of n points in $\mathbb{R}^d$, and let $1 < r \leq n/2$ be a given parameter. Then there exists a family of pairs $\Pi = \{(S_1, \Delta_1), \ldots, (S_m, \Delta_m)\}$ such that each $S_i \subseteq S$ lies inside the simplex Δ_i, $n/r \leq |S_i| \leq 2n/r$, $S_i \cap S_j = \emptyset$ for all $i \neq j$, and every hyperplane crosses at most $cr^{1-1/d}$ simplices of Π; here c is a constant. If $r \leq n^\alpha$ for some suitable constant $0 < \alpha < 1$, then Π can be constructed in $O(n \log r)$ time.*

Note that although S is being partitioned into a family of subsets, unlike the earlier results on partition trees, it does not partition $\mathbb{R}^d$ because Δ_i's may intersect. Theorem 3.2 is proved by constructing a "test set" Γ of $O(r^d)$ hyperplanes so that if each hyperplane in Γ crosses $O(r^{1-1/d})$ simplices in Π, then the same holds for an arbitrary hyperplane [76]. Given Γ and S, the pairs in Π are constructed one by one using an iterative reweighing scheme that assigns weights to the hyperplanes in Γ and computes a (weighted) $(1/cr^d)$-cutting of Γ at each stage, analogous to the construction of the spanning path in [43, 93]; see [76] for details.

Using Theorem 3.2, a partition tree T can be constructed as follows. Each interior node v of T is associated with a canonical subset $C_v \subseteq S$ and a simplex Δ_v containing C_v; if v is the root of T, then $C_v = S$ and $\Delta_v = \mathbb{R}^d$. Choose r to be a sufficiently large constant. If $|S| \leq 4r$, T consists of a single node, and it stores all points of S. Otherwise, v stores Δ_v and $|C_v|$, and we construct a family of pairs $\Pi_v = \{(S_1, \Delta_1), \ldots, (S_m, \Delta_m)\}$ using Theorem 3.2. We recursively construct a partition tree T_i for each S_i and attach T_i as the ith subtree of v. The total size of the data structure is linear, and it can be constructed in time $O(n \log n)$. A simplex range-counting query can be answered in the same way as with Willard's partition tree. Since any hyperplane intersects at most $cr^{1-1/d}$ simplices of Π, the query time is $O(n^{1-1/d+\log_r c})$; the $\log_r c$ term in the exponent can be reduced to any arbitrarily small positive constant ε by choosing r sufficiently large. The query time can be improved to $O(n^{1-1/d} \operatorname{polylog} n)$ by choosing $r = n^\varepsilon$.

In a subsequent paper, Matoušek [79] proved a stronger version of Theorem 3.2, using some additional sophisticated techniques (including Theorem 3.1), that gives a linear-size partition tree with $O(n^{1-1/d})$ query time. His scheme was subsequently simplified by Chan [28]. Unlike the recursive construction of the partition tree described above, Chan's algorithm does not construct the subtree at each node of the partition tree independently. Instead, roughly speaking, it constructs the partition tree level by level, where each level corresponds to a triangulation that "covers" all points of S (i.e., each point of S lie in a simplex of the triangulation), and each step partitions one of the simplices of the previous level into subsimplices. Therefore the triangulation at level $i+1$ is a refinement of the triangulation at level i, and the entire tree corresponds to a hierarchical triangulation in $\mathbb{R}^d$. The main result in [28] is the following theorem:

Theorem 3.3 (Chan [28]) *Let S be a set of n points in $\mathbb{R}^d$, H a set of m hyperplanes in $\mathbb{R}^d$, and $b \geq 1$ a parameter. Given t disjoint simplices that cover P such that each simplex contains at most $2n/t$ points of S and each hyperplane of H crosses at most ℓ cells, each simplex can be partitioned into $O(b)$ (disjoint) subsimplices, for a total of $O(bt)$ simplices such that each subsimplex contains at most $\frac{2n}{bt}$ points* *of S and each hyperplane of H crosses $O((bt)^{1-1/d} + b^{1-1/(d-1)}\ell + b \log t \log m)$ subsimplices.*

Note that the first term is the same as in Theorem 3.2. It is crucial that the coefficient $b^{1-1/(d-1)}$ in the second term is asymptotically smaller than $b^{1-1/d}$, for this ensures that the repeated application of this theorem does not incur a constant-factor blow up, as in the recursive construction based on Theorem 3.2. The partition tree based on Theorem 3.3 can be constructed in $O(n \log n)$ randomized expected time. An interesting byproduct of Chan's technique is that it can be used to construct a triangulation of S so that any hyperplane crosses the interior of $O(n^{1-1/d})$ simplices, thereby solving a long-standing open problem.

If the points in S lie on a k-dimensional algebraic surface of constant degree, the *crossing number* in Theorem 3.2 can be improved to $O(r^{1-1/\gamma})$, where $\gamma = 1/\lfloor (d+k)/2 \rfloor$ [7], which implies that in this case a simplex range query can be answered in time $O(n^{1-1/\gamma+\varepsilon})$ using linear space.

3.3 Trading Space for Query Time

In the previous two subsections we surveyed data structures for simplex range searching that either use near-linear space or answer a query in polylogarithmic time. By combining these two types of data structures, a tradeoff between the size and the query time can be obtained [9, 28, 42, 79]. For example, the hierarchical-cutting based data structure of size $O(n^d/\log^{d-1} n)$ and $O(\log n)$ query time by Chazelle [33] and the linear-size partition tree with $O(\sqrt{n})$ query time by Chan [28] can be combined to construct a halfspace range-countig data structure of size m and $O(\log(m/n) + n/m^{1/d})$ query time, for $n \le m \le n^d/\log^d n$, as follows: Choose a parameter $r = \lceil (\frac{m}{n})^{1/(d-1)} \rceil$. Construct a hierarchical $(1/r)$-cutting Ξ of size $O(r^d)$ on S^*, the set of hyperplanes dual to the points in S, using Theorem 3.1, and build a tree data structure $\mathcal{T}$ as described in Sect. 3.1. For each simplex $\triangle \in \Xi$, let $S_\triangle \subseteq S$ be the set of points who dual hyperplanes intersect $\triangle$; $|S_\triangle| = O(n/r)$. We construct Chan's partition tree $\mathcal{T}_\triangle$ on $S_\triangle$. The overall size of the data structure is $O(r^d + \sum_{\triangle \in \Xi} |S_\triangle|) = O(nr^{d-1}) = O(m)$. See Fig. 2. For a given query halfspace h^+ lying above the hyperplane h, we first query the top tree $\mathcal{T}$ with the point h^*, dual to h, and compute in $O(\log r)$ time: (i) the simplex $\triangle \in \Xi$ that contains h^*, and (ii) $\kappa_\triangle$, the number of hyperplanes of S^* lying below $\triangle$. Next, we query the partition tree on $S_\triangle$ and compute, in $O(|S_\triangle|^{1-1/d})$ time, $\sigma_{h,\triangle} = |S_\triangle \cap h^+|$. We return $\kappa_\triangle + \sigma_{h,\triangle}$ as $|S \cap h^+|$. The overall query time is $O(\log r + (n/r)^{1-1/d}) = O(\log(m/n) + n/m^{1/d})$. Note that by choosing $r = \frac{n}{\log^{d/(d-1)} n}$, we obtain a data structure of $O(n^d/\log^d n)$ with $O(\log n)$ query time, as mentioned in Sect. 3.1.

The approach just described is very general and works for any geometric-searching data structure that can be viewed as a hierarchical decomposition scheme (described in Sect. 2), provided it satisfies certain assumptions. We state the general

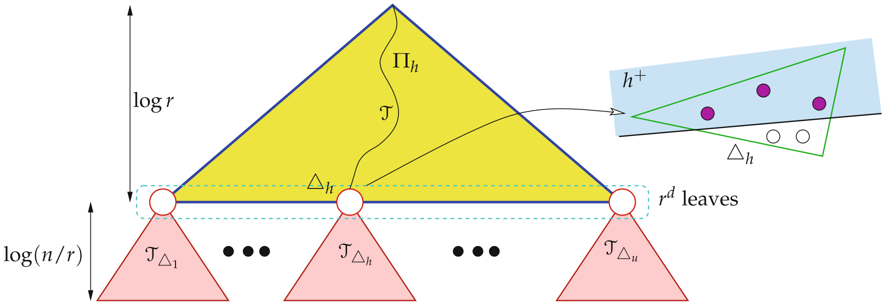


Fig. 2 Space/query-time trade-off for halfspace range counting. The top tree, constructed using hierarchical cutting, has $O(r^d)$ leaves, each corresponding to a simplex of the cutting Ξ, and each bottom tree $\mathcal{T}_\triangle$ built on $S_\triangle$, using Chan's algorithm, is of size $O(n/r)$; Π_h is the path in $\mathcal{T}$ followed by the query procedure and $\triangle_h$ is the simplex of Ξ that contains the query point h^*

result here, though a slightly better bounds (by a polylogarithmic factor) can be obtained by exploiting special properties of the data structures.

Theorem 3.4 *Let S be a set of n points in $\mathbb{R}^d$, and let r be a sufficiently large constant. Let $\mathcal{P}$ be a range-searching problem. Let $\mathcal{D}^1$ be a decomposition scheme for $\mathcal{P}$ of size $O(n^\alpha)$ and query time $O(\log n)$, and let $\mathcal{D}^2$ be another decomposition scheme of size $O(n)$ and query time $O(n^\beta)$. If either $\mathcal{D}^1$ or $\mathcal{D}^2$ is hierarchical, efficient, and r-convergent, then for any $n \le m \le n^\alpha$, then a decomposition scheme for $\mathcal{P}$ of size $O(m)$ can be constructed that has $O((\frac{n^\alpha}{m})^{\beta/(\alpha-1)} + \log \frac{m}{n})$ query time.*

For the d-dimensional halfspace range-counting problem, for example, we have $\alpha = d$ and $\beta = 1 - 1/d$. Thus, for any $n \le m \le n^d$, a halfspace range-counting query can be answered in time $O(n/m^{1/d} + \log(m/n))$ using $O(m)$ space.

We conclude this discussion by making a few remarks on Theorem 3.4.

(i) Theorem 3.4 can be refined to balance polylogarithmic factors in the sizes and query times of $\mathcal{D}^1$ and $\mathcal{D}^2$. For example, if the size of $\mathcal{D}^1$ is $O(n^\alpha \operatorname{polylog} n)$ and rest of the parameters are the same as in the theorem, then the query time of the new data structure is $O((\frac{n^\alpha}{m})^{\beta/(\alpha-1)} \operatorname{polylog}(\frac{m}{n}))$. For example, Matoušek [79] showed that for any $n \le m \le n^d$, a simplex range-counting query can be answered in time $O((n/m^{1/d}) \log^{d+1}(m/n))$ using $O(m)$ space.

(ii) it is not essential for $\mathcal{D}^1$ or $\mathcal{D}^2$ to be tree-based data structures. It is sufficient to have an efficient, r-convergent decomposition scheme with a partial order on the canonical subsets.

3.4 Lower Bounds

Fredman [61] showed that a sequence of n insertions, deletions, and halfplane queries on a set of points in the plane requires $\Omega(n^{4/3})$ time, in the semigroup model. His technique, however, does not extend to static data structures. In a series of papers, Chazelle has proved nontrivial lower bounds on the complexity of online simplex range searching, using various elegant mathematical techniques. He showed that any data structure of size m, for $n \leq m \leq n^d$, for simplex range searching in the semigroup model requires a query time of $\Omega(n/\sqrt{m})$ for $d = 2$ and $\Omega(n/(m^{1/d} \log n))$ for $d \geq 3$ in the worst case. It should be pointed out that this theorem holds even if the query ranges are wedges or strips, but not if the ranges are halfspaces. For halfspaces, Brönnimann et al. [24] proved a lower bound of $\Omega((\frac{n}{\log n})^{1-\frac{d-1}{d(d+1)}}/m^{1/d})$ on the query time of any data structure that uses $O(m)$ space, under the semigroup model. Arya et al. [19] improved the lower bound to $\Omega((\frac{n}{\log n})^{1-\frac{1}{d+1}}/m^{\frac{1}{d+1}})$. They also showed that if the semigroup is integral (i.e., for all non-zero elements of the semigroup and for all $k \geq 2$, the k-fold sum $x + \cdots + x \neq x$), then the lower bound can be improved to $\Omega(\frac{n}{m^{1/d}}/\log^{1+\frac{2}{d}} n)$.

A few lower bounds on simplex range searching have been proved under the group model. Chazelle [34] proved an $\Omega(n \log n)$ lower bound for off-line halfspace range searching (i.e., the time taken to answer n queries over a set of n points) under the group model. Exploiting a close-connection between range searching and discrepancy theory, Larsen [71] showed that for any dynamic data structure with t_u and t_q worst-case update and query time, respectively, $t_u \cdot t_q = \Omega(n^{1-1/d})$.

Chazelle and Rosenberg [41] proved a lower bound of $\Omega(\frac{n^{1-\varepsilon}}{m} + k)$ for simplex range reporting under the pointer-machine model. Afshani [1] improved their bound slightly by proving that the size of any data structure that answers a simplex range reporting query in time $O(t_q + k)$ is $\Omega((\frac{n}{t_q})^d/2^{O(\sqrt{\log t_q})})$.

A series of papers by Erickson established the first nontrivial lower bounds for on-line and off-line emptiness query problems, in the partition-graph model of computation. He first considered this model for Hopcroft's problem—Given a set of n points and m lines, does any point lie on a line?—for which he obtained a lower bound of $\Omega(n \log m + n^{2/3}m^{2/3} + m \log n)$ [58], almost matching the best known upper bound $O(n \log m + n^{2/3}m^{2/3}2^{O(\log^*(n+m))} + m \log n)$, due to Matoušek [79]. He later established lower bounds on a trade-off between space and query time, or preprocessing and query time, for on-line hyperplane emptiness queries. For d-dimensional hyperplane queries, $\Omega(n^d/\operatorname{polylog} n)$ preprocessing time is required to achieve polylogarithmic query time, and the best possible query time is $\Omega(n^{1-1/d}/\operatorname{polylog} n)$ if $O(n \operatorname{polylog} n)$ preprocessing time is allowed. For $d = 2$, if the preprocessing time is t_p, the query time is $\Omega(n/\sqrt{t_p})$.

4 Halfspace Range Reporting

A halfspace range-reporting query can be answered more quickly than a simplex range-reporting query using the so-called shallow cuttings and the filtering search technique. We begin by considering a simpler problem: the *halfspace-emptiness*

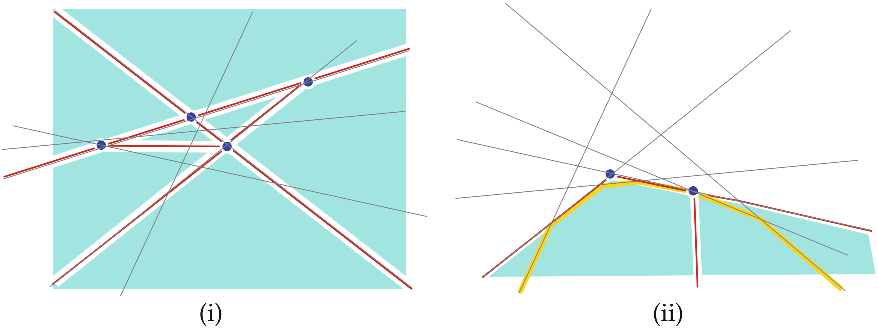


Fig. 3 (**i**) A (2/5)-cutting with seven triangles; (**ii**) a shallow $(0, 2/5)$-cutting with two triangles covering points that lie below all lines of H (lying below the *gold polygonal curve*)

query, which asks whether a query halfspace contains any input point. For simplicity, as in the previous section, assume the query halfspace to lie above its bounding hyperplane. By the duality transform, the halfspace-emptiness query in $\mathbb{R}^d$ can be formulated as asking whether a query point $q \in \mathbb{R}^d$ lies below all hyperplanes in a given set H of hyperplanes in $\mathbb{R}^d$. This query is equivalent to asking whether q lies inside a convex polyhedron $\mathcal{P}(H)$, defined by the intersection of halfspaces lying below the hyperplanes of H. For $d \leq 3$, a point-location query in $\mathcal{P}(H)$ can be answered optimally in $O(\log n)$ time using $O(n)$ space and $O(n \log n)$ preprocessing since $\mathcal{P}(H)$ has linear size [51]. For $d \geq 4$, point-location query in $\mathcal{P}(H)$ becomes more challenging. Clarkson [45] had described a random-sampling based data structure of size $O(n^{\lfloor d/2 \rfloor + \varepsilon})$ that could anser a point-location query in $O(\log n)$ time. Here we describe an approach, based on the concept of shallow-cutting, which can be viewed as a refinement of Clarkson's approach. Given a parameter $r \in [1, n]$ and an integer $q \in [0, n-1]$, a *shallow* $(q, 1/r)$*-cutting* of H is a set Ξ of (relatively open) disjoint simplices covering all points that lie above at most q hyperplanes of H so that at most n/r hyperplanes of H cross each simplex of Ξ (see Fig. 3(ii)).

The following theorem by Matoušek can be used to construct a point-location data structure for $\mathcal{P}(H)$:

Theorem 4.1 (Matoušek [77]) *Let H be a set of n hyperplanes and $r \leq n$ a parameter. A shallow $(0,1/r)$-cutting of H of size $O(r^{\lfloor d/2 \rfloor})$ can be computed in time $O(nr^c)$, where c is a constant.*

Choose r to be a sufficiently large constant and compute a shallow $(0,1/r)$-cutting Ξ of H using Theorem 4.1. For each simplex $\triangle \in \Xi$, let $H_\triangle \subseteq H$ be the set of hyperplanes that intersect $\triangle$. Recursively, construct the data structure for $H_\triangle$; the recursion stops when $|H_\triangle| \leq r$. The size of the data structure is $O(n^{\lfloor d/2 \rfloor + \varepsilon})$, where $\varepsilon > 0$ is an arbitrarily small constant, and it can be constructed in $O(n^{\lfloor d/2 \rfloor + \varepsilon})$ time. If a query point q does not lie in a simplex of Ξ, then one can conclude that $q \notin \mathcal{P}(H)$ and thus stop. Otherwise, if q lies in a simplex $\triangle \in \Xi$, recursively determine whether q lies below all the hyperplanes of $H_\triangle$. The query time is

$O(\log n)$. Matoušek and Schwarzkopf [83] showed that the space can be reduced to $O(\frac{n^{\lfloor d/2\rfloor}}{\log^{\lfloor d/2\rfloor-\varepsilon} n})$ while keeping the query time to $O(\log n)$.

A linear-size data structure can be constructed for answering halfspace-emptiness queries by constructing a simplicial partition analogous to Theorem 3.2, as follows. A hyperplane h is called *λ-shallow* if one of the halfspaces bounded by h contains at most λ points of S.

Theorem 4.2 (Matoušek [77]) *Let S be a set of n points in $\mathbb{R}^d$ and let $1 \le r < n$ be a given parameter. Then there exists a family of pairs $\Pi = \{(S_1, \Delta_1), \ldots, (S_m, \Delta_m)\}$ such that each $S_i \subseteq S$ lies inside the simplex Δ_i, $n/r \le |S_i| \le 2n/r$, $S_i \cap S_j = \emptyset$ for all $i \ne j$, and the number of simplices of Π crossed by any (n/r)-shallow hyperplane is $O(\log r)$ for $d \le 3$ and $O(r^{1-1/\lfloor d/2\rfloor})$ for $d > 3$. If $r \le n^\alpha$ for some suitable constant $0 < \alpha < 1$, then Π can be constructed in $O(n\log r)$ time.*

Using this theorem, a partition tree for S can be constructed in the same way as for simplex range searching, by choosing $r = n^\alpha$ for some $\alpha < 1/d$. While answering a query for a halfspace h^+, if h^+ crosses more than $O(r^{1-1/\lfloor d/2\rfloor})$ simplices of the partition Π_v associated with a node v, then we conclude that h^+ is not (n/r)-shallow, i.e., $h^+ \cap S \ne \emptyset$, and thus we stop. Similarly if for a pair (Δ_i, S_i), $\Delta_i \subseteq h^+$, we conclude that $h^+ \cap S \ne \emptyset$ and we stop. Otherwise, the procedure recursively visits the children of v corresponding to the pairs (Δ_i, S_i) for which h crosses Δ_i. The size of the data structure is $O(n)$, the query time is $O(n^{1-1/\lfloor d/2\rfloor}$ polylog $n)$, and the preprocessing time is $O(n\log n)$. The query time can be improved to $n^{1-1/\lfloor d/2\rfloor}2^{O(\log^* n)}$ by increasing the preprocessing time to $O(n^{1+\varepsilon})$. For even dimensions, Chan's approach [28] can be extended to construct a linear-size partition tree with query time $O(n^{1-1/\lfloor d/2\rfloor})$.

The halfspace-emptiness data structures can be adapted to answer halfspace range-reporting queries, using the so-called *filtering search* technique introduced by Chazelle [30]. All the data structures mentioned above answer a range-reporting query in two stages. The first stage "identifies" the k points of a query output, in time $f(n)$ that is independent of the output size, and the second stage explicitly reports these points in $O(k)$ time. Chazelle observes that since $\Omega(k)$ time will be spent in reporting k points, the first stage can compute in $f(n)$ time a superset of the query output of size $O(k)$, and the second stage can "filter" the actual k points that lie in the query range. This observation not only simplifies the data structure but also gives better bounds in many cases, including halfspace range reporting. See [2, 12, 30, 77] for some applications of filtering search.

An optimal halfspace reporting data structure in the plane was proposed by Chazelle et al. [38]. They compute *convex layers* $L_1, \ldots, L_m$ of S, where L_i is the set of points lying on the boundary of the convex hull of $S \setminus \bigcup_{j<i} L_j$, and store them in a linear-size data structure, so that a query can be answered in $O(\log n + k)$ time. For $d = 3$, after a series of papers with successive improvements (see e.g. [13, 40, 89]), a linear-size data structure with $O(\log n + k)$ query time was proposed by Afshani and Chan [2] who combine the shallow-cutting with Theorem 4.2 cleverly; this structure can be constructed in $O(n\log n)$ time [29].

For $d > 3$, the linear-size data partition tree for halfspace-emptiness queries can be adapted for answering reporting queries as follows. For each node v of the partition tree, we also preprocess the corresponding canonical subset S_v for simplex range searching and store the resulting partition tree as a secondary data structure of v. Because of the simplex range searching data structure being stored at each node of the tree, the total size of the data structure is $O(n \log \log n)$. For a query halfspace h^+, if its boundary hyperplane h crosses more than $O(r^{1-1/\lfloor d/2 \rfloor})$ simplices of Π_v at a node v, then h is not (n_v/r)-shallow, and the simplex range-reporting data structure stored at v is used to report all k_v points of $S_v \cap h^+$ in time $O(n_v^{1-1/d} + k_v) = O(k_v)$; the last equality follows because $r \leq n_v^{1/d}$ and $k_v = \Omega(n_v/r) = \Omega(n_v^{1-1/d})$. Otherwise, for each pair $(S_i, \Delta_i) \in \Pi_v$, if $\Delta_i \subseteq h^+$, it reports all points S_i; and if Δ_i is crossed by h, it recursively visits the corresponding child of v. The overall query time is $O(n^{1-1/\lfloor d/2 \rfloor} \operatorname{polylog}(n) + k)$. The size of the data structure was improved to $O(n)$ by Afshani and Chan [2], and the query time for even values of d was improved to $O(n^{1-1/\lfloor d/2 \rfloor} + k)$ by Chan [28].

The technique by Afshani for simplex range-reporting lower bound [1] also shows that the size of any halfspace range-reporting data structure in dimension $d(d+3)/2$ with query time t_q has size $\Omega((\frac{n}{t_q})^d / 2^{O(\sqrt{\log t_q})})$.

Finally, we comment that halfspace-emptiness data structures have been adapted to answer halfspace range-counting queries approximately [2, 14, 15, 64, 68, 88]. For example, a set S of n points in $\mathbb{R}^3$ can be preprocessed, in $O(n \log n)$ time, into a linear-size data structure that for a query halfspace γ in $\mathbb{R}^3$, can report in $O(\log n)$ time a number t such that $|\gamma \cap S| \leq t \leq (1+\delta)|\gamma \cap S|$, where $\delta > 0$ is a constant [2, 3]. For $d > 3$, such a query can be answered in $O((\frac{n}{t})^{1-1/\lfloor d/2 \rfloor} \operatorname{polylog}(n))$ time using linear space [88].

5 Semialgebraic Range Searching

So far we assumed that the ranges were bounded by hyperplanes, but many applications involve ranges bounded by nonlinear functions. For example, a query of the form "for a given point p and a real number r, find all points of S lying within distance r from p" is a range-searching problem in which ranges are balls. As shown below, range searching with balls in $\mathbb{R}^d$ can be formulated as an instance of halfspace range searching in $\mathbb{R}^{d+1}$. So a ball range-reporting (resp. range-counting) query in $\mathbb{R}^d$ can be answered in time $O((n/m^{1/\lceil d/2 \rceil}) \operatorname{polylog} n + k)$ (resp. $O((n/m^{1/(d+1)}) \log(m/n))$), using $O(m)$ space. (Somewhat better performance can be obtained using a more direct approach, which we will describe shortly.) However, data structures for more general ranges seem more challenging.

A natural class of more general ranges can be defined as follows. A *semialgebraic set* is a subset of $\mathbb{R}^d$ obtained from a finite number of sets of the form $\{x \in \mathbb{R}^d \mid g(x) \geq 0\}$, where g is a d-variate polynomial with real coefficients, by Boolean operations (union, intersection, and complement). Specifically, let $\Gamma_{d,\Delta,s}$ denote the family of all semialgebraic sets in $\mathbb{R}^d$ defined by at most s polynomial inequalities of degree at most Δ each. If d, Δ, s are all constants, we refer to the sets in $\Gamma_{d,\Delta,s}$ as *constant-complexity semialgebraic sets*; such sets are sometimes also called *Tarski cells*. The range-searching problem in which query ranges belong to $\Gamma_{d,\Delta,s}$ for constants d, Δ, s, is referred to as *semialgebraic range searching*.

It suffices to consider the ranges bounded by a single polynomial because the ranges bounded by multiple polynomials can be handled using multi-level data structures (see Sect. 6.1). We assume that the ranges are of the form

$$\gamma_f(a) = \{x \in \mathbb{R}^d \mid f(a,x) \geq 0\},$$

where f is a $(d+b)$-variate polynomial specifying the type of range (disks, cylinders, cones, etc.), and a is a b-tuple specifying a specific range of the given type (e.g., a specific disk). Let $\Gamma_f = \{\gamma_f(a) \mid a \in \mathbb{R}^b\}$. We will refer to the range-searching problem in which the ranges are from the set Γ_f as *Γ_f-range searching*. We describe two approaches for Γ_f-range searching.

Linearization One approach to answer Γ_f-range queries is to use *linearization*, originally proposed by Yao and Yao [98]. We represent the polynomial $f(a,x)$ in the form

$$f(a,x) = \psi_0(a)\varphi_0(x) + \psi_1(a)\varphi_1(x) + \cdots + \psi_\ell(a)\varphi_\ell(x)$$

where $\varphi_0, \ldots, \varphi_\ell, \psi_0, \ldots, \psi_\ell$ are polynomials. A point $x \in \mathbb{R}^d$ is mapped to the point

$$\varphi(x) = [\varphi_0(x), \varphi_1(x), \varphi_2(x), \ldots, \varphi_\ell(x)] \in \mathbb{R}^\ell,$$

represented in homogeneous coordinates. Then each range $\gamma_f(a) = \{x \in \mathbb{R}^d \mid f(x,a) \geq 0\}$ is mapped to a halfspace

$$\psi^{\#}(a) : \{y \in \mathbb{R}^\ell \mid \psi_0(a)y_0 + \psi_1(a)y_1 + \cdots + \psi_\ell(a)y_\ell \geq 0\},$$

where, again, $[y_0, y_1, \ldots, y_\ell]$ are homogeneous coordinates. The constant ℓ is called the *dimension* of the linearization. Agarwal and Matoušek [7] have described an algorithm for computing a linearization of the smallest dimension under certain assumptions on φ_i's and ψ_i's.

A well-known example of linearization is the so-called *lifting transform*, which maps a ball in $\mathbb{R}^d$ to a halfspace in $\mathbb{R}^{d+1}$. A ball in $\mathbb{R}^d$ with center $(a_1, \ldots, a_d)$ and radius a_{d+1} can be regarded as a set of the form $\gamma_f(a)$, where $a = (a_1, \ldots, a_d, a_{d+1})$ and f is a $(2d+1)$-variate polynomial

$$f(a_1, \ldots, a_{d+1}; x_1, \ldots, x_d) = -\sum_{i=1}^{d}(x_i - a_i)^2 + a_{d+1}^2.$$

This polynomial can be linearized in $d + 1$ dimensions by the following set of polynomials:

$$\begin{array}{ll}\psi_0(a) = a_{d+1}^2 - \sum_{i=1}^{d} a_i^2, & \psi_i(a) = 2a_i \; i = 1, \ldots, d, \; \psi_{d+1}(a) = -1, \\ \varphi_0(x) = 1, & \varphi_i(x) = x_i \quad i = 1, \ldots, d, \; \varphi_{d+1}(x) = \sum_{i=1}^{d} x_i^2.\end{array}$$

Another popular linearization in computational geometry is for lines in $\mathbb{R}^3$. Based on the so-called *Plücker coordinates*, it maps a line in $\mathbb{R}^3$ to a hyperplane in $\mathbb{R}^5$;

Using linearization, a Γ_f-range query can now be answered using an ℓ-dimensional halfspace range-searching data structure. Thus, for counting queries, we immediately obtain a linear-size data structure with query time $O(n^{1-1/\ell})$ [79], or a data structure of size $O(n^\ell / \log^\ell n)$ with logarithmic query time [33]. For $d < \ell$, the performance of the linear-size data structures can be improved by exploiting the fact that the points $\varphi(x)$ have only d degrees of freedom. As mentioned at the end of Sect. 3.2, the query time in this case can be reduced to $O(n^{1-1/\lfloor(d+\ell)/2\rfloor+\varepsilon})$. It is an open question whether one can similarly exploit the fact that the halfspaces $\psi^{\#}(a)$ have only b degrees of freedom to reduce the size of data structures with logarithmic query time when $b < \ell$.

Algebraic methods In cases where the linearization dimension is very large, semialgebraic queries can also be answered using the following more direct approach proposed by Agarwal and Matoušek [7]. Let S be a set of n points in $\mathbb{R}^d$. For each point p_i, we can define a b-variate polynomial $g_i(a) \equiv f(p_i, a)$. Then $\Gamma_f(a) \cap S$ is the set of points p_i for which $g_i(a) \geq 0$. Hence, the problem reduces to point location in the arrangement of algebraic surfaces $g_i = 0$ in $\mathbb{R}^b$. Let G be the set of resulting surfaces. Using a result by Koltun [70], a point-location query in an arrangement of n algebraic surfaces in $\mathbb{R}^b$ can be answered in $O(\log n)$ time using $O(n^{2b-4+\varepsilon})$ space.

Agarwal and Matoušek [7] extended Theorem 3.2 to Tarski cells and showed how to construct partition trees using this extension, obtaining a linear-size data structure with query time $O(n^{1-1/\gamma+\varepsilon})$, where $\gamma = 2$ if $d = 2$ and $\gamma = 2d - 3$ if $d \geq 3$.[3] Extending Matoušek's shallow-cutting based data structure for halfspace range-reporting queries, Sharir and Shaul [92] showed that the query time for Γ_f range-reporting query can be improved in some special cases.

A better linear-size data structure has been proposed [8, 82] based on the *polynomial partitioning scheme* introduced by Guth and Katz [63]. For a set $S \subset \mathbb{R}^d$ of n points and a real parameter r, $1 < r \leq n$, an *r-partitioning polynomial* for S is a nonzero d-variate polynomial f such that each connected component of $\mathbb{R}^d \setminus Z(f)$ contains at most n/r points of S, where $Z(f) := \{x \in \mathbb{R}^d \mid f(x) = 0\}$ denotes the zero set of f. The decomposition of $\mathbb{R}^d$ into $Z(f)$ and the connected components of $\mathbb{R}^d \setminus Z(f)$ is called a *polynomial partition* (induced by f). Guth and Katz showed that an r-partitioning polynomial of degree $O(r^{1/d})$ always exists.

[3] The paper by Agarwal and Matoušek [7] predates the result by Koltun [70]. Using Koltun's result, the value of γ can be improved to $\gamma = d$ for $d \leq 3$ and $\gamma = 2d - 4$ for $d > 3$.

Agarwal et al. [8] described a randomized algorithm to compute such a polynomial in expected time $O(nr + r^3)$. Recent results in real algebraic geometry [21] imply that an algebraic variety of dimension k defined by polynomials of constant-bounded degree crosses $O(r^{k/d})$ components of $\mathbb{R}^d \setminus Z(f)$, and that these components can be computed in time $r^{O(1)}$. Therefore, one can recursively construct the data structure for points lying in each component of $\mathbb{R}^d \setminus Z(f)$. The total time spent in recursively searching in the components crossed by a query range will be roughly $n^{1-1/d}$. However, this ignores the points in $S^* = S \cap Z(f)$. Using a scheme based on the so-called cylindrical algebraic decomposition, Agarwal et al. [8] project the points of S^* to $\mathbb{R}^{d-1}$ and recursively construct a $(d-1)$-dimensional data structure to preprocess S^*. A more elegant and simpler method was subsequently proposed by Matoušek and Patáková [82], which basically applies a generalized polynomial-partitioning scheme on S^* and $Z(f)$. Roughly speaking, they choose another parameter r' and partition S^* further by another polynomial g of degree $O(r'^{1/d})$ such that $Z(f, g) = Z(f) \cap Z(g)$ has dimension $d-2$ and each component of $Z(f) \setminus Z(g)$ contains at most n^*/r' points of S^*. If $Z(f, g)$ also contains many points, then they partition $Z(f, g)$ by another polynomial h so that $Z(f, g, h)$ has dimension $d-3$, and so on. The main contribution of their paper is proving the existence of such functions g and h and an algorithm for computing them. Putting everything together, a semialgebraic range-counting query can be answered in $O(n^{1-1/d} \operatorname{polylog}(n))$ time using a linear-size data structure; all k points lying inside the query range can be reported by spending an additional $O(k)$ time.

We conclude this section by noting that Arya and Mount [16] have presented a linear-size data structure for approximate range-searching queries. Let γ be a constant-complexity semialgebraic set and $\varepsilon > 0$ a parameter. Their data structure returns in $O(\frac{1}{\varepsilon^d} \log n + k_\varepsilon)$ time a subset S_ε of k_ε points such that $\gamma \cap S \subseteq S_\varepsilon \subseteq \gamma_\varepsilon \cap S$ where γ_ε is the set of points within distance $\varepsilon \cdot \operatorname{diam}(\gamma)$ of γ. If γ is convex, the query time improves to $O(\log n + \frac{1}{\varepsilon^{d-1}} + k_\varepsilon)$. A result by Larsen and Nguyen [72] implies that query time of a linear-size data structure is $\Omega(\log n + \varepsilon^{-\frac{d}{1+\delta}-1})$ for any arbitrarily small constant $\delta > 0$. The data structure in [16] can also return a value k_ε, with $|S \cap \gamma| \le k_\varepsilon \le |S \cap \gamma_\varepsilon|$ in time $O(\frac{1}{\varepsilon^d} \log n)$, or in $O(\log n + \frac{1}{\varepsilon^{d-1}})$ time if γ is convex. See also [50]. Chazelle et al. [39] studied the approximate halfspace range-counting problem in high dimensions, where d is not a constant, under a similar notion of approximation—the points within distance ε from the boundary hyperplane may be misclassified. They presented a data structure of size $dn^{O(\varepsilon^{-2})}$ that can answer a query in time $O((d/\varepsilon^2) \operatorname{polylog}(d/\varepsilon))$.

6 Variants and Extensions

In this section we review some extensions of range-searching data structures, including multi-level data structures, semialgebraic range searching, and dynamization. As in the previous section, the preprocessing time for each of the data structures we describe is at most a polylogarithmic or n^ε factor larger than its size.

6.1 Multi-level Data Structures

A rather powerful property of data structures based on decomposition schemes (described in Sect. 2) is that they can be cascaded together to answer more complex queries, at the increase of a logarithmic factor in their performance. This property has been implicitly used for a long time; see, for example, [55, 73, 91]. The real power of the cascading property was first observed by Dobkin and Edelsbrunner [52], who used this property to answer several complex geometric queries. Since their result, several papers have exploited and extended this property to solve numerous geometric-searching problems; see [9, 28, 79]. In this subsection we briefly sketch the general cascading scheme, as described in [79].

Let S be a set of weighted objects, let $\mathcal{R}$ be a set of ranges, and let $\Diamond \subseteq S \times \mathcal{R}$ a "spatial" relation between objects and ranges. A geometric-searching problem $\mathcal{P}$, with underlying relation $\Diamond$, requires computing $\sum_{p \Diamond \gamma} w(p)$ for a query range γ. Let $\mathcal{P}^1$ and $\mathcal{P}^2$ be two geometric-searching problems, and let $\Diamond^1$ and $\Diamond^2$ be the corresponding relations. Then we define $\mathcal{P}^1 \circ \mathcal{P}^2$ to be the conjunction of $\mathcal{P}^1$ and $\mathcal{P}^2$, whose relation is $\Diamond^1 \cap \Diamond^2$. That is, for a query range γ, we want to compute $\sum_{p \Diamond^1 \gamma, p \Diamond^2 \gamma} w(p)$. Suppose we have hierarchical decomposition schemes $\mathcal{D}^1$ and $\mathcal{D}^2$ for problems $\mathcal{P}^1$ and $\mathcal{P}^2$. Let $\mathcal{F}^1 = \mathcal{F}^1(S)$ be the set of canonical subsets constructed by $\mathcal{D}^1$, and for a range γ, let $\mathcal{C}^1_\gamma = \mathcal{C}^1(S, \gamma)$ be the corresponding partition of $\{p \in S \mid p \Diamond^1 \gamma\}$ into canonical subsets. For each canonical subset $C \in \mathcal{F}^1$, let $\mathcal{F}^2(C)$ be the collection of canonical subsets of C constructed by $\mathcal{D}^2$, and let $\mathcal{C}^2(C, \gamma)$ be the corresponding partition of $\{p \in C \mid p \Diamond^2 \gamma\}$ into level-two canonical subsets. The decomposition scheme $\mathcal{D}^1 \circ \mathcal{D}^2$ for the problem $\mathcal{P}^1 \circ \mathcal{P}^2$ consists of the canonical subsets $\mathcal{F} = \bigcup_{C \in \mathcal{F}^1} \mathcal{F}^2(C)$. For a query range γ, the query output is $\mathcal{C}_\gamma = \bigcup_{C \in \mathcal{C}^1_\gamma} \mathcal{C}^2(C, \gamma)$. Note that we can cascade any number of decomposition schemes in this manner.

If we view $\mathcal{D}^1$ and $\mathcal{D}^2$ as tree data structures, then cascading the two decomposition schemes can be regarded as constructing a two-level tree, as follows. We first construct the tree induced by $\mathcal{D}^1$ on S. Each node v of $\mathcal{D}^1$ is associated with a canonical subset C_v. We construct a second-level tree $\mathcal{D}^2_v$ on C_v and store $\mathcal{D}^2_v$ at v as its secondary structure. A query is answered by first identifying the nodes that correspond to the canonical subsets $C_v \in \mathcal{C}^1_\gamma$ and then searching the corresponding secondary trees to compute the second-level canonical subsets $\mathcal{C}^2(C_v, \gamma)$.

The $O(\text{polylog}(n))$ query-time data structures for simplex range counting fit in this framework. For example, a data structure for counting the number of points in a wedge (i.e., intersection of two halfspaces) in $\mathbb{R}^d$ can be constructed by cascading two d-dimensional halfspace range-counting data structures, as follows. Let S be a set of n points. We define two binary relations $\Diamond^1$ and $\Diamond^2$, where for any wedge $\gamma = \gamma_1 \cap \gamma_2$, where γ_1, γ_2 are half-spaces, $p \Diamond^i \gamma$ if $p \in \gamma_i$ $(i = 1, 2)$. Let $\mathcal{P}^i$ be the searching problem associated with $\Diamond^i$, and let $\mathcal{D}^i$ be the halfspace range-counting data structure corresponding to $\mathcal{P}^i$. Then the wedge range-counting problem is the same as $\mathcal{P}^1 \circ \mathcal{P}^2$. We can therefore cascade $\mathcal{D}^1$ and $\mathcal{D}^2$, as described above, to answer a wedge range-counting query. Similarly, a data structure for d-dimensional simplex range-counting can be constructed by cascading $d + 1$ halfspace range-counting data structures. The following theorem states a general result for multi-level data structures.

Theorem 6.1 *Let $S, \mathcal{P}^1, \mathcal{P}^2, \mathcal{D}^1, \mathcal{D}^2$ be as defined above, and let r be a constant. Suppose the size and query time of each decomposition scheme are at most $S(n)$ and $Q(n)$, respectively. If $\mathcal{D}^1$ is efficient and r-convergent, then we obtain a hierarchical decomposition scheme $\mathcal{D}$ for $\mathcal{P}^1 \circ \mathcal{P}^2$ whose size and query time are $O(S(n) \log_r n)$ and $O(Q(n) \log_r n)$. If $\mathcal{D}^2$ is also efficient and r-convergent, then $\mathcal{D}$ is also efficient and r-convergent.*

For example, the wedge range-counting data structure described above has $O(\frac{n^d}{\log^{d-1} n})$ space and $O(\log^2 n)$ query time, and a simplex range counting query can be answered in $O(\log^{d+1} n)$ time using $O(n^d)$ space [79].

The real power of multi-level data structures stems from the fact that there are no restrictions on the relations $\Diamond^1$ and $\Diamond^2$. Hence, any query that can be represented as a conjunction of a constant number of "primitive" queries, each of which admits an efficient, r-convergent decomposition scheme, can be answered by cascading individual decomposition schemes. The next two subsections will describe a few multi-level data structures.

6.2 *Intersection Searching*

A general intersection-searching problem can be formulated as follows. Given a set S of objects in $\mathbb{R}^d$, a semigroup $(\mathbf{S}, +)$, and a weight function $w : S \to \mathbf{S}$, preprocess S into a data structure so that for a query object γ, the weighted sum $\sum_{p \cap \gamma \neq \emptyset} w(p)$, where the sum is taken over all objects $p \in S$ that intersect γ, can be computed quickly. Range searching is a special case of intersection searching in which S is a set of points.

Efficient data structures for many intersection-searching problems have been developed by expressing the intersection test as a conjunction of simple primitive tests (in low dimensions) and using a multi-level data structure to perform these tests. For example, a segment γ intersects another segment e if the endpoints of e lie on the opposite sides of the line containing γ and vice-versa. Suppose we want to report those segments of S whose left endpoints lie below the line supporting a query segment (the other case can be handled in a similar manner). We define three searching problems $\mathcal{P}^1$, $\mathcal{P}^2$, and $\mathcal{P}^3$, with relations $\Diamond^1$, $\Diamond^2$, $\Diamond^3$, as follows:

$e \Diamond^1 \gamma$: The left endpoint of e lies below the line supporting γ.

$e \Diamond^2 \gamma$: The right endpoint of e lies above the line supporting γ.

$e \Diamond^3 \gamma$: The line ℓ_e supporting e intersects γ; equivalently, in the dual plane, the point dual to ℓ_e lies in the double wedge dual to e.

Table 1 Asymptotic upper bounds for intersection-counting queries, with polylogarithmic factors omitted. Reporting queries can be answered by paying an additional $O(k)$ cost

d	Objects	Range	Size	Query time	Source
$d = 2$	Disk	Point	m	$(n/\sqrt{m})^{4/3}$	[28]
	Segment	Segment	m	$n/\sqrt{m}$	[28, 79]
	Triangle	Point	m	$n/\sqrt{m}$	[28, 79]
	Circular arc	Segment	m	$n/m^{1/3}$	[12]
$d = 3$	Tetrahedron	Point	m	$n/m^{1/3}$	[79]
	Sphere	Segment	m	$n/m^{1/3}$	[7]
	Triangle	Segment	m	$n/m^{1/4}$	[7]
$d > 3$	Simplex	Point	m	$n/m^{1/d}$	[79]

For $1 \leq i \leq 3$, let $\mathcal{D}^i$ denote a data structure for $\mathcal{P}^i$. Then $\mathcal{D}^1$ (resp. $\mathcal{D}^2$) is a halfplane range-searching structure on the left (resp. right) endpoints of segments in S, and $\mathcal{D}^3$ is (essentially) a triangle range-searching structure for points dual to the lines supporting S. By cascading $\mathcal{D}^1$, $\mathcal{D}^2$, and $\mathcal{D}^3$, we obtain a data structure for segment-intersection queries. Therefore, by Theorem 6.1, a segment-intersection query can be answered in time $O(n^{1/2} \log^2 n)$ using $O(n \log^2 n)$ space, or in $O(\log^3 n)$ time using $O(n^2)$ space. Similarly, the condition for a query point p lying inside a triangle can be expressed as the conjunction of three tests, each testing whether p lies in a halfplane. So point-stabbing queries amid triangles, i.e., counting all input triangles that intersect a query point, can also be answered in $O(\sqrt{n} \log^2 n)$ (resp. $O(\log^3 n)$) time using $O(n \log^2 n)$ (resp. $O(n^2)$) space. Table 1 summarizes a few intersection searching results.

6.3 Ray-Shooting Queries

Preprocess a set S of objects in $\mathbb{R}^d$ into a data structure so that the first object (if any) hit by a query ray can be reported efficiently. Originally motivated by the ray-tracing problem in computer graphics, this problem has found many applications and has been studied extensively in computational geometry.

A general approach to the ray-shooting problem, using segment intersection-detection structures and Megiddo's parametric searching technique [85], was proposed by Agarwal and Matoušek [6]. Suppose we have a segment intersection-detection data structure for S. Let ρ be a query ray. Their algorithm maintains a segment $\overline{ab} \subseteq \rho$ such that the first intersection point of $\overline{ab}$ with S is the same as that of ρ. If a lies on an object of S, it returns a. Otherwise, it picks a point $c \in \overline{ab}$ and determines, using the segment intersection-detection data structure, whether the interior of the segment $\overline{ac}$ intersects any object of S. If the answer is yes, it recursively finds the first intersection point of $\overline{ac}$ with S; otherwise, it recursively finds the first intersection point of $\overline{cb}$ with S. Using parametric

Table 2 Asymptotic upper bounds for ray shooting queries, with polylogarithmic factors omitted

d	Objects	Size	Query time	Source
$d = 2$	Simple polygon	n	$\log n$	[66]
	s disjoint simple polygons	n	$\sqrt{s}$	[10, 66]
	s disjoint convex polygons	$s^2 + n$	$\log n$	[87]
	Segments	m	$n/\sqrt{m}$	[9, 44]
	Circular arcs	m	$n/m^{1/3}$	[12]
	Disjoint arcs	n	$\sqrt{n}$	[12]
$d = 3$	Convex polytope	n	$\log n$	[53]
	s convex polytopes	$s^2 n^{2+\varepsilon}$	$\log^2 n$	[11, 69]
	Halfplanes	m	$n/m^{1/3}$	[6]
	Triangles	m	$n/m^{1/4}$	[7]
	Spheres	m	$n/m^{1/3}$	[7, 92]
$d > 3$	Hyperplanes	m	$n/m^{1/d}$	[6]
	Convex polytope	m	$n/m^{1/\lfloor d/2 \rfloor}$	[6, 84]

searching, the point c at each stage can be chosen in such a way that the algorithm terminates after $O(\log n)$ steps. In some cases, the query time can be improved by a polylogarithmic factor by exploiting the additional structure of input objects, e.g., replacing parametric search with a randomized technique [26, 89].

Another approach for answering ray-shooting queries is the locus approach. A ray in $\mathbb{R}^d$ can be represented as a point in $\mathbb{R}^d \times \mathbb{S}^{d-1}$. Given a set S of objects, we can partition the parametric space $\mathbb{R}^d \times \mathbb{S}^{d-1}$ into cells so that all points within each cell correspond to rays that hit the same object first; this partition is called the *visibility map* of S. Using this approach and some other techniques, Chazelle and Guibas [37] showed that a ray-shooting query in a simple polygon can be answered in $O(\log n)$ time using $O(n)$ space. A simpler data structure was subsequently proposed by Hershberger and Suri [66]. Table 2 gives a summary of known ray-shooting results.

6.4 Nearest-Neighbor Queries

The *nearest-neighbor (NN) query* problem is defined as follows: Preprocess a set S of points in $\mathbb{R}^d$ into a data structure so that a point in S closest to a query point ξ can be reported quickly. This is one of the most widely studied problems not only in computational geometry but in many other fields such as machine learning, computer vision, database systems, information retrieval, and geographic information systems.

For simplicity, we assume that the distance between points is measured in the Euclidean metric, though a more complicated metric can be used depending on the application. For $d = 2$, one can construct the Voronoi diagram of S and answer NN queries by performing point-location queries in the Voronoi diagram. This approach gives an optimal data structure with $O(\log n)$ query time, $O(n)$ size, and $O(n \log n)$ preprocessing [51]. No such data structure is known for even $d = 3$. A NN query for a set of points under the Euclidean metric in $\mathbb{R}^d$ can be formulated as an instance of the ray-shooting problem in a convex polyhedron in $\mathbb{R}^{d+1}$, as follows. We map each point $p = (p_1, \ldots, p_d) \in S$ to a hyperplane $\hat{p}$ in $\mathbb{R}^{d+1}$, which is the graph of the function

$$f_p(x_1, \ldots, x_d) = 2p_1x_1 + \cdots + 2p_dx_d - (p_1^2 + \cdots + p_d^2).$$

Then p is a closest neighbor of a point $\xi = (\xi_1, \ldots, \xi_d)$ if and only if

$$f_p(\xi_1, \ldots, \xi_d) = \max_{q \in S} f_q(\xi_1, \ldots, \xi_d).$$

That is, if and only if f_p is the first hyperplane intersected by the vertical ray $\rho(\xi)$ emanating from the point $(\xi_1, \ldots, \xi_d, 0)$ in the negative x_{d+1}-direction. If we define $P = \bigcap_{p \in S} \{(x_1, \ldots, x_{d+1}) \mid x_{d+1} \geq f_p(x_1, \ldots, x_d)\}$, then p is the nearest neighbor of ξ if and only if the intersection point of $\rho(\xi)$ and ∂P lies on the graph of f_p. Thus a nearest-neighbor query can be answered in time roughly $n/m^{1/\lceil d/2 \rceil}$ using $O(m)$ space [6, 46, 84]. This approach can be extended to answer farthest-neighbor and k-nearest-neighbor queries also. In general, if we have an efficient data structure for answering disk-emptiness queries for disks under a given metric ρ, we can apply parametric searching to answer nearest-neighbor queries under the ρ-metric, provided the data structure satisfies certain mild assumptions [6].

Since answering NN queries is expensive even for moderate values of d, there is extensive work on computing an ε-approximate nearest neighbor (ε-ANN), i.e., returning a point $\tilde{p} \in S$ such that $\|\xi\tilde{p}\| \leq (1 + \varepsilon) \min_{p \in S} \|\xi p\|$, for a given parameter $\varepsilon > 0$; see e.g. [17, 18] and the references therein. The best known data structure can answer an ε-ANN query in $O(\log(n/\varepsilon))$ time using $O(n/\varepsilon^{d/2})$ space. More generally, for a parameter $\log \frac{1}{\varepsilon} \leq \theta \leq \frac{1}{\varepsilon^{d/2} \log(1/\varepsilon)}$, a query can be answered in $O(\log n + \frac{1}{\theta \varepsilon^{d/2}})$ time using $O(n\theta)$ space [18]. The performance of this and many earlier data structures for answering ANN queries depends exponentially on d, so they are not efficient for large values of d. There is much work on data structures for ANN-queries whose query time and size have polynomial dependence on d; see [67] for a survey of higher dimensional NN queries.

6.5 Linear-Programming Queries

Let S be a set of n halfspaces in $\mathbb{R}^d$. We wish to preprocess S into a data structure so that for a direction vector $\vec{v}$, we can determine the first point of $\bigcap_{h \in S} h$ in the direction $\vec{v}$. For $d \leq 3$, such a query can be answered in $O(\log n)$ time using $O(n)$ storage, by constructing the normal diagram of the convex polytope $\bigcap_{h \in S} h$ and preprocessing it for point-location queries. For higher dimensions, Matoušek [78] showed that, using multidimensional parametric searching and a data structure for answering halfspace emptiness queries, a linear-programming query can be answered in $O((n/m^{1/\lfloor d/2 \rfloor})\, \text{polylog}\, n)$ with $O(m)$ storage. Using a randomized LP algorithm by Clarkson [47], Chan [25] described a randomized procedure that reduces LP queries to halfspace-emptiness queries. Using the best-known data structures for halfspace-emptiness queries, there exists a linear-size data structure that can answer LP-queries in expected time $O(n^{1-1/\lfloor d/2 \rfloor})$ for even values of d and in $n^{1-1/\lfloor d/2 \rfloor} 2^{O(\log^* n)}$ time for odd values of d. See [27, 90] for other similar reductions.

7 Concluding Remarks

This chapter reviewed data structures for range searching and a few related problems. The quest for efficient range-searching data structures has resulted in several elegant geometric techniques that have enriched computational geometry as a whole. It would have been impossible to describe all techniques and results on range searching and their applications, so we focused on a few of them. Notwithstanding extensive research in this area, many challenging problems remain open, including the following:

- Although near-optimal bounds are known for simplex range searching in the semigroup model, the known lower bounds in the group model are far from optimal. Also, can a lower bound of $n/m^{1/2}$ be proved for simplex range counting in the real RAM model of computation?
- Can a semialgebraic range query in $\mathbb{R}^d$ be answered in $O(\log n)$ time using $O(n^d)$ space? The simplest question in this direction is whether a disk range-counting query in $\mathbb{R}^2$ be answered in $O(\log n)$ time using $O(n^2)$ space?
- From a practical standpoint, simplex range searching is still largely open; known data structures are rather complicated and do not perform well in practice. Lower bounds suggest that we cannot hope for data structures that do significantly better than the naïve algorithm in the worst case (and for some problems, even in the average case). An interesting open question is to develop simple data structures that work well under some assumptions on input points and query ranges.

Finally, let me conclude this chapter by noting that Jirka Matoušek played a pivotal role not only in shaping the area of range searching but computational geometry more broadly. The impact of his work is profound and will be felt for many years to come.

References

1. P. Afshani, Improved pointer machine and I/O lower bounds for simplex range reporting and related problems. Int. J. Comput. Geom. Appl. **23**, 233–252 (2013)
2. P. Afshani, T.M. Chan, Optimal halfspace range reporting in three dimensions, in *Proceedings of 20th Annual ACM-SIAM Symposium on Discrete Algorithm*, New York, 2009, pp. 180–186
3. P. Afshani, C.H. Hamilton, N. Zeh, A general approach for cache-oblivious range reporting and approximate range counting. Comput. Geom. Theory Appl. **43**, 700–712 (2010)
4. P.K. Agarwal, Range searching, in *Handbook of Discrete and Computational Geometry*, ed. by J.E. Goodman, J. O'Rourke, C. Toth (CRC Press LLC, Boca Raton, 2017, to appear)
5. P.K. Agarwal, J. Erickson, Geometric range searching and its relatives, in *Advances in Discrete and Computational Geometry* ed. By B. Chazelle, J.E. Goodman, R. Pollack. Contemporary Mathematics, vol. 223 (American Mathematical Society, Providence, 1999), pp. 1–56
6. P.K. Agarwal, J. Matoušek, Ray shooting and parametric search. SIAM J. Comput. 22, 794–806 (1993)
7. P.K. Agarwal, J. Matoušek, On range searching with semialgebraic sets. Discret. Comput. Geom. **11**, 393–418 (1994)
8. P.K. Agarwal, J. Matousek, M. Sharir, On range searching with semialgebraic sets. II. SIAM J. Comput. **42**, 2039–2062 (2013)
9. P.K. Agarwal, M. Sharir, Applications of a new space-partitioning technique. Discret. Comput. Geom. **9**, 11–38 (1993)
10. P.K. Agarwal, M. Sharir, Ray shooting amidst convex polygons in 2D. J. Algorithms **21**, 508–519 (1996)
11. P.K. Agarwal, M. Sharir, Ray shooting amidst convex polyhedra and polyhedral terrains in three dimensions. SIAM J. Comput. **25**, 100–116 (1996)
12. P.K. Agarwal, M. van Kreveld, M. Overmars, Intersection queries in curved objects. J. Algorithms **15**, 229–266 (1993)
13. A. Aggarwal, M. Hansen, T. Leighton, Solving query-retrieval problems by compacting Voronoi diagrams, in *Proceedings of 22nd Annual ACM Symposium on Theory computing*, Baltimore, 1990, pp. 331–340
14. B. Aronov, S. Har-Peled, On approximating the depth and related problems. SIAM J. Comput. **38**, 899–921 (2008)
15. B. Aronov, M. Sharir, Approximate halfspace range counting. SIAM J. Comput. **39**, 2704–2725 (2010)
16. S. Arya, D.M. Mount, Approximate range searching. Comput. Geom. Theory Appl. **17**, 135–152 (2000)
17. S. Arya, G.D. da Fonseca, D.M. Mount, Approximate polytope membership queries. CoRR (2016). abs/1604.01183
18. S. Arya, G.D. da Fonseca, D.M. Mount, Optimal approximate polytope membership, in *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms* (2017), pp. 270–288
19. S. Arya, D.M. Mount, J. Xia, Tight lower bounds for halfspace range searching. Discret. Comput. Geom. **47**, 711–730 (2012)
20. D. Avis, Non-partitionable point sets. Inform. Process. Lett. **19**, 125–129 (1984)
21. S. Barone, S. Basu, Refined bounds on the number of connected components of sign conditions on a variety. Discret. Comput. Geom. **47**, 577–597 (2012)
22. J.L. Bentley, Multidimensional binary search trees used for associative searching. Commun. ACM **18**, 509–517 (1975)
23. J.L. Bentley, Multidimensional divide-and-conquer. Commun. ACM **23**, 214–229 (1980)
24. H. Brönnimann, B. Chazelle, J. Pach, How hard is halfspace range searching. Discret. Comput. Geom. **10**, 143–155 (1993)
25. T.M. Chan, Fixed-dimensional linear programming queries made easy, in *Proceedings of 12th Annual ACM Symposium on Computational Geometry*, Philadelphia, 1996, pp. 284–290

26. T.M. Chan, Geometric applications of a randomized optimization technique. Discret. Comput. Geom. **22**, 547–567 (1999)
27. T.M. Chan, Random sampling, halfspace range reporting, and construction of ($\leq$ k)-levels in three dimensions. SIAM J. Comput. **30**, 561–575 (2000)
28. T.M. Chan, Optimal partition trees. Discret. Comput. Geom. **47**, 661–690 (2012)
29. T.M. Chan, K. Tsakalidis, Optimal deterministic algorithms for 2-D and 3-D shallow cuttings, in *Proceedings of 31st International Symposium on Computational Geometry*, Eindhoven, 2015, pp. 719–732
30. B. Chazelle, Filtering search: a new approach to query-answering. SIAM J. Comput. **15**, 703–724 (1986)
31. B. Chazelle, A functional approach to data structures and its use in multidimensional searching. SIAM J. Comput. **17**, 427–462 (1988)
32. B. Chazelle, Lower bounds for orthogonal range searching, II: the arithmetic model. J. ACM **37**, 439–463 (1990)
33. B. Chazelle, Cutting hyperplanes for divide-and-conquer. Discret. Comput. Geom. **9**, 145–158 (1993)
34. B. Chazelle, A spectral approach to lower bounds with applications to geometric searching. SIAM J. Comput. **27**, 545–556 (1998)
35. B. Chazelle, *The Discrepancy Method: Randomness and Complexity* (Cambridge University Press, New York, 2001)
36. B. Chazelle, J. Friedman, A deterministic view of random sampling and its use in geometry. Combinatorica **10**, 229–249 (1990)
37. B. Chazelle, L.J. Guibas, Visibility and intersection problems in plane geometry. Discret. Comput. Geom. **4**, 551–581 (1989)
38. B. Chazelle, L.J. Guibas, D.T. Lee, The power of geometric duality. BIT **25**, 76–90 (1989)
39. B. Chazelle, D. Liu, A. Magen, Approximate range searching in higher dimension. Comput. Geom. Theory Appl. **39**, 24–29 (2008)
40. B. Chazelle, F.P. Preparata, Halfspace range search: an algorithmic application of *k*-sets. Discret. Comput. Geom. **1**, 83–93 (1986)
41. B. Chazelle, B. Rosenberg, Simplex range reporting on a pointer machine. Comput. Geom. Theory Appl. **5**, 237–247 (1996)
42. B. Chazelle, M. Sharir, E. Welzl, Quasi-optimal upper bounds for simplex range searching and new zone theorems. Algorithmica **8**, 407–429 (1992)
43. B. Chazelle, E. Welzl, Quasi-optimal range searching in spaces of finite VC-dimension. Discret. Comput. Geom. **4**, 467–489 (1989)
44. S.W. Cheng, R. Janardan, Algorithms for ray-shooting and intersection searching. J. Algorithms **13**, 670–692 (1992)
45. K.L. Clarkson, New applications of random sampling in computational geometry. Discret. Comput. Geom. **2**, 195–222 (1987)
46. K.L. Clarkson, A randomized algorithm for closest-point queries. SIAM J. Comput. **17**, 830–847 (1988)
47. K.L. Clarkson, Las Vegas algorithms for linear and integer programming. J. ACM **42**, 488–499 (1995)
48. R. Cole, Partitioning point sets in 4 dimensions, in *Proceedings of 12th International Colloquium on Automata, Languages, and Programming*, Nafplio. Lecture Notes Computer Science, vol. 194 (Springer, 1985), pp. 111–119
49. R. Cole, C.K. Yap, Geometric retrieval problems. Inform. Control **63**, 39–57 (1984)
50. G.D. da Fonseca, D.M. Mount, Approximate range searching: the absolute model. Comput. Geom. Theory Appl. **43**, 434–444 (2010)
51. M. de Berg, M. van Kreveld, M. Overmars, O. Schwarzkopf, *Computational Geometry: Algorithms and Applications* (Springer, Berlin, 1997)
52. D.P. Dobkin, H. Edelsbrunner, Space searching for intersecting objects. J. Algorithms **8**, 348–361 (1987)

53. D.P. Dobkin, D.G. Kirkpatrick, A linear algorithm for determining the separation of convex polyhedra. J. Algorithms **6**, 381–392 (1985)
54. H. Edelsbrunner, D.G. Kirkpatrick, H.A. Maurer, Polygonal intersection searching. Inform. Process. Lett. **14**, 74–79 (1982)
55. H. Edelsbrunner, H.A. Maurer, A space-optimal solution of general region location. Theoret. Comput. Sci. **16**, 329–336 (1981)
56. H. Edelsbrunner, E. Welzl, Halfplanar range search in linear space and $O(n^{0.695})$ query time. Inform. Process. Lett. **23**, 289–293 (1986)
57. J. Erickson, New lower bounds for halfspace emptiness, in *Proceedings of 37th Annual IEEE Symposium on Foundations of Computer Science*, Burlington, 1996, pp. 472–481
58. J. Erickson, New lower bounds for Hopcroft's problem. Discret. Comput. Geom. **16**, 389–418 (1996)
59. J. Erickson, Space-time tradeoffs for emptiness queries. SIAM J. Comput. **19**, 1968–1996 (2000)
60. M.L. Fredman, A lower bound on the complexity of orthogonal range queries. J. ACM **28**, 696–705 (1981)
61. M.L. Fredman, Lower bounds on the complexity of some optimal data structures. SIAM J. Comput. **10**, 1–10 (1981)
62. V. Gaede, O. Günther, Multidimensional access methods. ACM Comput. Surv. **30**, 170–231 (1998)
63. L. Guth, N.H. Katz, On the Erdős distinct distances problem in the plane. Annals Math. **181**, 155–190 (2015)
64. S. Har-Peled, M. Sharir, Relative (p, ϵ)-approximations in geometry. Discret. Comput. Geom. **45**, 462–496 (2011)
65. D. Haussler, E. Welzl, Epsilon-nets and simplex range queries. Discret. Comput. Geom. **2**, 127–151 (1987)
66. J. Hershberger, S. Suri, A pedestrian approach to ray shooting: shoot a ray, take a walk. J. Algorithms 18, 403–431 (1995)
67. P. Indyk, Nearest neighbors in high-dimensional spaces, in *Handbook of Discrete and Computational Geometry*, ed. by J.E. Goodman, J.O'Rourke, C. Toth (CRC Press LLC, Boca Raton, 2017), p. to appear
68. H. Kaplan, E. Ramos, M. Sharir, Range minima queries with respect to a random permutation, and approximate range counting. Discret. Comput. Geom. **45**, 3–33 (2011)
69. H. Kaplan, N. Rubin, M. Sharir, Linear data structures for fast ray-shooting amidst convex polyhedra. Algorithmica **55**, 283–310 (2009)
70. V. Koltun, Sharp bounds for vertical decompositions of linear arrangements in four dimensions. Discret. Comput. Geom. **31**, 435–460 (2004)
71. K.G. Larsen, On range searching in the group model and combinatorial discrepancy. SIAM J. Comput. **43**, 673–686 (2014)
72. K.G. Larsen, H.L. Nguyen, Improved range searching lower bounds, in *Proceedings of 28th Annual Symposium Computational Geometry*, Chapel Hill, 2012, pp. 171–178
73. G.S. Lueker, A data structure for orthogonal range queries, in *Proceedings of 19th Annual IEEE Symposium on Foundations of Computer Science*, 1978, Ann Arbor, pp. 28–34
74. J. Matoušek, Construction of ϵ-nets. Discret. Comput. Geom. **5**, 427–448 (1990)
75. J. Matoušek, Cutting hyperplane arrangements. Discret. Comput. Geom. **6**, 385–406 (1991)
76. J. Matoušek, Efficient partition trees. Discret. Comput. Geom. **8**, 315–334 (1992)
77. J. Matoušek, Reporting points in halfspaces. Comput. Geom. Theory Appl. **2**, 169–186 (1992)
78. J. Matoušek, Linear optimization queries. J. Algorithms **14**, 432–448 (1993)
79. J. Matoušek, Range searching with efficient hierarchical cuttings. Discret. Comput. Geom. **10**, 157–182 (1993)
80. J. Matoušek, Geometric range searching. ACM Comput. Surv. **26**, 421–461 (1994)
81. J. Matoušek, *Using the Borsuk-Ulam Theorem* (Springer, Heidelberg, 2003)
82. J. Matousek, Z. Patáková, Multilevel polynomial partitions and simplified range searching. Discret. Comput. Geom. **54**, 22–41 (2015)

83. J. Matousek, O. Schwarzkopf, Linear optimization queries, in *Proceedings of 8th Annual Symposium on Computational Geometry*, Berlin, 1992, pp. 16–25
84. J. Matoušek, O. Schwarzkopf, On ray shooting in convex polytopes. Discret. Comput. Geom. **10**, 215–232 (1993)
85. N. Megiddo, Applying parallel computation algorithms in the design of serial algorithms. J. ACM **30**, 852–865 (1983)
86. J. Nievergelt, P. Widmayer, Spatial data structures: concepts and design choices, in *Handbook of Computational Geometry*, ed. By J.-R. Sack, J. Urrutia (Elsevier Science Publishers B.V. North-Holland, Amsterdam, 2000), pp. 725–764
87. M. Pocchiola, G. Vegter, Pseudo-triangulations: theory and applications, in *Proceedings of 12th Annual ACM Symposium on Computational Geometry*, Philadelphia, 1996, pp. 291–300
88. S. Rahul, Approximate range counting revisited. CoRR, abs/1512.01713 (2015)
89. E.A. Ramos, On range reporting, ray shooting and k-level construction, in *Proceedings of the 15th Annual Symposium on Computational Geometry* (1999), pp. 390–399
90. E.A. Ramos, Linear programming queries revisited, in *Proceedings of 16th Symposium on Computational Geometry*, Hong Kong, 2000, pp. 176–181
91. H.W. Scholten, M.H. Overmars, General methods for adding range restrictions to decomposable searching problems. J. Symb. Comput. **7**, 1–10 (1989)
92. M. Sharir, H. Shaul, Semialgebraic range reporting and emptiness searching with applications. SIAM J. Comput. **40**, 1045–1074 (2011)
93. E. Welzl, Partition trees for triangle counting and other range searching problems, in *Proceedings of 4th Annual ACM Symposium on Computational Geometry*, Urbana-Champaign, 1988, pp. 23–33
94. D.E. Willard, Polygon retrieval. SIAM J. Comput. **11**, 149–165 (1982)
95. F.F. Yao, A 3-space partition and its applications, in *Proceedings of 15th Annual ACM Symposium on Theory of Computing*, Boston, 1983, pp. 258–263
96. A.C. Yao, On the complexity of maintaining partial sums. SIAM J. Comput. **14**, 277–288 (1985)
97. F.F. Yao, D.P. Dobkin, H. Edelsbrunner, M.S. Paterson, Partitioning space for range queries. SIAM J. Comput. **18**, 371–384 (1989)
98. A.C. Yao, F.F. Yao, A general approach to *D*-dimensional geometric queries, in *Proceedings of 17th Annual ACM Symposium on Theory of Computing*, Providence, 1985, pp. 163–168

Fair Representation by Independent Sets

Ron Aharoni, Noga Alon, Eli Berger, Maria Chudnovsky, Dani Kotlar, Martin Loebl, and Ran Ziv

Abstract For a hypergraph H let $\beta(H)$ denote the minimal number of edges from H covering $V(H)$. An edge S of H is said to represent *fairly* (resp. *almost fairly*) a partition $(V_1, V_2, \ldots, V_m)$ of $V(H)$ if $|S \cap V_i| \geqslant \left\lfloor \frac{|V_i|}{\beta(H)} \right\rfloor$ (resp. $|S \cap V_i| \geqslant \left\lfloor \frac{|V_i|}{\beta(H)} \right\rfloor - 1$) for all $i \leqslant m$. In matroids any partition of $V(H)$ can be represented fairly by some

Research supported by BSF grant no. 2006099, an ISF grant, and by the Discount Bank Chair at the Technion.

Research supported by a USA-Israeli BSF grant, by an ISF grant, and by the Israeli I-Core program.

Research supported by BSF grant no. 2006099 and by an ISF grant.

Research partially supported by BSF Grant No. 2020345.

Gratefully acknowledges support by the Czech Science Foundation GACR, contract number P202-13-21988S.

R. Aharoni (✉)
Department of Mathematics, Technion, 32000 Haifa, Israel
e-mail: raharoni@gmail.com

N. Alon
Sackler School of Mathematics and Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel
e-mail: nogaa@tau.ac.il

E. Berger
Department of Mathematics, Haifa University, 31999 Haifa, Israel
e-mail: berger@math.haifa.ac.il

M. Chudnovsky
Department of Mathematics, Princeton University, 08544 Princeton, NJ, USA
e-mail: mchudnov@math.princeton.edu

D. Kotlar • R. Ziv
Department of Computer Science, Tel-Hai College, Upper Galilee, Israel
e-mail: dannykot@telhai.ac.il; ranziv@telhai.ac.il

M. Loebl
Department of Applied Mathematics, Charles University, Malostranské n. 25, 118 00 Praha, Czech Republic
e-mail: loebl@kam.mff.cuni.cz

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_2

independent set. We look for classes of hypergraphs H in which any partition of $V(H)$ can be represented almost fairly by some edge. We show that this is true when H is the set of independent sets in a path, and conjecture that it is true when H is the set of matchings in $K_{n,n}$. We prove that partitions of $E(K_{n,n})$ into three sets can be represented almost fairly. The methods of proofs are topological.

1 Introduction

1.1 Terminology and Main Theme

A hypergraph $\mathcal{C}$ is called a *simplicial complex* (or just a "complex") if it is closed down, namely $e \in \mathcal{C}$ and $f \subseteq e$ imply $f \in \mathcal{C}$. We denote by $V(\mathcal{C})$ the vertex set of $\mathcal{C}$, and by $E(\mathcal{C})$ its edge set. Let $\beta(\mathcal{C})$ be the minimal number of edges ("simplices") of $\mathcal{C}$ whose union is $V(\mathcal{C})$. For any hypergraph H we denote by $\Delta(H)$ the maximal degree of a vertex in H.

We say that $S \in \mathcal{C}$ represents a set A of vertices *fairly* if $|S \cap A| \geqslant \left\lfloor \frac{|A|}{\beta(\mathcal{C})} \right\rfloor$, and that it represents A *almost fairly* if $|S \cap A| \geqslant \left\lfloor \frac{|A|}{\beta(\mathcal{C})} \right\rfloor - 1$. We say that S represents fairly (almost fairly) a collection of sets if it does so to each set in the collection, reminiscent of the way a parliament represents fairly the voters of the different parties.

Clearly, every set A is fairly represented by some edge $S \in \mathcal{C}$. The aim of this paper is to study complexes $\mathcal{C}$ in which for every partition $V_1, \ldots, V_m$ of $V(\mathcal{C})$ there is an edge $S \in \mathcal{C}$ representing all V_i's fairly, or almost fairly.

In matroids, fair representation is always possible. The following can be proved, for example, by the use of Edmonds' matroids intersection theorem.

Theorem 1.1 *If $\mathcal{M}$ is a matroid then for every partition $V_1, \ldots, V_m$ of $V(\mathcal{M})$ there exists a set $S \in \mathcal{M}$ satisfying $|S \cap V_i| \geqslant \left\lfloor \frac{|V_i|}{\beta(\mathcal{C})} \right\rfloor$ for all i.*

Classical examples which do not always admit fair representation are complexes of the form $\mathcal{I}(G)$, the complex of independent sets in a graph G. In this case $\beta(\mathcal{I}(G)) = \chi(G)$, the chromatic number of G, which by Brooks' theorem is at most $\Delta(G) + 1$. Indeed, there are classes of graphs for which the correct proportion of representation is $\frac{1}{\Delta(G)+1}$. In [3] it was proved that if G is chordal and $|V_i| \geqslant \Delta(G)+1$ then there is an independent set representing all sets V_i, from which there follows:

Theorem 1.2 *If G is chordal and $V_1, \ldots, V_m$ is a partition of its vertex set, then there exists an independent set of vertices S such that $|S \cap V_i| \geqslant \lfloor \frac{|V_i|}{\Delta(G)+1} \rfloor$ for all $i \leqslant m$.*

However, in general graphs this is not always true. The following theorem of Haxell [16] pinpoints the correct parameter.

Theorem 1.3 *If $\mathcal{V} = (V_1, V_2, \ldots, V_m)$ is a partition of the vertex set of a graph G, and if $|V_i| \geqslant 2\Delta(G)$ for all $i \leqslant m$, then there exists a set S, independent in G, intersecting all V_i's.*

This was an improvement over earlier results of Alon, who proved the same with $25\Delta(G)$ [8] and then with $2e\Delta(G)$ [9]. The result is sharp, as shown in [17, 22, 23].

Corollary 1.4 *If the vertex set V of a graph G is partitioned into independent sets $V_1, V_2, \ldots, V_m$ then there exists an independent subset S of V, satisfying $|S \cap V_i| \geqslant \left\lfloor \frac{|V_i|}{2\Delta(G)} \right\rfloor$ for every $i \leqslant m$.*

Proof For each $i \leqslant m$ let V_i^j ($j \leqslant \left\lfloor \frac{|V_i|}{2\Delta(G)} \right\rfloor$) be disjoint subsets of size $2\Delta(G)$ of V_i. By Theorem 1.3 there exists an independent set S meeting all V_i^j, and this is the set desired in the theorem. □

1.2 The Special Behavior of Matching Complexes

Matching complexes, namely the independence complexes of line graphs, behave better than independence complexes of general graphs. For example, the following was proved in [1]:

Theorem 1.5 *If G is the line graph of a graph and $V_1, \ldots, V_m$ is a partition of $V(G)$ then there exists an independent set S such that $|S \cap V_i| \geqslant \left\lfloor \frac{|V_i|}{\Delta(G)+2} \right\rfloor$ for every $i \leqslant m$.*

This follows from a bound on the topological connectivity of the independence complexes of line graphs,

$$\eta(\mathcal{I}(G)) \geqslant \frac{|V|}{\Delta(G) + 2}. \tag{1}$$

Here $\eta(\mathcal{C})$ is a connectivity parameter of the complex $\mathcal{C}$ (for the definition see, e.g., [1]). The way from (1) to Theorem 1.5 goes through a topological version of Hall's theorem, proved in [4]. A hypergraph version of (1) was proved in Aharoni, Gorelik, Narins, (Connectivity of the independence complex of line graphs, unpublished). Theorem 1.2 follows from the fact that if G is chordal then $\eta(\mathcal{I}(G)) \geqslant \frac{|V|}{\Delta(G)+1}$.

So, matching complexes are more likely to admit fair representations. We suggest four classes of complexes as candidates for having almost fair representation of disjoint sets.

1. The matching complex of a path.
2. The matching complex of $K_{n,n}$.
3. The matching complex of any bipartite graph.
4. The intersection of two matroids.

Since the third class contains the first two and the fourth contains the third, conjecturing almost fair representation for them goes in ascending order of daring. In fact, we only dare make the conjecture for the first two. As to the fourth, let us just remark that intersections of matroids often behave unexpectedly well with respect to partitions. For example, no instance is known to the authors in which, given two matroids $\mathcal{M}$ and $\mathcal{N}$, there holds $\beta(\mathcal{M} \cap \mathcal{N}) > \max(\beta(\mathcal{M}), \beta(\mathcal{N})) + 1$.

1.3 Independence Complexes of Paths

In Sect. 2 we prove that the independence complex of a path always admits almost fair representation. In fact, possibly more than that is true. Since the matching complex of a path is the independence complex of a path one vertex shorter, a conjecture in this direction (in a slightly stronger form) can be formulated as follows:

Conjecture 1.6 *Given a partition of the vertex set of a path into sets $V_1, \ldots, V_m$ there exists an independent set S and integers b_i, $i \leqslant m$, such that $|S \cap V_i| \geqslant \frac{|V_i|}{2} - b_i$ for all i, and*

1. $\sum_{i \leqslant m} b_i \leqslant \frac{m}{2}$
   and
2. $b_i \leqslant 1$ *for all* $i \leqslant m$.

We prove the existence of sets satisfying either condition of Conjecture 1.6 (but not necessarily both simultaneously).

Theorem 1.7 *Given a partition of the vertex set of a path into sets $V_1, \ldots, V_m$ there exists an independent set S and integers b_i, $i \leqslant m$, such that $\sum_{i \leqslant m} b_i \leqslant \frac{m}{2}$ and $|S \cap V_i| \geqslant \frac{|V_i|}{2} - b_i$ for all i.*

The proof of Theorem 1.7 uses the Borsuk-Ulam theorem.

Theorem 1.8 *Given a partition of the vertex set of a cycle into sets $V_1, \ldots, V_m$ there exists an independent set S such that $|S \cap V_i| \geqslant \frac{|V_i|}{2} - 1$ for all i.*

The proof uses a theorem of Schrijver, strengthening a famous theorem of Lovász on the chromatic number of Kneser graphs. This means that it, too, uses indirectly the Borsuk-Ulam theorem, since the Lovász-Schrijver proof uses the latter. We refer the reader to Matoušek's book [18] for background on topological methods in combinatorics, in particular applications of the Borsuk-Ulam theorem.

1.4 The Matching Complex of $K_{n,n}$

Conjecture 1.9 *For any partition $E_1, E_2, \ldots, E_m$ of $E(K_{n,n})$ and any $j \leqslant m$ there exists a perfect matching F in $K_{n,n}$ satisfying $|F \cap E_i| \geqslant \left\lfloor \frac{|E_i|}{n} \right\rfloor$ for all $i \neq j$, and $|F \cap E_j| \geqslant \left\lfloor \frac{|E_i|}{n} \right\rfloor - 1$.*

We shall prove:

Theorem 1.10 *Conjecture 1.9 is true for $m = 2, 3$.*

For $m = 2$ the result is simple, and the weight of the argument is in a characterization of those cases in which there necessarily exists an index j for which $|F \cap E_j| = \left\lfloor \frac{|E_j|}{n} \right\rfloor - 1$. The proof of the case $m = 3$ is topological, using Sperner's lemma.

1.5 Relationship to Known Conjectures

Conjecture 1.6 is related to a well known conjecture of Ryser on Latin squares. Given an $n \times n$ array A of symbols, a *partial transversal* is a set of entries taken from distinct rows and columns, and containing distinct symbols. A partial transversal of size n is called simply a *transversal*. Ryser's conjecture [19] is that if A is a Latin square, and n is odd, then A necessarily has a transversal. The oddness condition is indeed necessary - for every even $n > 0$ there exist $n \times n$ Latin squares not possessing a transversal. An example is the addition table of $\mathbb{Z}_n$: if a transversal T existed for this Latin square, then the sum of its elements, modulo n, is $\sum_{k \leqslant n} k = \frac{n(n+1)}{2} \pmod{n}$. On the other hand, since every row and every column is represented in this sum, the sum is equal to $\sum_{i \leqslant n} i + \sum_{j \leqslant n} j = n(n+1) \pmod{n}$, and for n even the two results do not agree. Arsovski [12] proved a closely related conjecture, of Snevily, that every square submatrix (whether even or odd) of the addition table of an odd order abelian group possesses a transversal.

Brualdi [13] and Stein [21] conjectured that for any n, any Latin square of order n has a partial transversal of order $n-1$. Stein [21] observed that the same conclusion may follow from weaker conditions – the square does not have to be Latin, and it may suffice that the entries of the $n \times n$ square are equally distributed among n symbols. Re-formulated, this becomes a special case of Conjecture 1.9:

Conjecture 1.11 *If the edge set of $K_{n,n}$ is partitioned into sets $E_1, E_2, \ldots, E_n$ of size n each, then there exists a matching in $K_{n,n}$ consisting of one edge from all but possibly one E_i.*

Here, even for n odd there are examples without a full transversal. In matrix terminology, take a matrix M with $m_{i,j} = i$ for $j < n$, $m_{i,n} = i + 1$ for $i < n$, and $m_{n,n} = 1$.

A related conjecture to Conjecture 1.9 was suggested in [2]:

Conjecture 1.12 *If $E_1, E_2, \ldots, E_m$ are sets of edges in a bipartite graph, and $|E_i| > \Delta(\bigcup_{i \leqslant m} E_i) + 1$ then there exists a rainbow matching.*

Re-phrased, this conjecture reads: If H is a bipartite multigraph, $G = L(H)$ and $V_i \subseteq V(G)$ satisfy $|V_i| \geqslant \Delta(H) + 2$ for all i, then there exists an independent set in G (namely a matching in H) meeting all V_i's.

Remark 1.13

1. We know only one example, taken from [17, 23], in which $|V_i| \geqslant \Delta(H) + 1$ does not suffice. Take three vertex disjoint copies of C_4, say A_1, A_2, A_3. Number the edges of A_i cyclically as a_i^j $(j = 1 \ldots 4)$. Let $E_1 = \{a_1^1, a_1^3, a_3^1\}$, $E_2 = \{a_1^2, a_1^4, a_3^3\}$, $E_3 = \{a_2^1, a_2^3, a_3^2\}$ and $E_4 = \{a_2^2, a_2^4, a_3^4\}$. Then $\Delta(\bigcup_{i \leqslant m} E_i) = 2$, $|E_i| = 3$ and there is no rainbow matching.
2. The conjecture is false if the sets E_i are allowed to be multisets. We omit the example showing this.

An even stronger version of the conjecture is:

Conjecture 1.14 *If the edge set of a graph H is partitioned into sets $E_1, \ldots, E_m$ then there exists a matching M satisfying $|M \cap E_i| \geqslant \left\lfloor \frac{|E_i|}{\Delta(H)+2} \right\rfloor$ for all $i \leqslant m$*

1.6 Over-Representation Vs. Under-Representation and Representing General Systems of Sets

It is easy to find examples falsifying the above conjectures when the sets that are to be fairly represented do not form a partition. Why is that? A possible explanation is that a more natural formulation of our conjectures is not in terms of over-representation, but in terms of under-representation by a large set. Here is a conjecture in this direction:

Conjecture 1.15 *For every m there exists a number $c(m)$ for which the following is true: if G is a bipartite graph and $E_1, \ldots, E_m$ are any sets of edges, then there exists a matching S in G of size at least $\frac{|E(G)|}{\Delta(G)} - c(m)$ such that*

$$|S \cap E_i| \leqslant \left\lceil \frac{|E_i|}{\Delta(G)} \right\rceil \quad \textit{for all} \quad i \leqslant m \tag{2}$$

Possibly $c(m) = \frac{m}{2}$ may suffice. When the E_i's form a partition, condition (2) implies that all but $c(m)$ sets are fairly represented. Of course, a stronger condition is required to imply Conjecture 1.9. The reason that the under-representation formulation is natural is that if the sets E_i form a partition, the condition in (2) defines a generalized partition matroid. The conjecture thus concerns representation by a set belonging to the intersection of three matroids.

2 Fair Representation by Independent Sets in Paths: A Borsuk-Ulam Approach

In this section we prove Theorem 1.7. Following an idea from the proof of the "necklace theorem" [7], we shall use the Borsuk-Ulam theorem. In the necklace problem two thieves want to divide a necklace with m types of beads, each occurring in an even number of beads, so that the beads of every type are evenly split between the two. The theorem is that the thieves can achieve this goal using at most m cuts of the necklace. In our case, we shall employ as "thieves" the sets of odd and even points, respectively, in a sense to be explained below.

We first quote the Borsuk-Ulam theorem. As usual, for $n \geqslant 1$, S^n denotes the set of points $\vec{x} = (x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1}$ satisfying $\sum_{i \leqslant n+1} x_i^2 = 1$.

Theorem 2.1 (Borsuk-Ulam) *For all $n \geqslant 1$, if $f : S^n \to \mathbb{R}^n$ is a continuous odd function, then there exists $\vec{x} \in S^n$ such that $f(\vec{x}) = 0$.*

Proof of Theorem 1.7 Let $v_1, \ldots, v_n$ be the vertices of P_n, ordered along the path. In order to use the Borsuk-Ulam theorem, we first make the problem continuous, by replacing each vertex v_p by the characteristic function of the pth of n intervals of length $\frac{1}{n}$ in $[0, 1]$, open on the left and closed on the right, except for the firs interval which is closed on both sdeis. We call the interval $(\frac{p-1}{n}, \frac{p}{n}]$ ($[0, \frac{1}{n}]$ for $p = 1$) a *bead* and denote it by B_p. Let χ_i be the characteristic function of $\bigcup_{v_p \in V_i} B_p$. Let g be the characteristic function of the union of odd beads on the path, and let $h(y) = 1 - g(y)$.

Given a point $\vec{x} \in S^m$, let $z_k = \sum_{j \leqslant k} x_j^2$, for all $k = 0, \ldots, m+1$ (where $z_0 = 0, z_{m+1} = 1$).

For each $i \leqslant m$ define a function $f_i : S^m \to \mathbb{R}$ by:

$$f_i(x_1, \ldots, x_m) = \sum_{1 \leqslant k \leqslant m} \int_{z_{k-1}}^{z_k} (g(y) - h(y))\chi_i(y)\text{sign}(x_k)dy$$

Here, as usual, $\text{sign}(x) = 0$ if $x = 0$, $\text{sign}(x) = 1$ if $x > 0$, and $\text{sign}(x) = -1$ if $x < 0$. Since the set of points of discontinuity of the sign function is discrete, the functions f_i are continuous. The sign term guarantees that $f_i(-\vec{x}) = -f_i(\vec{x})$. Hence, by the Borsuk-Ulam theorem there exists a point $\vec{w} = (w_1, \ldots, w_{m+1}) \in S^m$ such that $f_i(\vec{w}) = 0$ for all $i \in [m]$, where $z_k = \sum_{j \leqslant k} w_j^2$, for all $k = 0, \ldots, m+1$.

For $y \in [0, 1]$ such that $y \in (z_{k-1}, z_k]$ define $POS(y) = 1$ if $w_k \geqslant 0$ and $POS(y) = 0$ otherwise. Let $NEG(y) = 1 - POS(y)$. Let

$$J_1(y) = POS(y)g(y) + NEG(y)h(y), \quad J_2(y) = POS(y)h(y) + NEG(y)g(y).$$

For fixed $i \in [m]$, the fact that $f_i(\vec{w}) = 0$ means that

$$\int_{y=0}^{1} \chi_i(y)POS(y)[g(y) - h(y)]dy = \int_{y=0}^{1} \chi_i(y)NEG(y)[g(y) - h(y)]dy$$

Shuffling terms this gives:

$$\int_{y=0}^{1} \chi_i(y)[POS(y)g(y) + NEG(y)h(y)]dy = \int_{y=0}^{1} \chi_i(y)[POS(y)h(y) + NEG(y)g(y)]dy \tag{3}$$

Denoting the integral $\int_0^1 u(y)dy$ of a function u by $|u|$, and noting that $J_1(y) + J_2(y) = 1$ for all $y \in [0, 1]$, Equation (3) says that

$$|\chi_i J_1| = |\chi_i J_2| = \frac{|\chi_i|}{2} \tag{4}$$

for every $i \leqslant m$.

A bead contained in an interval $(z_{k-1}, z_k]$ is called *positive* if $w_k \geqslant 0$ and *negative* otherwise. For $k = 1, \ldots, m$ let T_k be the bead containing z_k. The beads that are equal to T_k for some k are called *transition beads*. Let F be the set of transition beads, and let $Z = \bigcup F$. We next remove the transition beads from J_j, by defining:

$$\tilde{J}_j(y) = \min(J_j(y), 1 - \chi_Z(y))$$

Thus $\tilde{J}_1$ is the characteristic function of the union of those beads that are either positive and odd, or negative and even, and $\tilde{J}_2$ is the characteristic function of the union of those beads that are either positive and even, or negative and odd. Let I_j $(j = 1, 2)$ be the set of vertices v_p on whose bead B_p the function $\tilde{J}_j$ is positive. Since the transition beads have been removed, I_1 and I_2 are independent.

For $i \leqslant m$ and $j = 1, 2$ let $c(i,j)$ be the amount of loss of $\tilde{J}_j$ with respect to J_j on beads belonging to V_i, namely beads $B_p \in F$ such that $v_p \in V_i$. Formally,

$$c(i,j) = \sum_{v_p \in V_i, B_p \in F} |\chi_{B_p} \cdot (J_j - \tilde{J}_j)|$$

Then

$$c(i, 1) + c(i, 2) = \frac{1}{n}|\{v_p \in V_i, B_p \in F\}| \tag{5}$$

and $\sum_{i \leqslant m} c(i, 1) + \sum_{i \leqslant m} c(i, 2) = \frac{m}{n}$. Hence for either $j = 1$ or $j = 2$ we have $\sum_{i \leqslant m} c(i,j) \leqslant \frac{m}{2n}$. Let $I = I_j$ for this particular j, and denote $c(i,j)$ by b_i. Then, by (4) and (5) we have $|I \cap V_i| \geqslant \frac{|V_i|}{2} - b_i$, while $\sum_{i \leqslant m} b_i \leqslant \frac{m}{2}$. Namely, the set I satisfies the conditions of the theorem. □

Example 2.2 Let $P = P_4$, the path with 4 vertices $v_i, 1 \leqslant i \leqslant 4$, and let $V_1 = \{v_1, v_2, v_4\}$ and $V_2 = \{v_3\}$. Then one possible set of points given by the Borsuk-Ulam theorem is $z_1 = \frac{1}{8}$, $z_2 = \frac{5}{8}$, and $w_1 > 0$, $w_2 < 0$, $w_3 > 0$ (or with all three signs reversed), as illustrated in Fig. 1. Thus, J_1 is the characteristic function

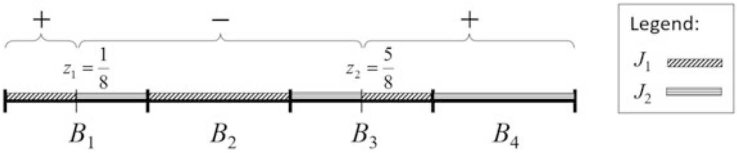


Fig. 1 An example with four vertices divided into two sets

of $[0, \frac{1}{8}] \cup (1, 2] \cup (\frac{5}{8}, \frac{3}{4}]$ and J_2 is the characteristic function of $(\frac{1}{8}, 1] \cup (\frac{1}{2}, \frac{5}{8}] \cup (\frac{3}{4}, 1]$. The set I_1 is obtained from J_1 by removing the z_i-infected beads, namely $I_1 = \{v_2\}$, and then $I_2 = \{v_4\}$. In this case $\sum_{i \leqslant m} c(i, j) = \frac{m}{2n} = \frac{2}{4} = \frac{1}{2}$ for both $j = 1$ and $j = 2$, and thus we can choose I as either I_1 or I_2. This is what the proof gives, but in fact in this example we can do better – we can take $I = \{v_1, v_3\}$, in which only V_1 is under-represented.

Remark 2.3

1. The inequality $\sum_{i \leqslant m} b_i \leqslant \frac{m}{2}$ can possibly be improved, but not much. Namely, there are examples in which the minimum of the sum $\sum_{i \leqslant m} b_i$ in the theorem is $\frac{m-1}{2}$. To see this, let $m = 2k + 1$, and let each V_i be of size $2k$. Consider a sequence of length $2k \times (2k + 1)$, in which the $((i - 1)m + 2j - 1)$-th element belongs to V_i $(i = 1, \ldots, 2k,\ j = 1, \ldots, k + 1)$ and the rest of the elements are chosen in any way so as to satisfy the condition $|V_i| = 2k$. For example, if $k = 2$, then the sequence is of the form:

$$1 * 1 * 1 - 2 * 2 * 2 - 3 * 3 * 3 - 4 * 4 * 4$$

   where the $*$'s can be filled in any way that satisfies $|V_i| = 4$ (namely, four of them are replaced by the symbol 5 and one is replaced by i, for each symbol $i = 1, 2, 3, 4$. The dashes are there to facilitate the reference to the four stretches). If S is an independent set in the path, then we may assume that S contains no more than k elements from the same V_i from each stretch (for example, in the first stretch of the example above choosing all three 1s will result in deficit of 2 in the other sets), Thus $|S| \leqslant 2k \times k$, which is $\frac{m-1}{2}$ short of half the length of the path.
2. It may be of interest to find the best bounds as a function of the sizes of the sets V_i and their number. Note that in the example above the size of the sets is almost equal to their number. As one example, if all V_i's are of size 2, then the inequality can be improved to: $\sum_{i \leqslant m} b_i \leqslant \frac{m}{3}$. To see this, look at the multigraph obtained by adding to P_n the pairs forming the sets V_i as edges. In the resulting graph the maximum degree is 3, and hence by Brooks' theorem it is 3-colorable. Thus there is an independent set of size at least $\frac{n}{3}$, which represents all V_i's apart from at most $\frac{m}{3}$ of them.

3 Fair Representation by Independent Sets in Cycles: Using a Theorem of Schrijver

In this section we shall prove Theorem 1.8. The proof uses a result of Schrijver [20], which is a strengthening of a theorem of Lovász:

Theorem 3.1 (Schrijver [20]) *For integers k, n satisfying $n > 2k$ let $K = K(n, k)$ denote the graph whose vertices are all independent sets of size k in a cycle C of length n, where two such vertices are adjacent iff the corresponding sets are disjoint. Then the chromatic number of K is $n - 2k + 2$.*

The hard part of this inequality is that the chromatic number of K is at least $n - 2k + 2$, which can be formulated as follows:

Theorem 3.2 *The family $\mathcal{I}(n, k)$ of independent sets of size k in the cycle C_n cannot be partitioned into fewer than $n - 2k + 2$ intersecting families.*

We start with a simple case, in which all V_i's but one are odd:

Theorem 3.3 *Let $m, r_1, r_2, \ldots, r_m$ be positive integers, and put $n = \sum_{i=1}^{m}(2r_i + 1) - 1$. Let $G = (V, E)$ be a cycle of length n, and let $V = V_1 \cup V_2 \cup \ldots \cup V_m$ be a partition of its vertex set, where $|V_i| = 2r_i + 1$ for all $1 \leqslant i < m$ and $|V_m| = 2r_m$. Then there is an independent set S of G satisfying $|S| = \sum_{i=1}^{m} r_i$ and $|S \cap V_i| = r_i$ for all $1 \leqslant i \leqslant m$.*

Proof of Theorem 3.3 Put $k = \sum_{i=1}^{m} r_i$ and note that $n - 2k + 2 = m + 1 > m$. Assume, for contradiction, that there is no $S \in \mathcal{I}(n, k)$ satisfying the assertion of the theorem. Then for every $S \in \mathcal{I}(n, k)$ there is at least one index i for which $|S \cap V_i| \geqslant r_i + 1$. Indeed, otherwise $|S \cap V_i| \leqslant r_i$ for all i and hence $|S \cap V_i| = r_i$ for all i, contradicting the assumption. Let $\mathcal{F}_i$ be the family of sets $S \in \mathcal{I}(n, k)$ for which $|S \cap V_i| \geqslant r_i + 1$. Clearly, $\mathcal{F}_i$ is intersecting (in fact, intersecting within V_i), contradicting the conclusion of Theorem 3.2. □

Corollary 3.4 *Let $V = V_1 \cup V_2 \cup \cdots \cup V_m$ be a partition of the vertex set of a cycle C.*

(i) For every i such that $|V_i|$ is even there exists an independent set S_i of C satisfying:

1. $|S_i \cap V_i| = |V_i|/2$.
2. $|S_i \cap V_j| = (|V_j| - 1)/2$ for all j for which $|V_j|$ is odd.
3. $|S \cap V_j| = |V_j|/2 - 1$ for every $j \neq i$ for which $|V_j|$ is even. .

(ii) If $|V_i|$ is odd for all $i \leqslant m$ then for any vertex v of C there is an independent set S of C not containing v and satisfying $|S \cap V_i| = (|V_i| - 1)/2$ for all i.

Proof of Corollary 3.4 Part (i) in case all sets V_j besides V_i are of odd sizes is exactly the assertion of Theorem 3.3. If there are additional indices $j \neq i$ for which $|V_j|$ is even, choose an arbitrary vertex from each of them and contract an edge incident with it. The result follows by applying the theorem to the shorter cycle obtained. Part (ii) is proved in the same way, contracting an edge incident with v. □

4 More Applications of Schrijver's Theorem and Its Extensions

4.1 Hypergraph Versions

The results above can be extended by applying known hypergraph variants of Theorem 3.1. For integers $n \geqslant s \geqslant 2$, let C_n^{s-1} denote the $(s-1)$-th power of a cycle of length n, that is, the graph obtained from a cycle of length n by connecting every two vertices whose distance in the cycle is at most $s-1$. Thus if $s = 2$ this is simply the cycle of length n whereas if $n \leqslant 2s-1$ this is a complete graph on n vertices. For integers n, k, s satisfying $n > ks$, let $K(n, k, s)$ denote the following s-uniform hypergraph. The vertices are all independent sets of size k in C_n^{s-1}, and a collection $V_1, V_2, \ldots, V_s$ of such vertices forms an edge iff the sets V_i are pairwise disjoint. Note that for $s = 2$, $K(n, k, 2)$ is exactly the graph $K(n, k)$ considered in Theorem 3.1. The following conjecture appears in [10].

Conjecture 4.1 *For $n > ks$, the chromatic number of $K(n, k, s)$ is $\lceil \frac{n-ks+s}{s-1} \rceil$.*

This is proved in [10] if s is any power of 2. Using this fact we can prove the following.

Theorem 4.2 *Let $s \geqslant 2$ be a power of 2, let m and $r_1, r_2, \ldots, r_m$ be integers, and put $n = s\sum_{i=1}^m r_i + (s-1)(m-1)$. Let $V_1, V_2, \ldots, V_m$ be a partition of the vertex set of C_n^{s-1}, where $|V_i| = sr_i + s - 1$ for all $1 \leqslant i < m$, and $|V_m| = sr_m$. Then there exists an independent set S in C_n^{s-1} satisfying $|S \cap V_i| = r_i$ for all $1 \leqslant i \leqslant m$.*

Proof Put $k = \sum_{i=1}^m r_i$ and note that the chromatic number of $K(n, k, s)$ is $\lceil (n - ks + s)/(s-1) \rceil > m$. Assume, for contradiction, that there is a partition of the vertex set of C_n^{s-1} with parts V_i as in the theorem, with no independent set of C_n^{s-1} of size $k = \sum_{i=1}^m r_i$ satisfying the assertion of the theorem. In this case, for any such independent set S there is at least one index i so that $|S \cap V_i| \geqslant r_i + 1$. We can thus define a coloring f of the independent sets of size k of C_n^{s-1} by letting $f(S)$ be the smallest i such that $|S \cap V_i| \geqslant r_i + 1$. Since the chromatic number of $K(n, k, s)$ exceeds m, there are s pairwise disjoint sets $S_1, S_2, \ldots, S_s$ and an index i such that $|S_j \cap V_i| \geqslant r_i + 1$ for all $1 \leqslant j \leqslant s$. But this implies that $|V_i| \geqslant sr_i + s$, contradicting the assumption on the size of the set V_i, and completing the proof. □

Just as in the previous section, this implies the following.

Corollary 4.3 *Let $s > 1$ be a power of 2. Let $V_1, V_2, \ldots, V_m$ be a partition of the vertex set of C_n^{s-1}, where $n = \sum_{i=1}^m |V_i|$. Then there is an independent set S in C_n^{s-1}*

satisfying

$$|S \cap V_i| = \left\lfloor \frac{|V_i| - s + 1}{s} \right\rfloor$$

for all $1 \leqslant i < m$*, and*

$$|S \cap V_m| = \left\lfloor \frac{|V_i|}{s} \right\rfloor.$$

The proof is by contracting edges, reducing each set V_i to one of size $s\left\lfloor \frac{|V_i|-s+1}{s} \right\rfloor + s - 1$ for $1 \leqslant i < m$, and reducing V_m to a set of size $s\left\lfloor \frac{|V_m|}{s} \right\rfloor$. The result follows by applying Theorem 4.2 to this contracted graph.

4.2 The Du-Hsu-Wang Conjecture

Du, Hsu and Wang [14] conjectured that if a graph on $3n$ vertices is the edge disjoint union of a Hamilton cycle of length $3n$ and n vertex disjoint triangles then its independence number is n. Erdős conjectured that in fact any such graph is 3-colorable. Using an algebraic approach introduced in [11], Fleischner and Stiebitz [15] proved this conjecture in a stronger form – any such graph is in fact 3-choosable.

The original conjecture, in a slightly stronger form, can be derived from Theorem 3.3: omit any vertex and apply the theorem with $r_i = 1$ for all i. So, for every vertex v there exists a representing set as desired in the conjecture omitting v. The derivation of the statement of Theorem 3.3 from the result of Schrijver in [20] actually supplies a quick proof of the following:

Theorem 4.4 *Let* $C_{3n} = (V, E)$ *be cycle of length* $3n$ *and let* $V = A_1 \cup A_2 \cup \ldots \cup A_n$ *be a partition of its vertex set into n pairwise disjoint sets, each of size* 3*. Then there exist two disjoint independent sets in the cycle, each containing one point from each* A_i*.*

Proof Define a coloring of the independent sets of size n in C_{3n} as follows. If S is such an independent set and there is an index i so that $|S \cap A_i| \geqslant 2$, color S by the smallest such i. Otherwise, color S by the color $n + 1$. By [20] there are two disjoint independent sets S_1, S_2 with the same color. This color cannot be any $i \leqslant n$, since if this is the case then

$$|(S_1 \cup S_2) \cap A_i| = |S_1 \cap A_i| + |S_2 \cap A_i| \geqslant 2 + 2 = 4 > 3 = |A_i|,$$

which is impossible. Thus S_1 and S_2 are both colored $n + 1$, meaning that each of them contains exactly one element of each A_i. □

The Fleischner–Stiebitz theorem implies that the representing set in the DHW conjecture can be required to contain any given vertex. This can also be deduced from the topological version of Hall's Theorem first proved in [4] (for this derivation see e.g. [5]). The latter shows also that the cycle of length $3n$ can be replaced by a union of cycles, totalling $3n$ vertices, none being of length 1 mod 3. Simple examples show that the Fleischner–Stiebitz theorem on 3-colorability does not apply to this setting.

Note that none of the above proofs supplies an efficient algorithm for finding the desired independent set.

5 Fair Representation by Matchings in $K_{n,n}$, the Case of Two Parts

The case $m = 2$ of Conjecture 1.9 is easy. Here is its statement in this case:

Theorem 5.1 *If F is a subset of $E(K_{n,n})$, then there exists a perfect matching N such that $|N \cap F| \geqslant \left\lfloor \frac{|F|}{n} \right\rfloor - 1$ and $|N \setminus F| \geqslant \left\lfloor \frac{|E(G)\setminus F|}{n} \right\rfloor - 1$.*

Partitioning $E(K_{n,n})$ into n perfect matchings shows that there exist two perfect matchings, N_1 and N_2, such that $|N_1 \cap F| \leqslant \frac{|F|}{n} \leqslant |N_2 \cap F|$. The fact that any permutation can be reached from any other by a sequence of transpositions means that it is possible to reach N_2 from N_1 by a sequence of exchanges, replacing at each step two edges of the perfect matching by two other edges. Thus, by a mean value argument, at some matching in the process the condition is satisfied.

The question remains of determining the cases in which the (-1) term is necessary. That this term is sometimes necessary is shown, for example, by the case of $n = 2$ and F being a perfect matching. Another example – $n = 6$ and $F = ([3] \times [3]) \cup (\{4, 5, 6\} \times \{4, 5, 6\})$: it is easy to see that there is no perfect matching containing precisely 3 edges from F, as required in Conjecture 1.9.

The appropriate condition is given by the following concept:

Definition 5.2 A subset F of $E(K_{n,n})$ is said to be *rigid* if there exist subsets K and L of $[n]$ such that $F = K \times L \cup ([n] \setminus K) \times ([n] \setminus L)$.

The rigidity in question is with respect to F-parity of perfect matchings:

Theorem 5.3 ([6]) *A subset F of $E(K_{n,n})$ is rigid if and only if $|P \cap F|$ has the same parity for all perfect matchings P in $K_{n,n}$.*

This characterization shows that when F is rigid, it is not always possible to drop the "minus 1" term in Theorem 5.1. Conversely, if F is not rigid, then the "minus 1" term can indeed be dropped, as indicated by Corollary 5.5 below.

We shall show:

Theorem 5.4 *Let $a < c < b$ be three integers and suppose that $F \subseteq E(K_{n,n})$ is not rigid. If there exists a perfect matching P_a such that $|P_a \cap F| = a$ and a perfect* *matching P_b such that $|P_b \cap F| = b$, then there exists a perfect matching P_c satisfying $|P_c \cap F| = c$.*

It follows from Theorem 5.4 that if a subset F of $E(K_{n,n})$ is not rigid then for every integer c such that $n - \nu(E(K_{n,n}) \setminus F) \leqslant c \leqslant \nu(F)$ there exists a perfect matching N satisfying $|N \cap F| = c$. This implies,

Corollary 5.5 *If a subset F of $E(K_{n,n})$ is not rigid, or if $n \nmid |F|$, then there exists a perfect matching N such that $|N \cap F| \geqslant \left\lfloor \frac{|F|}{n} \right\rfloor$ and $|N \setminus F| \geqslant \left\lfloor \frac{|E(K_{n,n})\setminus F|}{n} \right\rfloor$.*

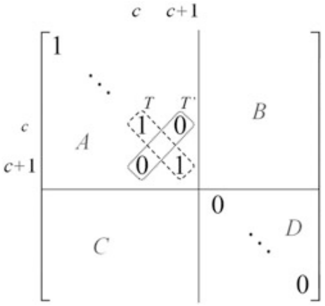
Proof of Theorem 5.4 We use the matrix language of the original Ryser conjecture (Sect. 1.5). Let M be the $n \times n$ matrix in which $m_{i,j} = 1$ if $(i,j) \in F$ and $m_{i,j} = 0$ if $(i,j) \notin F$. A perfect matching in G corresponds to a *generalized diagonal* (abbreviated *GD*) in M, namely a set of n entries belonging to distinct rows and columns. A GD will be called a *k-GD* if exactly k of its entries are 1. By assumption there exist an a-GD T^a and a b-GD T^b. Assume, for contradiction, that there is no c-GD. The case $n = 2$ is trivial, and hence, reversing the roles of 0s and 1s if necessary, we may assume that $c > 1$. Since a GD corresponds to a permutation in S_n, and since every permutation can be obtained from any other permutation by a sequence of transpositions, there exists a sequence of GD's $T^a = T_1, T_2, \ldots, T_k = T^b$, where each pair T_i and T_{i+1}, $i = 1, \ldots, k-1$, differ in two entries. By the contradictory assumption there exists i such that $T := T_{i+1}$ is a $(c+1)$-GD and $T' := T_i$ is a $(c-1)$-GD. Without loss of generality we may assume that T lies along the main diagonal, its first $c+1$ entries are 1, and the rest of its entries are 0.

Let $I = [c+1]$, $J = [n] \setminus I$ and let $A = M[I \mid I]$, $B = M[I \mid J]$, $C = M[J \mid I]$, $D = M[J \mid J]$ (we are using here a common notation – $M[I \mid J]$ denotes the submatrix of M induced by the row set I and column set J). We may assume that the GD T' is obtained from T by replacing the entries (c,c) and $(c+1,c+1)$ by $(c+1,c)$ and $(c,c+1)$ (Fig. 2).

Claim 1 *The matrices A and D are symmetric.*

Proof of Claim 1 To prove that A is symmetric, assume, for contradiction, that there exist $i_1 \neq i_2 \in I$ such that $m_{i_1,i_2} \neq m_{i_2,i_1}$. Then, we can replace the entries (i_1,i_1)

Fig. 2 Obtaining T' from T



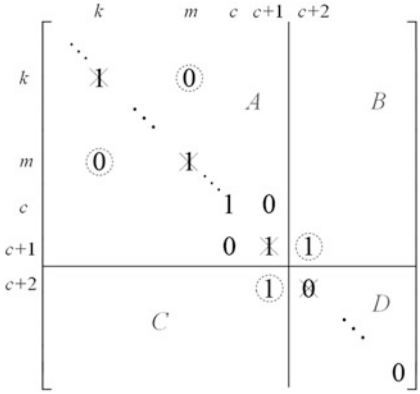


Fig. 3 Subcase II_2. Removed entries are struck out by × and added entries are *circled*

and (i_2, i_2) in T by (i_1, i_2) and (i_2, i_1) to obtain a c-GD. The proof for D is similar, applying the replacement in this case to T'.

Claim 2 *If $i \in I$ and $j \in J$ then $m_{i,j} \neq m_{j,i}$.*

Proof of Claim 2

Case I: $m_{i,j} = m_{j,i} = 0$. Replacing (i, i) and (j, j) in T by (i, j) and (j, i) results in a c-GD.
Case II: $m_{i,j} = m_{j,i} = 1$.
Subcase II_1: $i \not\in \{c, c + 1\}$. Replacing in T' the entries (i, i) and (j, j) by (i, j) and (j, i) results in a c-GD.
Subcase II_2: $i \in \{c, c + 1\}$. Without loss of generality we may assume $i = c + 1$ and $j = c + 2$ (Fig. 3). If $m_{k,\ell} = m_{\ell,k} = 0$ for some $1 \leqslant k < \ell \leqslant c$ then replacing in T the entries (k, k), (ℓ, ℓ), $(c + 1, c + 1)$ and $(c + 2, c + 2)$ by (k, ℓ), (ℓ, k), $(c + 1, c + 2)$ and $(c + 2, c + 1)$ results in a c-GD (Fig. 3). Thus, we may assume that $m_{k,\ell} = m_{\ell,k} = 1$ for all $k, \ell \leqslant c$.

We now consider three sub-subcases:

(i) $m_{c,c+2} = 0, m_{c+2,c} = 1$. In this case we may replace the entries (c, c), $(c + 1, c + 1)$ and $(c + 2, c + 2)$ in T by $(c, c + 2)$, $(c + 1, c)$ and $(c + 2, c + 1)$ and obtain a c-GD (Fig. 4a).

(ii) $m_{c,c+2} = 1, m_{c+2,c} = 0$. Replace the same entries as in Case (i) by $(c, c + 1)$, $(c + 1, c + 2)$ and $(c + 2, c)$, again obtaining a c-GD (Fig. 4b).

(iii) $m_{c,c+2} = m_{c+2,c} = 1$. If $m_{c-1,c+1} = 0$ then, remembering that $m_{c-1,c-1} = 1$, we can replace $(c - 1, c - 1)$, (c, c), $(c + 1, c + 1)$ and $(c + 2, c + 2)$ in T by $(c - 1, c + 1)$, $(c, c + 2)$, $(c + 1, c - 1)$, $(c + 2, c)$ and obtain a c-GD (Fig. 5a). If $m_{c-1,c+1} = 1$, we can replace $(c - 1, c - 1)$, (c, c) and $(c + 1, c + 1)$ in T by $(c - 1, c + 1)$, $(c, c - 1)$ and $(c + 1, c)$ and obtain a c-GD (Fig. 5b.). This proves Claim 2.

For a matrix K indexed by any set of indices X and indices $i, j \in X$, denote by $K_{(i)}$ the row of K indexed by i, and by $K^{(j)}$ the column of K indexed by j.

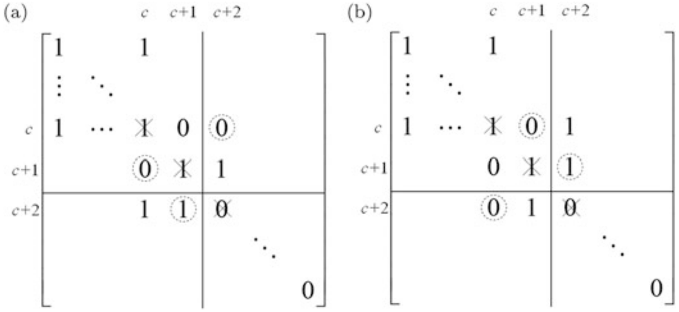


Fig. 4 Subcases II_2(i)–(ii)

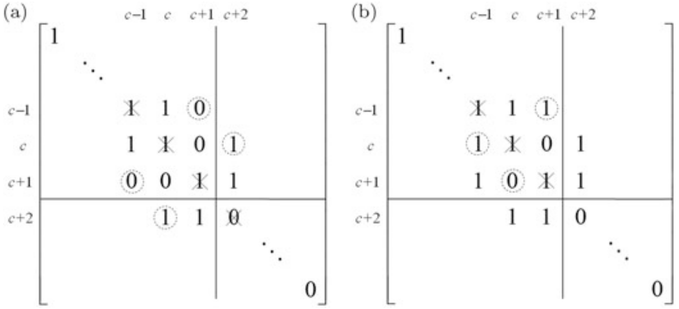


Fig. 5 Subcase II_2(iii)

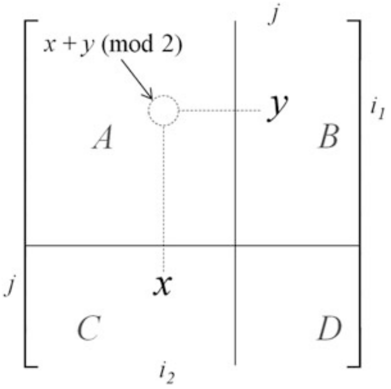


Fig. 6 The submatrix A is the addition (*modulo* 2) table of row $C_{(j)}$ and column $B^{(j)}$

Claim 3 *For any $j \in J$, the submatrix A is the addition table modulo 2 of the row $C_{(j)}$ and the column $B^{(j)}$ (See illustration in Fig. 6).*

Proof of Claim 3 We need to show that for any $i_1, i_2 \in I$ and $j \in J$ we have $m_{i_1,i_2} = m_{j,i_2} + m_{i_1,j} \pmod 2$. We may assume that $i_1 \neq i_2$ since the case $i_1 = i_2$ follows

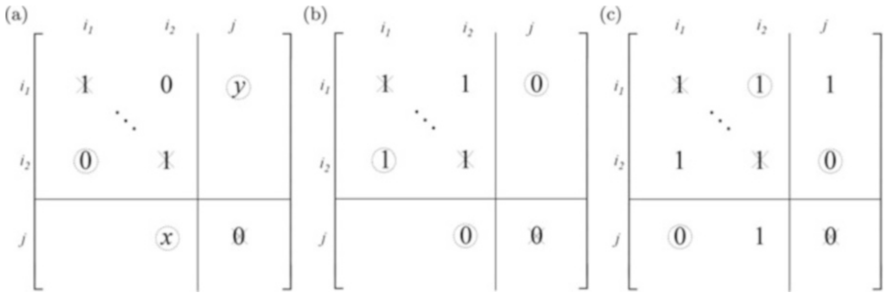


Fig. 7 The three cases in the proof of Claim 3

from Claim 2 and the fact that A has 1's in the main diagonal. Let $x = m_{j,i_2} \in C_{(j)}$ and $y = m_{i_1,j} \in B^{(j)}$. We consider three cases: (i) $x \neq y$, (ii) $x = y = 0$, and (iii) $x = y = 1$.

(i) Assume, for contradiction, that $m_{i_1,i_2} = 0$. Then, by Claim 1, $m_{i_2,i_1} = 0$ and we can replace (i_1, i_1), (i_2, i_2) and (j,j) in T by (i_2, i_1), (i_1,j) and (j, i_2) and obtain a c-GD (Fig. 7a). (ii) Assume, for contradiction, that $m_{i_1,i_2} = 1$. We perform the same exchange as in Case (i) and, again, obtain a c-GD (Fig. 7b). (iii) By Claim 2, we have $m_{i_2,j} = m_{j,i_1} = 0$. Assume, for contradiction, that $m_{i_1,i_2} = 1$. We replace (i_1, i_1), (i_2, i_2) and (j,j) in T by (i_1, i_2), (i_2,j) and (j, i_1) and obtain a c-GD (Fig. 7c). This proves Claim 3.

We say that two (0,1)-vectors u and v of the same length are *complementary* (denoted $u \bowtie v$) if their sum is the vector $(1, 1, \ldots, 1)$. By Claim 3, for every $i_1, i_2 \in I$, if for some $j \in J$, it is true that $m_{i_1,j} = m_{i_2,j}$ then the two rows $A_{(i_1)}, A_{(i_2)}$ are identical, and if $m_{i_1,j} \neq m_{i_2,j}$ then these two rows are complementary. Furthermore - the rows $M_{(i_1)}, M_{(i_2)}$ are identical or complementary. We summarize this in:

Claim 4 *Any two rows in $M[I \mid [n]]$ are either identical or complementary.*

Next we show that the property in Claim 4 holds for any two rows in M.

For $x, y \in \{0, 1\}$ we define the operation $x \circ y = x + y + 1 \pmod 2$ (Fig. 8).

Claim 5 *The submatrix D is the $\circ$-table between the column $C^{(i)}$ and the row $B_{(i)}$, for any $i \in I$.*

Proof of Claim 5 We first consider i such that $1 \leqslant i \leqslant c - 1$ (we assumed $c > 1$). Let $j_1, j_2 \in J$. We may assume that $j_1 \neq j_2$ since the case $j_1 = j_2$ follows from Claim 2 and the fact that D has 0's in the diagonal. Let $x = m_{j_2,i}$ and $y = m_{i,j_1}$. We consider three cases: (i) $x = y = 0$, (ii) $x = y = 1$, and (iii) $x \neq y$.

(i) Assume, for contradiction, that $m_{j_2,j_1} = 0$. By Claim 1, $m_{j_1,j_2} = 0$, and we can replace (i, i), (j_1,j_1) and (j_2,j_2) in T by (i,j_1), (j_1,j_2) and (j_2, i) and obtain a c-GD (Fig. 9a). (ii) By Claim 2, $m_{j_1,i} = m_{i,j_2} = 0$, and we can replace the same entries as in Case 1 by (i,j_2), (j_1, i) and (j_2,j_1) and obtain a c-GD (Fig. 9b). (iii) Here is where we need the assumption $i \leqslant c - 1$. We perform the same replacement as in Case 1, but this time on the GD T', and obtain a c-GD (Fig. 9c. Recall that T' is a $(c - 1)$-GD).

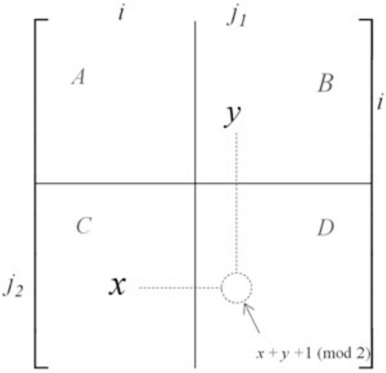


Fig. 8 The submatrix D is the $\circ$-table between column $C^{(i)}$ and row $B_{(i)}$

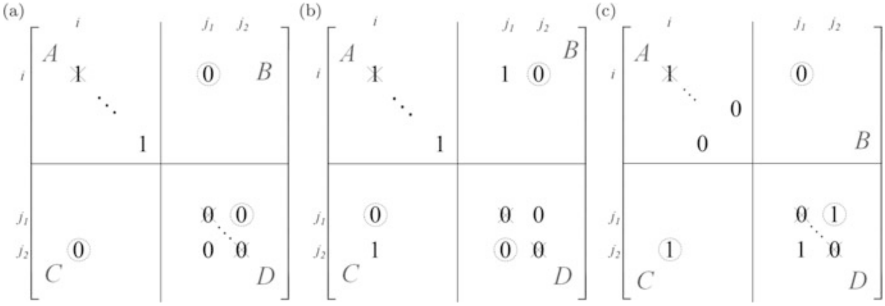


Fig. 9 The three cases in the proof of Claim 5

It remains to prove the claim for $i = c, c+1$. It follows from Claim 4 that any two rows of B are either identical or complementary. Thus, by Claim 2, any two columns of C are either identical or complementary. If there exists $j < c$ such that $B_{(c)} = B_{(j)}$, then $C^{(c)} = C^{(j)}$. Since D is the $\circ$-table between $C^{(j)}$ and $B_{(j)}$, it is also the $\circ$-table between $C^{(c)}$ and $B_{(c)}$. If all $j < c$ satisfy $B_{(c)} \bowtie B_{(j)}$, then for any such j, we have $C^{(c)} = B^T_{(j)}$ and $C^{(j)} = B^T_{(c)}$ by Claim 2. Since $\circ$ is commutative we again have that D is the $\circ$-table between $C^{(c)}$ and $B_{(c)}$. A similar argument holds for $i = c+1$.

Claim 6 *Any two rows of M are either identical or complementary.*

Proof of Claim 6 The fact that any two rows in $M[J|[n]]$ are either identical or complementary follows in the same manner as Claim 4. Now, assume $i \in I, j \in J$. We want to show that $M_{(i)}$ is either identical or complementary to $M_{(j)}$. From Claim 3 we know that $A_{(i)}$ is either identical or complementary to $C_{(j)}$ and from Claim 5 we have that $B_{(i)}$ is either identical or complementary to $D_{(j)}$. We need to show that $A_{(i)}$ is identical to $C_{(j)}$ if and only if $B_{(i)}$ is identical to $D_{(j)}$. Note that

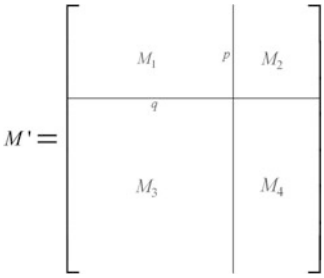


Fig. 10 The four regions in the matrix M'

$m_{ii} = 1, m_{jj} = 0$ and $m_{ij} \neq m_{ji}$. So, if $m_{ji} = 1$ we have identity in both cases and if $m_{ji} = 0$ we have complementarity in both cases.

Suppose all the rows of M are identical. Then, the first $c+1$ columns are all-1 and the rest of the columns are all-0. So, any GD has exactly $c+1$ 1s. So, $a = b = c+1$, which is obviously not the case. Thus, by Claim 6, we can permute the rows and columns to obtain a matrix M' consisting of four submatrices M_1, M_2, M_3 and M_4 of positive dimensions, where M_1 and M_4 are all-1, and M_2 and M_3 are all-0 (Fig. 10).

Thus, F is rigid (Definition 5.2), contrary to the hypothesis. We conclude that there must be a c-GD in M. □

In the case that the partition $E(G) = F \cup (E(G) \setminus F)$ is rigid, if there exists a partition P_{c+1} such that $|P_{c+1} \cap F| = c + 1$, then clearly there is no partition P_c such that $|P_c \cap F| = c$. The proof of Theorem 5.4 shows that in this case, for any c between a and b there is a partition $P_{c'}$ such that $0 \leqslant |P_{c'} \cap F| - c \leqslant 1$.

Corollary 5.6 *Let $G = K_{n,n}$ and assume the partition $E(G) = F \cup (E(G) \setminus F)$ is not rigid. Then, there exist perfect matchings P_1 and P_2 such that $|P_1 \cap F| = \left\lfloor \frac{|F|}{n} \right\rfloor$ and $|P_2 \cap F| = \left\lceil \frac{|F|}{n} \right\rceil$.*

6 Fair Representation by Perfect Matchings in $K_{n,n}$, the Case of Three Parts

In this section we prove Conjecture 1.9 for $m = 3$, namely:

Theorem 6.1 *Suppose that the edges of $K_{n,n}$ are partitioned into sets E_1, E_2, E_3. Then, there exists a perfect matching F in $K_{n,n}$ satisfying $\left\lceil \frac{|E_i|}{n} \right\rceil + 1 \geqslant |F \cap E_i| \geqslant \left\lfloor \frac{|E_i|}{n} \right\rfloor - 1$ for every $i = 1, 2, 3$.*

It clearly suffices to prove the theorem for partitions of $E(K_{n,n})$ into sets E_1, E_2, E_3 such that $|E_i| = k_i n$, for k_i integers $(i = 1, 2, 3)$. Assuming negation of Theorem 6.1 there is no perfect matching with exactly k_i edges from each E_i. As already mentioned, the theorem is patently true if one of the sets E_i is empty, so we may assume $k_1, k_2, k_3 \in \{1, \ldots, n-2\}$.

We identify perfect matchings in $K_{n,n}$ with permutations in S_n. For $\sigma, \tau \in S_n$, the *Hamming distance* (or plainly *distance*) $d(\sigma, \tau)$ between σ and τ is $|\{i \mid \sigma(i) \neq \tau(i)\}|$. We write $\sigma \sim \tau$ if $d(\sigma, \tau) \leqslant 3$. Let $\mathcal{C}$ be the simplicial complex of the cliques of this relation. So, the vertices of $\mathcal{C}$ are the permutations in S_n and the simplexes are the sets of permutations each two of which have distance at most 3 between them. The core of the proof of the theorem will be in showing that $\mathcal{C}$ is simply connected, which will enable us to use Sperner's lemma.

Here is a short outline of the proof of the theorem. Clearly, for each $i \leqslant 3$ there exits a matching F_i representing E_i fairly, namely $|F_i \cap E_i| \geqslant \left\lfloor \frac{|E_i|}{n} \right\rfloor$. We shall connect every pair F_i, F_j $(1 \leqslant i < j \leqslant 3)$ by a path consisting of perfect matchings representing fairly $E_i \cup E_j$, in such a way that every two adjacent matchings are $\sim$-related. This generates a triangle D that is not necessarily simple (namely it may have repeating vertices), together with a triangulation T of its circumference, and an assignment A of matchings to its vertices. We shall then show that there exists a triangulation T' extending T and contained in $\mathcal{C}$ (meaning that there is an assignment A' extending A of perfect matchings to the vertices of T'), such that the perfect matchings assigned to adjacent vertices are $\sim$-related. We color a vertex v of T' by color i if $A'(v)$ represents fairly the set E_i. By our construction, this coloring satisfies the conditions of the 2-dimensional version of Sperner's lemma, and applying the lemma we obtain a multicolored triangle. We shall then show that at least one of the matchings assigned to the vertices of this triangle satisfies the condition required in the theorem.

6.1 Topological Considerations

Let us recall the 2-dimensional version of Sperner's lemma:

Lemma 6.2 *Let T be a triangulation of a triangle ABC and suppose that the vertices of T are colored* 1, 2, 3. *Assume that*

- *The vertex A has color 1.*
- *The vertex B has color 2.*
- *The vertex C has color 3.*
- *Every vertex in the subdivision of the edge AB has either color 1 or color 2.*
- *Every vertex in the subdivision of the edge BC has either color 2 or color 3.*
- *Every vertex in the subdivision of the edge CA has either color 3 or color 1.*

Then T contains a region triangle with three vertices colored 1, 2 and 3.

We shall need a "hexagonal" version of the lemma:

Lemma 6.3 *Let T be a triangulation of a hexagon, whose outer cycle is the union of six paths* $p_1, \dots, p_6$ *(which are, in a cyclic order, subdivisions of the six edges of the hexagon). Suppose that the vertices of T are colored* 1, 2, 3, *in such a way that*

- *No vertex in* p_1 *has color 1.*

- *No edge in p_2 is between two vertices of colors 1 and 2.*
- *No vertex in p_3 has color 2.*
- *No edge in p_4 is between two vertices of colors 2 and 3.*
- *No vertex in p_5 has color 3.*
- *No edge in p_6 is between two vertices of colors 3 and 1.*

Then T contains a region triangle with three vertices colored 1, 2 and 3.

Proof Add three vertices to T outside the circumference of the hexagon in the following way. Add a vertex A of color 1 adjacent to all vertices in p_4, a vertex B of color 2 adjacent to all vertices in p_6 and a vertex of color 3 adjacent to all vertices in p_2. Using Sperner's Lemma on this augmented triangulation yields the lemma. □

Our strategy for the proof of Theorem 6.1 is the following. First we form a triangulation of a hexagon and assign a permutation in S_n to each vertex of the triangulatin, where adjacent permutations are $\sim$ related. Afterwards we color each permutation σ with some color i, where E_i is fairly represented in σ. We then apply Lemma 6.3 to get three permutations $\sigma_1, \sigma_2, \sigma_3$ which are pairwise $\sim$ related, and fairly represent E_1, E_2, E_3 respectively. We then show that how to use this to construct a permutation almost fairly representing all three sets E_1, E_2, E_3, simultaneously.

For $i \in [n]$ let $shift_i : S_n \to S_n$ be a function defined as follows. For every $\sigma \in S_n$, if $\sigma(i) = j$ then

$$shift_i(\sigma)(k) = \begin{cases} i & \text{if } k = i \\ j & \text{if } \sigma(k) = i \\ \sigma(k) & \text{otherwise} \end{cases}$$

Remark 6.4 Note that if $\sigma(i) = i$ then $shift_i(\sigma) = \sigma$.

Lemma 6.5 *If $\sigma \sim \tau$ then $shift_i(\sigma) \sim shift_i(\tau)$.*

Proof Without loss of generality let $i = 1$. If $shift_1(\sigma) = \sigma$ and $shift_1(\tau) = \tau$ then we are done.

Case I: $shift_1(\tau) = \tau$ and $shift_1(\sigma) \neq \sigma$. Without loss of generality $\tau = I$, the identity permutation. For every $k \in [n]$, if $\sigma(k) = k$ then also $shift_1(\sigma)(k) = k$ and thus the distance between $shift_1(\sigma)$ and I is at most the distance between σ and I, yielding $shift_1(\sigma) \sim I = shift_1(\tau)$.

Case II: $shift_1(\sigma) \neq \sigma$ and $shift_1(\tau) \neq \tau$. Without loss of generality $\tau = (12)$ and hence $shift_1(\tau) = I$. As in the previous case, for every $k \in [n]$ if $\sigma(k) = k$ then also $shift_1(\sigma)(k) = k$. We also note that $shift_1(\sigma)(1) = 1$ but $\sigma(1) \neq 1$ (since $shift_1(\sigma) \neq \sigma$). Therefore $d(shift_1(\sigma), I) < d(\sigma, I)$. If $d(\sigma, I) \leqslant 4$ then $shift_1(\sigma) \sim I = shift_1(\tau)$ and we are done. Since $\sigma \sim \tau$, we have $d(\sigma, I) \leqslant 5$ so we may assume that $d(\sigma, I) = 5$. Note that if $\sigma(1) = j \neq 2$, then σ and τ differ on 1,2 and j, and thus $\sigma(k) = k$ for all $k \notin \{1, 2, j\}$, so $d(\sigma, I) \leqslant 3$, contrary to the assumption that this distance is 5. Thus, we must have that $\sigma(1) = 2$. It follows that

$A := \{i \in [n] : \sigma(i) \neq \tau(i)\}$ is a set of size 3 disjoint from $\{1, 2\}$. But then also $\{i \in [n] : shift_1(\sigma(i)) \neq shift_1(\tau(i))\} = A$, yielding $shift_1(\sigma) \sim shift_1(\tau)$. □

At this point we need a connectivity result. This is best formulated in matrix language.

Lemma 6.6 *Let $A = (a_{ij})$ be an $n \times n$ 0-1 matrix and let $k \in [n-1]$. Let G be the graph whose vertices are the permutations $\sigma \in S_n$ satisfying $\sum_{i=1}^{n} a_{i\sigma(i)} \geqslant k$ and whose edges correspond to the $\sim$ relation. If there exists $\rho \in S_n$ with $\sum_{i=1}^{n} a_{i\rho(i)} > k$, then G is connected.*

Proof Without loss of generality $\rho = I$, meaning that $\sum_{i=1}^{n} a_{ii} > k$. We shall show that there is a path in G from ρ to σ for any $\sigma \in V(G) \setminus \{\rho\}$. We prove this claim by induction on $d(\sigma, \rho)$. Write $\ell = \sum_{i=1}^{n} a_{i\sigma(i)}$. Our aim is to find distinct $j \in [n]$ for which $\sigma(j) \neq j$ and $\sigma' = shift_j(\sigma) \in V(G)$. Then the induction hypothesis can be applied since $\sigma \sim \sigma'$ and σ' is closer to ρ than σ.

If $\ell \geqslant k + 2$ choose any $j \in [n]$ with $\sigma(j) \neq j$. Then we have $\sum_{i=1}^{n} a_{i\sigma'(i)} \geqslant \sum_{i=1}^{n} a_{i\sigma(i)} - 2 \geqslant k$, so $\sigma' \in V(G)$.

Suppose next that $\ell = k + 1$. By the assumption that $\sum_{i=1}^{n} a_{ii} > k$ we have $\sum_{i=1}^{n} a_{i\sigma(i)} \leqslant \sum_{i=1}^{n} a_{ii}$ and since $\sigma \neq \rho$ there must be some $j \in [n]$ for which $\sigma(j) \neq j$ and $a_{jj} \geqslant a_{j\sigma(j)}$. Taking $\sigma' = shift_j(\sigma) \in V(G)$ yields $\sum_{i=1}^{n} a_{i\sigma'(i)} \geqslant \sum_{i=1}^{n} a_{i\sigma(i)} - 1 = k$, so $\sigma' \in V(G)$.

Finally, if $\ell = k$ then $\sum_{i=1}^{n} a_{i\sigma(i)} < \sum_{i=1}^{n} a_{ii}$ and hence there must be some $j \in [n]$ for which $a_{jj} > a_{j\sigma(j)}$. Taking $\sigma' = shift_j(\sigma) \in V(G)$ we get $\sum_{i=1}^{n} a_{i\sigma'(i)} \geqslant \sum_{i=1}^{n} a_{i\sigma(i)} + 1 - 1 = k$, so $\sigma' \in V(G)$. □

Corollary 6.7 *Let $A = (a_{ij})$ be an $n \times n$ 0-1 matrix and let $k \in [n]$. Let G be the graph whose vertices are the permutations $\sigma \in S_n$ with $\sum_{i=1}^{n} a_{i\sigma(i)} \geqslant k$ and whose edges correspond to the $\sim$ relation. If $\sum_{i,j \leqslant n} a_{ij} \geqslant kn$ then G is connected.*

Proof If there exists a permutation ρ with $\sum_{i=1}^{n} a_{i\rho(i)} > k$ then we are done by Lemma 6.6. If not, by König's theorem there exist sets $A, B \subseteq [n]$ with $|A| + |B| \leqslant k$ such that $a_{ij} = 0$ for $i \notin A$ and $j \notin B$. This is compatible with the condition $\sum_{i,j \leqslant n} a_{ij} \geqslant kn$ only if $|A| = 0$ and $|B| = k$ or $|B| = 0$ and $|A| = k$, and $a_{ij} = 1$ for all $(i, j) \in A \times [n] \cup [n] \times B$. In both cases $V(G) = S_n$, implying that the relation $\sim$ is path connected since every permutation is reachable from every other permutation by a sequence of transpositions. □

In the next two lemmas let $i \in [n]$ and $\sigma,\ \tau \in S_n$. We write *shift* for $shift_i$.

Lemma 6.8 *If $d(\sigma, \tau) = 2$, then the 4-cycle $\sigma - \tau - shift(\tau) - shift(\sigma) - \sigma$ is null-homotopic in $\mathcal{C}$ (i.e., it can be triangulated.)*

Proof If either $\sigma \sim shift(\tau)$ or $\tau \sim shift(\sigma)$ then we are done. So, we may assume this does not happen and in particular $\sigma \neq shift(\sigma)$ and $\tau \neq shift(\tau)$. We may assume, without loss of generality, that $i = 1, \sigma = (12), \tau = (12)(34), shift(\sigma) = I$ and $shift(\tau) = (34)$. We can now fill the cycle as in Fig. 11. □

Lemma 6.9 *If $d(\sigma, \tau) = 3$ then the 4-cycle $\sigma - \tau - shift(\tau) - shift(\sigma) - \sigma$ is a null cycle in $\mathcal{C}$.*

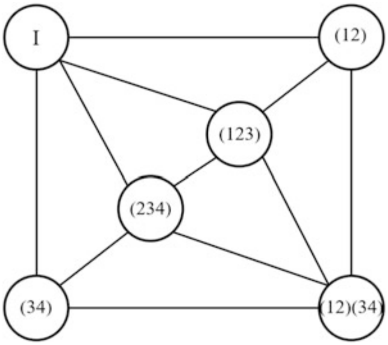


Fig. 11 *Triangulation* of the 4-cycle $\sigma - \tau - \mathit{shift}(\tau) - \mathit{shift}(\sigma) - \sigma$ when $d(\sigma, \tau) = 2$

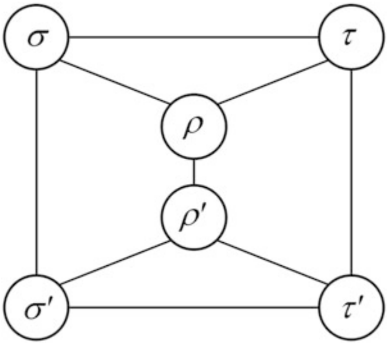


Fig. 12 *Triangulation* of the 4-cycle $\sigma - \tau - \mathit{shift}(\tau) - \mathit{shift}(\sigma) - \sigma$ when $d(\sigma, \tau) = 3$

Proof Let $\rho \in S_n$ have distance 2 from both σ and τ. Denote $\sigma' = \mathit{shift}(\sigma)$, $\tau' = \mathit{shift}(\tau)$ and $\rho' = \mathit{shift}(\rho)$. We use the previous lemma to fill the cycle as in Fig. 12.□

As a corollary from the above two lemmas we get

Corollary 6.10 *Let C be a cycle and let $f : C \to \mathcal{C}$ be a simplicial map, i.e., mapping each edge to an edge or a vertex. Let $\bar{f} : C \to \mathcal{C}$ be defined by $\bar{f}(v) = \mathit{shift}_i(f(v))$ for every $v \in V(C)$. Then $\bar{f}$ is also simplicial and is homotopic to f. (See Fig. 13. As above, $\mathit{shift}(\sigma)$ is denoted by σ'.)*

Lemma 6.11 *The simplicial complex $\mathcal{C}$ is simply connected.*

Proof Let C be a cycle and let $f_0 : C \to \mathcal{C}$ be a simplicial map. We need to show that f_0 is null-homotopic. For each $i \in [n]$, we define $f_i : C \to \mathcal{C}$ by $f_i(v) = \mathit{shift}_i(f_{i-1}(v))$ for every $v \in V(C)$. Then by Corollary 6.10 $f_0, \ldots, f_n$ are all homotopic to each other. But $f_n(v) = I$ for every $v \in V(C)$. This means that $f_0, \ldots, f_n$ are all null-homotopic. □

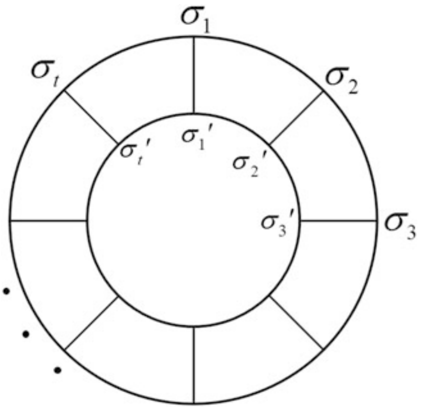


Fig. 13 Obtaining $\bar{f}$ from f in Corollary 6.10

6.2 Associating a Complex with the Graph

Lemma 6.12 *Let the set E of edges of $K_{n,n}$ be partitioned to three sets $E = E_1 \dot{\cup} E_2 \dot{\cup} E_3$. Then there exists a perfect matching M with at least $\left\lceil \frac{|E_1|}{n} \right\rceil$ edges of E_1 and at most $\left\lceil \frac{|E_3|}{n} \right\rceil$ edges of E_3.*

Proof Let H be the graph with the edge set $E_1 \cup E_2$. König's edge coloring theorem, combined with an easy alternating paths argument, yields that H can be edge colored with n colors in a way that each color class is of size either $\left\lfloor \frac{|E(H)|}{n} \right\rfloor$ or $\left\lceil \frac{|E(H)|}{n} \right\rceil$. Clearly, at least one of these classes contains at least $\frac{|E_1|}{n}$ edges from E_1. A matching with the desired property can be obtained by completing this color class in any way we please to a perfect matching of $K_{n,n}$. □

In fact, a stronger property may hold:

Conjecture 6.13 *Let $G = (V, E)$ be a bipartite graph with maximal degree Δ and let $f : E \to \{1, 2, 3, \dots, k\}$ for some positive integer k. Then there exists a matching M in G such that every number $j \in \{1, 2, 3, \dots, k\}$ satisfies*

$$|\{e \in M \mid f(e) \leqslant j\}| \geqslant \left\lfloor \frac{|\{e \in E : f(e) \leqslant j\}|}{\Delta} \right\rfloor$$

Clearly, we only need to see to it that the condition holds for $j < k$.

In Berger, et al. (Fair representation, unpublished) this conjecture was proved for $G = K_{6,6}$.

We shall say that a perfect matching F has property $i^{(+)}$ if $|F \cap E_i| \geqslant k_i$, property $i^{(++)}$ if $|F \cap E_i| > k_i$, and property $i^{(-)}$ if $|F \cap E_i| \leqslant k_i$.

Lemma 6.14 *There exists a triangulation of the boundary of a hexagon, and an assignment of a perfect matching M_v and a color $i_v \in \{1, 2, 3\}$ to each vertex v*

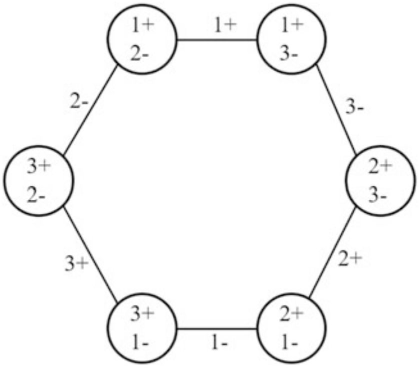


Fig. 14 Assigning perfect matchings to the vertices of the *hexagon* in Lemma 6.14

of the triangulation, such that M_v has property $i_v^{(++)}$ and the coloring satisfies the conditions of Lemma 6.3.

Proof By Lemma 6.12 there exists a perfect matching M with properties $1^{(+)}$ and $3^{(-)}$. We assign it to one vertex of the hexagon. By permuting the roles of E_1, E_2, E_3 we can find six such perfect matchings and assign them to the six vertices of the hexagon as in Fig. 14.

By Corollary 6.7, we can fill the path between the two permutations with property $i^{(-)}$ in a way that all perfect matchings in the path have property $i^{(-)}$. Similarly, we can fill the path between the two permutations with property $i^{(+)}$. For each vertex v we assign a color i_v such that M_v has property $i_v^{(++)}$. If Lemma 6.3 does not hold, then without loss of generality we have two perfect matchings $M_1 \sim M_2$, where M_1 has properties $3^{(+)}$ and $1^{(++)}$ and M_2 has properties $3^{(+)}$ and $2^{(++)}$. This yields Lemma 6.14. □

Since $\mathcal{C}$ is simply connected, we can extend the mapping we got in Lemma 6.14 to a triangulation of the hexagon. Applying Lemma 6.3 we obtain a triangle in the triangulation whose vertices are colored 1, 2 and 3. This means that there exist $\sigma_1, \sigma_2, \sigma_3 \in S_n$, pairwise $\sim$ related and fairly representing E_1, E_2, E_3 respectively.

6.3 Proof of Theorem 6.1

We form a matrix $A = (a_{ij})_{i,j \leqslant n}$, where $a_{ij} = p \ \ (p = 1, 2, 3)$ if the edge ij belongs to E_p.

For each $\ell \in \{1, 2, 3\}$ and $\sigma \in S_n$ we write $d_\ell(\sigma) = |\{i : a_{i\sigma(i)} = \ell\}| - k_\ell$.

Lemma 6.15 *Suppose that the triple $\{\sigma_1, \sigma_2, \sigma_3\}$ is in $\mathcal{C}$, and that $d_\ell(\sigma_\ell) > 0$ for each $\ell \in \{1, 2, 3\}$. Then there exists $\sigma \in S_n$ with $|d_\ell(\sigma)| \leqslant 1$ for each $\ell \in \{1, 2, 3\}$.*

Since the existence of such $\sigma_1, \sigma_2, \sigma_3$ follows from Lemmas 6.3, 6.14 and 6.11, this will finish the proof of Theorem 6.1.

Proof Define a 3×3 matrix $B = (b_{ij})$ by $b_{ij} = d_i(\sigma_j)$. We know that the diagonal entries in B are positive, the sum in each column is zero, and any two entries in the same row differ by at most 3. This means that the minimal possible entry in B is -2. We may assume each column has some entry not in $\{-1, 0, 1\}$.

Let us start with the case that all of the diagonal entries of B are at least 2. This implies that all off-diagonal entries are at least -1. Since each column must sum up to zero, we must have

$$B = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$$

This implie that the distance between any two of $\sigma_1, \sigma_2, \sigma_3$ is exactly 3, and without loss of generality $\sigma_1 = I$, $\sigma_2 = (123)$, $\sigma_3 = (132)$, and the matrix A has the form

$$A = \begin{pmatrix} 1 & 2 & 3 & * & \dots & * \\ 3 & 1 & 2 & * & \dots & * \\ 2 & 3 & 1 & * & \dots & * \\ * & * & * & * & \dots & * \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ * & * & * & * & \dots & * \end{pmatrix}$$

We can now take $\sigma = (12)$ and we are done.

We are left with the case that some diagonal entry of B is 1. Without loss of generality $b_{11} = 1$. We also assume without loss of generality that $b_{21} \leqslant b_{31}$. Since the first column must sum up to zero, we have $b_{21} + b_{31} = -1$, and thus $-0.5 = 0.5(b_{21} + b_{31}) \leqslant b_{31} = -1 - b_{21} \leqslant 1$. In other words, either $b_{21} = -1$ and $b_{31} = 0$ or $b_{21} = -2$ and $b_{31} = 1$. In the first case we can just take $\sigma = \sigma_1$ and we are done. Therefore we assume the second case.

$$B = \begin{pmatrix} 1 & * & * \\ -2 & * & * \\ 1 & * & * \end{pmatrix}$$

Since $d_3(\sigma_1) > 0$, we may assume $\sigma_3 = \sigma_1$, and due to the -2 entries in the second row, we must have $b_{22} = 1$. We now get

$$B = \begin{pmatrix} 1 & * & 1 \\ -2 & 1 & -2 \\ 1 & * & 1 \end{pmatrix}$$

Without loss of generality $b_{12} \leqslant b_{32}$ and by arguments similar to the above we can fill the second column

$$B = \begin{pmatrix} 1 & -2 & 1 \\ -2 & 1 & -2 \\ 1 & 1 & 1 \end{pmatrix}$$

The distance between σ_1 and σ_2 is exactly 3, so without loss of generality $\sigma_1 = I$ and $\sigma_2 = (123)$. In order to achieve the values of $b_{12} = -2, b_{11} = 1, b_{21} = -2, b_{22} = 1$ we must have $a_{ii} = 1$ and $a_{i\sigma_2(i)} = 2$ for each $i \in \{1, 2, 3\}$.

The only case in which none of the choices $\sigma = (12)$ or $\sigma = (23)$ or $\sigma = (13)$ works is if $a_{13} = a_{21} = a_{32} = 3$, so once again we get

$$A = \begin{pmatrix} 1 & 2 & 3 & * & \dots & * \\ 3 & 1 & 2 & * & \dots & * \\ 2 & 3 & 1 & * & \dots & * \\ * & * & * & * & \dots & * \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ * & * & * & * & \dots & * \end{pmatrix}$$

We have $b_{31} = 1$ which means that 3 appears $k_3 + 1$ times on the diagonal. Without loss of generality $a_{44} = a_{55} = \ldots = a_{k_3+4 k_3+4} = 3$. In any of the following cases one can easily find some $\sigma \in S_n$ with $|d_\ell(\sigma)| \leqslant 1$ for each $\ell \in \{1, 2, 3\}$:

- If either $a_{ij} \neq 3$ or $a_{ji} \neq 3$ for some $i \in \{4, \ldots, k_3 + 4\}$ and $j \in \{1, 2, 3\}$.
- If $a_{ij} \neq 3$ for some $i, j \in \{4, \ldots, k_3 + 4\}$
- If both $a_{ij} \neq 3$ and $a_{ji} \neq 3$ for some $i \in \{4, \ldots, k_3 + 4\}$ and $j \in \{k_3 + 5, \ldots, n\}$.

If none of the above occurs then

$$k_3 n = |\{(i,j) : a_{ij} = 3\}| \geqslant 2 \cdot 3 \cdot (1 + k_3) + (1 + k_3)^2 + \frac{1}{2} \cdot 2(k_3 + 1)(n - k_3 - 4)$$

which is a contradiction. □

Remark 6.16 After the above topological proof of Theorem 6.1 was found, a combinatorial proof was given in Berger, et al. (Fair representation, unpublished).

Acknowledgements The authors are grateful to Frédéric Meunier for pointing out an inaccuracy in a previous version of the paper.

References

1. R. Aharoni, N. Alon, E. Berger, Eigenvalues of $K_{1,k}$-free graphs and the connectivity of their independence complexes. J. Graph Theory **83**(4), 384–391 (2016)
2. R. Aharoni, J. Barat, I. Wanless, Multipartite hypergraphs achieving equality in Ryser's conjecture. Graphs Comb. **32**, 1–15 (2016)
3. R. Aharoni, E. Berger, R. Ziv, Independent systems of representatives in weighted graphs. Combinatorica **27**, 253–267 (2007)
4. R. Aharoni, P. Haxell, Hall's theorem for hypergraphs. J. Graph Theory **35**, 83–88 (2000)
5. R. Aharoni, R. Holzman, D. Howard, P. Sprüssel, Cooperative colorings and independent systems of representatives. Electron. J. Comb. **22** (2015)
6. R. Aharoni, R. Manber, B. Wajnryb, Special parity of perfect matchings in bipartite graphs. Discret. Math. **79**, 221–228 (1989/1990)
7. N. Alon, Splitting necklaces. Adv. Math. **63**, 247–253 (1987)
8. N. Alon, The linear arboricity of graphs. Isr. J. Math. **62**, 311–325 (1988)
9. N. Alon, Probabilistic methods in coloring and decomposition problems. Discret. Math. **127**, 31–46 (1994)
10. N. Alon, L. Drewnowski, T. Łuczak, Stable Kneser hypergraphs and ideals in N with the Nikodym property. Proc. Am. Math. Soc. **137**, 467–471 (2009)
11. N. Alon, M. Tarsi, Chromatic numbers and orientations of graphs. Combinatorica **12**, 125–134 (1992)
12. B. Arsovski, A proof of Snevily's conjecture. Isr. J. Math. **182**, 505–508 (2011)
13. R.A. Brualdi, H.J. Ryser, *Combinatorial Matrix Theory* (Cambridge University Press, Cambridge, 1991)
14. D.Z. Du, D.F. Hsu, F.K. Hwang, The Hamiltonian property of consecutive-d digraphs, in Graph-theoretic models in computer science, II (Las Cruces, NM, 1988. 1990). Math. Comput. Model. **17**, 61–63 (1993)
15. H. Fleischner, M. Stiebitz, A solution to a colouring problem of P. Erdös. Discret. Math. **101**, 39–48 (1992)
16. P.E. Haxell, A condition for matchability in hypergraphs. Graphs Comb. **11**, 245–248 (1995)
17. G. Jin, Complete subgraphs of r-partite graphs. Comb. Probab. Comput. **1**, 241–250 (1992)
18. J. Matoušek, Using the Borsuk—Ulam theorem. Lectures on topological methods in *combinatorics and geometry* (Springer, New York, 2003)
19. H.J. Ryser, *Neuere probleme der kombinatorik*, Vorträge über Kombinatorik, Oberwolfach, *Matematisches Forschungsinstitute* (Oberwolfach, Germany), July 1967, pp. 69–91
20. A. Schrijver, Vertex-critical subgraphs of Kneser graphs. Nieuw Arch. Wisk. **26**, 454–461 (1978)
21. S.K. Stein, Transversals of Latin squares and their generalizations. Pac. J. Math. **59**, 567–575 (1975)
22. T. Szábo, G. Tardos, Extremal problems for transversals in graphs with bounded degrees. Combinatorica **26**, 333–351 (2006)
23. R. Yuster, Independent transversals in *r*-partite graphs. Discret. Math. **176**, 255–261 (1997)

Computing Heegaard Genus is NP-Hard

David Bachman, Ryan Derby-Talbot, and Eric Sedgwick

Dedicated to the memory of Jiří Matoušek

Abstract We show that HEEGAARD GENUS $\leq g$, the problem of deciding whether a triangulated 3-manifold admits a Heegaard splitting of genus less than or equal to g, is NP-hard. The result follows from a quadratic time reduction of the NP-complete problem CNF-SAT to HEEGAARD GENUS $\leq g$.

1 Introduction

While there is a tradition of studying decision problems in 3-manifold topology, the historical focus has been showing that problems are decidable [9, 13–15, 20, 21, 31, 37]. More recently, the computational complexity of these and related problems has gained attention [1, 5–7, 10, 18, 34]. Here we show that one of the most basic decision problems for 3-manifolds, the problem of determining Heegaard genus, is NP-hard.

Every closed, orientable 3-manifold M has a *Heegaard surface*: a closed surface that splits the manifold into a pair of handlebodies (i.e., thickened graphs). The

The authors would like to thank Martin Tancer, whose guidance was instrumental at the genesis of this project, and Iyad Kanj and Marcus Schaefer for helpful conversations along the way.

D. Bachman
Pitzer College, Claremont, CA, USA
e-mail: bachman@pitzer.edu

R. Derby-Talbot
Quest University, Squamish, BC, Canada
e-mail: rdt@questu.ca

E. Sedgwick (✉)
School of Computing, DePaul University, 243 S. Wabash Ave, 60604 Chicago, IL, USA
e-mail: esedgwick@cdm.depaul.edu

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_3

Heegaard genus, $g(M)$, is the minimal genus of a Heegaard surface for M, and is one of the most basic 3-manifold invariants. Because Heegaard surfaces are generic, they have been studied extensively and have been effectively classified for large classes of manifolds [16, 23]. It is thus natural to ask (phrased as a decision problem):

Problem 1.1 HEEGAARD GENUS $\leq g$: *Given a triangulated 3-manifold M and a natural number g, does M have a Heegaard surface of genus $\leq g$?*

HEEGAARD GENUS $\leq g$ was shown to be decidable (computable) by Johannson [14, 15] in the Haken case and by Li in the non-Haken case [20]. Our main result is the following:

Theorem 1.2 HEEGAARD GENUS $\leq g$ *is* NP-*hard.*

One way of obtaining a Heegaard surface in certain 3-manifolds is to *amalgamate* Heegaard surfaces in submanifolds. This approach allows us to relate Heegaard genus to satisfiability of Boolean formulas in *conjunctive normal form*, that is Boolean formulas stated as a conjunction of disjunctions, for example:

$$Q = (a \vee c) \wedge (\neg a \vee b) \wedge (b \vee c)$$

We will let $|Q|$ denote the length of Q without counting parentheses, e.g. $|Q| = 12$ for the above example.

Problem 1.3 CNF-SAT: *Given Q, a Boolean formula in conjunctive normal form, is there a satisfying assignment (i.e., an assignment of truth values to the variables) that makes the formula true?*

CNF-SAT is well known to be NP-complete. We prove Theorem 1.2 by giving a polynomial (quadratic) time reduction of CNF-SAT to HEEGAARD GENUS $\leq g$. Our reduction will proceed in two steps, first proving that there are manifolds M_Q that encode a formula Q:

Proposition 3.1 *Let Q be an instance of* CNF-SAT. *Then there is a manifold M_Q with Heegaard genus $g(M_Q) \geq |Q| + 2$, with equality holding if and only if Q has a satisfying assignment.*

The proof of Proposition 3.1 is based on constructing M_Q as a direct translation of the formula Q (a schematic of M_Q for the aforementioned Q is shown in Fig. 1), formed by taking a collection of Heegaard genus two "block" manifolds, one block for each term (VAR(iable), REP(licate), NOT, AND, OR) in Q, and gluing them together along torus boundary components via high distance maps. Each gluing surface then represents a sub-statement of Q. The high-distance gluings guarantee that any minimal genus Heegaard surface for M_Q is an amalgamation of Heegaard surfaces of the blocks (we provide a proof of this fact in the appendix of this paper), and this allows us to compute the Heegaard genus of M_Q.

Every Heegaard surface induces a *bipartition*, a partition into two sets, of its manifold's boundary components. The blocks are constructed so that each block emulates its logical operator via the way its minimal genus Heegaard surfaces bipartition its boundary components. The OR block is flexible, in that every non-trivial bipartition is possible, whereas all other block types have a fixed bipartition of

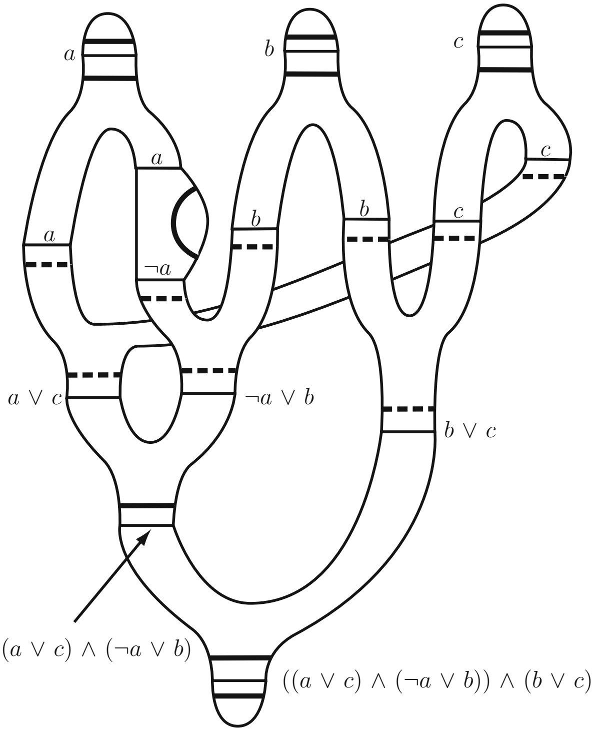

a
b
c
a
a
$\neg a$
b
b
c
c
$a \vee c$
$\neg a \vee b$
$b \vee c$
$(a \vee c) \wedge (\neg a \vee b)$
$((a \vee c) \wedge (\neg a \vee b)) \wedge (b \vee c)$

Fig. 1 The construction of M_Q, where $Q = ((a \vee c) \wedge (\neg a \vee b)) \wedge (b \vee c)$

boundary components determined by the minimal genus Heegaard surfaces. When Q is satisfiable, there is a minimal genus Heegaard surface for each block so that the complementary pieces can be bicolored in a particular way (see Definition 2.7) so that the Heegaard surfaces for the blocks can be amalgamated to a genus $|Q| + 2$ Heegaard surface for M_Q. The converse uses the same setup. We show that the genus of M_Q is at least $|Q| + 2$, and that when equality is achieved it is possible to read off a satisfying assignment for Q from a bicoloring induced by Heegaard surfaces for the block manifolds.

There are many manifolds that fit the above description of M_Q. The second step, from which Theorem 1.2 follows, is that we can construct a triangulation for one efficiently.

Proposition 4.1 *A triangulated M_Q can be produced in quadratic time (and tetrahedra) in $|Q|$.*

The essential ingredient for our main result is our ability to choose block manifolds whose minimal genus Heegaard surfaces bipartition their boundary components in a way that emulates the required logical operators. It is then worth asking: given a set of bipartitions, is there a 3-manifold whose minimal genus Heegaard surfaces induce precisely that set? In fact, this is an easy corollary of the techniques we use here.

Corollary 3.8 *Let $\mathcal{P}$ be a non-empty set of bipartitions of* $1, 2, \ldots, n$*. Then there is a 3-manifold X and a numbering of its boundary components,* $1, 2, \ldots, n$*, so that the set of bipartitions of ∂X induced by minimal genus Heegaard splittings of X is precisely $\mathcal{P}$.*

This paper is organized as follows: Sect. 2 contains the required background on Heegaard splittings, surfaces, and amalgamation. Section 3 gives a recipe for producing M_Q and proves Proposition 3.1 and Corollary 3.8. Section 4 shows how to triangulate M_Q and proves Proposition 4.1. Section 5 lists some related open questions. The appendix proves Proposition 1, which explains how high distance gluings ensure that minimal genus Heegaard surfaces are amalgamations.

2 Heegaard Splittings and Amalgamations

Definition 2.1 Consider a 3-ball B, and attach 1-handles to ∂B. The resulting 3-manifold is a *handlebody*. Alternatively, let F be a closed, not necessarily connected, orientable surface such that each component of F has genus greater than zero. Take the product $F \times [0, 1]$ and attach 1-handles along $F \times \{1\}$. Assuming it is connected, the resulting 3-manifold V is a *compression body*, and we denote $\partial_- V = F \times \{0\}$ and $\partial_+ V = \partial V - \partial_- V$. (We will consider a handlebody as a compression body with $\partial_- V = \emptyset$.)

Let M denote a compact, connected, orientable 3-manifold.

Definition 2.2 A *Heegaard splitting* for M is a decomposition $M = V \cup W$ where V and W are compression bodies such that $\partial_+ V = \partial_+ W = V \cap W$. The surface $H = \partial_+ V = \partial_+ W$ in M is called a *Heegaard surface*, and when needed we may include this surface in the notation for the Heegaard splitting as $V \cup_H W$. The *genus* of $V \cup_H W$ is the genus of H, denoted $g(H)$.

Remark 2.3 Note that the compression bodies V and W bipartition the boundary of M into $\partial_V M = \partial M \cap V = \partial_- V$ and $\partial_W M = \partial M \cap W = \partial_- W$. In particular, a Heegaard splitting for M always induces a bipartition $\{\partial_V M | \partial_W M\}$ of the boundary components of M, and thus it is proper to say that $V \cup W$ is a Heegaard splitting of M with respect to the bipartition $\{\partial_V M | \partial_W M\}$.

Given M, one can find Heegaard splittings of M in several ways. For example, if M is triangulated with t tetrahedra, then one can obtain a Heegaard splitting of M of genus $t + 1$, taking the boundary of a regular neighborhood of the 1-skeleton as the Heegaard surface. Alternatively, if M can be decomposed as a union of submanifolds $M = \bigcup M_i$, so that M is obtained by gluing the M_i together along their boundary components (including possible self-gluings), one can potentially *amalgamate* Heegaard splittings of the M_i to form a Heegaard splitting of M:

Example 2.4 Let M_1 and M_2 be 3-manifolds such that $\partial M_1 \cong \partial M_2 \cong F$, and let $V_1 \cup W_1$ be a Heegaard splitting of M_1 with respect to the bipartition $\{\emptyset | \partial M_1\}$ and

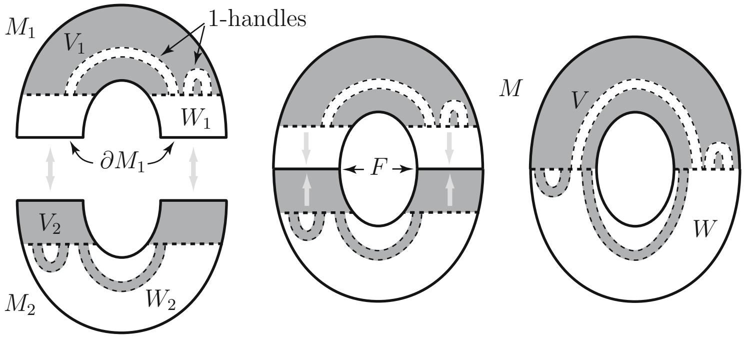


Fig. 2 A schematic for the amalgamation given in Example 2.4. The *light* and *dark regions* represent compression bodies, with W_1 and V_2 expressed as $F \times [0,1] \cup$ (1-handles). The *dotted lines* represent Heegaard surfaces

$V_2 \cup W_2$ a Heegaard splitting of M_2 with respect to the bipartition $\{\partial M_2|\emptyset\}$. Note that both W_1 and V_2 are compression bodies of the form $F \times [0,1] \cup \{\text{1-handles}\}$. Form the 3-manifold M by gluing M_1 to M_2 along their boundaries, and, abusing notation slightly, let F be the image of the boundary components in M. Collapse the product structures in W_1 and V_2 so that in each, $F \times [0,1]$ is mapped to $F \times \{0\} = F$, and so that the 1-handles of each of W_1 and V_2 are attached disjointly on F. We then obtain a new Heegaard splitting $V \cup W$ of M, where $V = V_1 \cup \{\text{1-handles in } V_2\}$, and $W = \{\text{1-handles in } W_1\} \cup W_2$. The splitting $V \cup W$ is called the *amalgamation* of $V_1 \cup W_1$ and $V_2 \cup W_2$ along F. See Fig. 2.

Constructing an amalgamation of $M = \bigcup M_i$ from component Heegaard splittings of M_i, however, is not always possible.

Example 2.5 Suppose M is formed by taking $M_1 = T^2 \times [0,1]$ and gluing the two components of ∂M_1 together. Let F be the image of ∂M_1 (an embedded torus) in M.

It is well known that M_1 admits two irreducible Heegaard surfaces up to isotopy [32]: a "Type 1" surface that is a level torus $T^2 \times \{\frac{1}{2}\}$ and induces the non-trivial bipartition of boundary components $\{T^2 \times \{0\}|T^2 \times \{1\}\}$, and a "Type 2" surface that is a genus two Heegaard surface obtained by tubing together two disjoint copies, say $T^2 \times \{\frac{1}{4}\}$ and $T^2 \times \{\frac{3}{4}\}$, of the level surface. Note that this latter surface induces the trivial bipartition of boundary components $\{T^2 \times \{0\}, T^2 \times \{1\}|\emptyset\}$.

One *cannot* form an amalgamated splitting for M by taking a Type 2 Heegaard splitting of M_1 and amalgamating it to itself (See Fig. 3a). This is because in attempting to apply the construction of Example 2.4, we do not end up with two resulting compression bodies once we collapse the product structure of $F \times [0,1]$ (i.e. the resulting "Heegaard surface" is not separating).

Example 2.6 Let M_1 and M_2 each be copies of $T^2 \times [0,1]$, and form $M = M_1 \cup M_2$ by gluing ∂M_1 to ∂M_2 component-wise. Let $F = \partial M_1 = \partial M_2$, so that F consists of two disjoint tori embedded in M. Then, one cannot form an amalgamated Heegaard splitting of M from Type 1 Heegaard splittings of M_1 and M_2 (See Fig. 3b). The issue

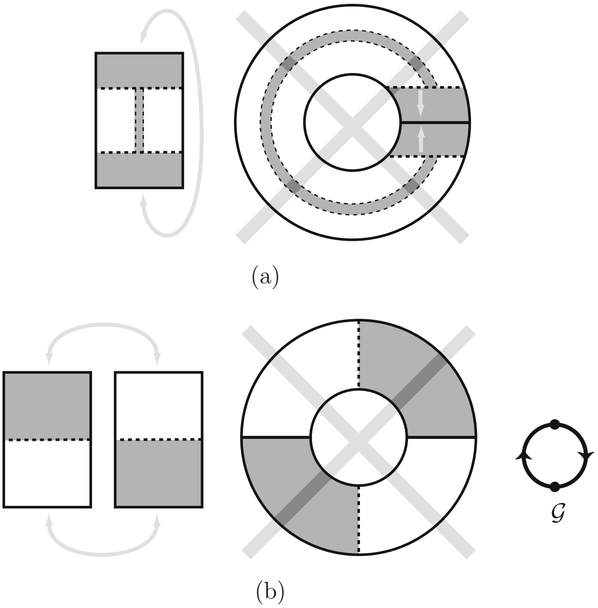


Fig. 3 (**a**) A Type 2 Heegaard splitting of $T^2 \times [0, 1]$ cannot be amalgamated to itself; (**b**) two Type 1 Heegaard splittings of $T^2 \times [0, 1]$ cannot be amalgamated together (Note that $\mathcal{G}$ here is not a DAG)

here is that the Heegaard splitting of M_i, $i = 1, 2$, does not partition the components of ∂M_i into a single compression body, and thus one cannot simultaneously collapse the product structure $F \times [0, 1]$ along each component of F as in Example 2.4 to form an amalgamation.

Assume that $M = \bigcup M_i$ where the M_i meet along boundary components. Rather than thinking of the M_i in a linear order, it is more natural to consider the following construction. Let $\mathcal{G}$ be the dual graph of $\bigcup M_i$, so that each submanifold M_i is assigned a vertex x, and two vertices corresponding to M_i and M_j are connected by an edge for each component of $\partial M_i \cap \partial M_j$. (Note that i may equal j, in the case of self-gluings.) Relabelling the submanifolds M_i as M_x, one for each vertex x of $\mathcal{G}$, we can consider $M = \bigcup_{x \in \mathcal{G}} M_x$. The following definition provides the conditions under which Heegaard splittings of the M_x can form an amalgamated Heegaard splitting of M.

Definition 2.7 A *generalized Heegaard splitting* of $M = \bigcup_{x \in \mathcal{G}} M_x$ is a choice, for each M_x, of a Heegaard splitting $M_x = V_x \cup W_x$, so that:

(1) The compression bodies are bicolored "black" and "white" (or "V" and "W"). That is $V_x \cap V_{x'} = \emptyset$, $W_x \cap W_{x'} = \emptyset$, for all $x \neq x'$.
(2) Given this bicoloring, the graph $\mathcal{G}$ becomes a directed acyclic graph (DAG) after assigning edges of $\mathcal{G}$ to point toward "white": as each edge e of $\mathcal{G}$ is dual

to a surface in M that has a black compression body V_x on one side and a white compression body $W_{x'}$ on the other, assign an orientation to e that points from x to x' ("black" to "white"). We require that the resulting directed graph has no directed cycles.

Theorem 2.8 *If $\bigcup_{x\in\mathcal{G}}(V_x \cup W_x)$ is a generalized Heegaard splitting of $M = \bigcup_{x\in\mathcal{G}} M_x$, then the Heegaard splittings $V_x \cup W_x$ can be amalgamated to form a Heegaard splitting of M.*

Proof We construct the desired Heegaard splitting in stages. Assume that the graph $\mathcal{G}$ is directed as per Definition 2.7. As $\mathcal{G}$ contains no directed cycles, the graph has a vertex which is a sink (all edges meeting it point "in"). Remove this vertex and all edges meeting it from the graph. In the remaining (potentially disconnected) graph, find another sink, and repeat the process. Continue until all such sinks have been removed. As $\mathcal{G}$ is a DAG, this means we are left only with a collection of vertices (the sources of the original graph).

Now add back the last removed sink x_0, along with the edges $e_1, \ldots, e_m$ that point in toward it. Let $x_1, \ldots, x_n$ be the set of vertices that bound the edges $e_1, \ldots, e_m$ along with x_0. Since x_0 is a sink, the bicoloring of the compression bodies of $\bigcup V_{x_i} \cup W_{x_i}$ in the generalized Heegaard splitting implies M_{x_0} meets each M_{x_i} only in W_{x_0} and V_{x_i}, $i = 1, \ldots, n$, respectively. In particular, the components $F_{e_1}, \ldots, F_{e_m}$ of ∂M_{x_0} corresponding to the edges $e_1, \ldots, e_m$ are all contained in W_{x_0} and $\bigcup V_{x_i}$. Thus, we may carry out the procedure of Example 2.4 and collapse the product structures $F_{e_j} \times [0, 1]$ to F_{e_j} simultaneously for all j in the compression bodies $W_{x_0}, V_{x_1}, \ldots, V_{x_n}$ and obtain a new Heegaard splitting $V' \cup W'$ of $M' = M_{x_0} \cup \ldots \cup M_{x_n}$. Note that this new Heegaard splitting preserves the original bicoloring given by $\bigcup V_x \cup W_x$ for boundary components of M': if F' is a component of $\partial M'$, then $F' \subset \partial V'$ if and only if $F' \subset \partial V_{x_i}$ for some x_i. (Boundary components of M' stay "black" or "white.")

Add back in the next sink x_0'. If $M_{x_0'}$ does not meet M', then we simply repeat the above process for the subset of $\mathcal{G}$ that consists of edges and bounding vertices that meet x_0'. If $M_{x_0'}$ meets M', then we consider M' as a whole with the Heegaard splitting $V' \cup W'$ obtained above. Since $V' \cup W'$ preserves the bicoloring of boundary components of M' given by the original generalized Heegaard splitting, we can repeat the above process to obtain a new Heegaard splitting of $M_{x_0'} \cup M' \cup \{M_y \mid y \text{ is a new vertex directed towards } x_0'\}$.

Building in this way, we can continue to obtain new Heegaard splittings of larger collections of submanifolds of M, until we complete the graph $\mathcal{G}$ and produce a Heegaard splitting $V \cup W$ of M. □

As before, the Heegaard splitting $V \cup W$ obtained in the above proof is called the amalgamation of the Heegaard splittings of the M_x along the surfaces F, where F is the collection of components of the ∂M_x that are dual to edges in $\mathcal{G}$ (i.e. $F = \left(\bigcup_{x\in\mathcal{G}} \partial M_x\right) \setminus \partial M$). Note that $V \cup W$ is obtained by sequential applications of the technique in Example 2.4 to amalgamations of Heegaard splittings of "sink" submanifolds to their adjacent submanifolds. The critical feature of a generalized Heegaard splitting that allows one to construct $V \cup W$ is that each component Heegaard splitting bipartitions the boundary components of the M_x suitably so that we can bicolor the set of compression bodies (this allows us to end up with two compression bodies in the amalgamated Heegaard splitting, avoiding the problem of Example 2.5), and can use the bicoloring to direct the edges of $\mathcal{G}$ so that we can amalgamate in sequence "outward" from sinks at each stage (thereby avoiding the problem of Example 2.6 – recall Fig. 3).

Theorem 2.9 *Suppose $\bigcup_{x\in\mathcal{G}}(V_x \cup_{H_x} W_x)$ is a generalized Heegaard splitting of $M = \bigcup_{x\in\mathcal{G}} M_x$. For every edge e of $\mathcal{G}$, let F_e denote the component of $\bigcup \partial M_x$ dual to e in M. Let $V \cup_H W$ be the amalgamation of $\bigcup_{x\in\mathcal{G}}(V_x \cup_{H_x} W_x)$. Then*

$$g(H) = \sum_{x\in\mathcal{G}} g(H_x) - \sum_{e\in\mathcal{G}} g(F_e) + 1 - \chi(\mathcal{G}).$$

Proof Proceed with the same setup and notation as in the proof of Theorem 2.8. In particular, for the first step in constructing an amalgamation of M, consider Heegaard splittings $V_{x_i} \cup_{H_{x_i}} W_{x_i}$ of M_{x_i}, $i = 0, \dots, n$, respectively, and their corresponding vertices $x_0, \dots, x_n$ and connecting edges $e_1, \dots, e_m$ in $\mathcal{G}$. Let $F_{e_1}, \dots, F_{e_m}$ denote the corresponding surfaces in M dual to $e_1, \dots, e_m$. Let $M' = \bigcup_{i=0}^{n} M_{x_i}$.

By construction, the genus of the amalgamated Heegaard splitting is obtained by adding the genus of H_{x_0} to the *handle numbers* of V_{x_i}, $i = 1, \dots, n$. If V is a compression body, then the handle number of V is the number of 1-handles added to $\partial_- V \times [0, 1]$ along $\partial_- V \times \{1\}$ to obtain V (see Fig. 4). There are two types of potential such 1-handles: a minimal set that connects components of $\partial_- V \times [0, 1]$ (essentially fulfilling the role of "connected sum" of components of $\partial_- V \times \{1\}$), and those that increase the genus of $\partial_+ V$. Thus, the handle number of V equals

$$\#_{handle}(V) = g(\partial_+ V) - \sum_{F\in\partial_- V} g(F) + |\partial_- V| - 1.$$

Let $V' \cup_{H'} W'$ be the amalgamation of $\bigcup_{i=0}^{n}\left(V_{x_i} \cup_{H_{x_i}} W_{x_i}\right)$. Using the handle number, the genus of the Heegaard surface H' is

$$g(H') = g(H_{x_0}) + \sum_{i=1}^{n} \#_{handle}(V_{x_i}).$$

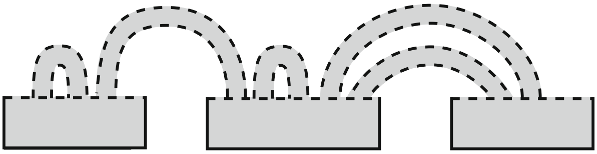


Fig. 4 A schematic of a compression body V with $\#_{handle}(V) = 5$ (Note that $\partial_+ V$ is denoted by *dotted lines*)

Plugging in the equations for the handle numbers for the V_{x_i} produces

$$\begin{aligned} g(H') &= g(H_{x_0}) + \sum_{i=1}^{n} g(H_{x_i}) - \sum_{j=1}^{m} g(F_{e_j}) + \left| \bigcup_{j=1}^{m} F_{e_j} \right| - n \\ &= \sum_{i=0}^{n} g(H_{x_i}) - \sum_{j=1}^{m} g(F_{e_j}) + m - n. \end{aligned}$$

Let $\mathcal{G}'$ denote the graph connecting x_0 to $x_1, \ldots, x_n$. Since m is the number of edges in $\mathcal{G}'$ and n is the number of vertices minus one, we conclude $m - n = 1 - \chi(\mathcal{G}')$. Hence

$$g(H') = \sum_{i=0}^{n} g(H_{x_i}) - \sum_{j=1}^{m} g(F_{e_j}) + 1 - \chi(\mathcal{G}').$$

For any new submanifold that is included in the amalgamation at a subsequent stage, the above relationship is preserved. That is, suppose that $M' = V' \cup_{H'} W'$ has already been obtained as above by amalgamating component Heegaard splittings, and suppose $M_y = V_y \cup_{H_y} W_y$ is a submanifold and Heegaard splitting being newly amalgamated to $V' \cup_{H'} W'$ along surfaces $F_{e'_1}, \ldots, F_{e'_{m'}}$. Let $\mathcal{G}'$ and $\mathcal{G}'_y$ be the dual graphs for M' and $M' \cup M_y$, respectively. Repeating the above argument implies that the genus of the resulting amalgamation of $M' \cup M_y$ increases by

$$g(H_y) - \sum_{k=1}^{m'} g(F_{e'_k}) + m' - 1.$$

Note that m' is the number of edges of $\mathcal{G}'_y \setminus \mathcal{G}'$, and so $m' - 1 = -\chi(\mathcal{G}'_y \setminus \mathcal{G}')$. In particular, this means that $m - n + m' - 1 = 1 - \chi(\mathcal{G}') - \chi(\mathcal{G}'_y \setminus \mathcal{G}') = 1 - \chi(\mathcal{G}'_y)$. Thus, the resulting genus of the amalgamation of $M' \cup M_y$ is

$$\sum_{x \in \mathcal{G}'_y} g(H_x) - \sum_{e \in \mathcal{G}'_y} g(F_e) + 1 - \chi(\mathcal{G}'_y).$$

Amalgamating thusly along all remaining submanifolds $M_{y'}$, $y' \in \mathcal{G}$, produces the desired result. □

It is important to note that one can find examples of (minimal genus) Heegaard splittings of 3-manifolds that are not amalgamations. For example, by gluing the bridge surface of a tunnel number $n - 1$, n-bridge knot complement to vertical annuli in a Seifert fibered space over a disk with n exceptional fibers, one can obtain a Heegaard surface of the resulting 3-manifold of genus n, whereas the minimal genus amalgamation along the gluing surface has genus $2n$. (See [36].) Note that this Heegaard surface results from a very specific gluing map between the boundary components of the two submanifolds. In general, gluing maps between boundary components can be chosen to be "sufficiently complicated" to ensure that all minimal genus Heegaard splittings are amalgamations along the gluing surfaces. (See the appendix.) Exploiting this property in the next sections allows us to ensure that the minimal genus Heegaard splittings of our constructed 3-manifolds M_Q are amalgamations, to which we can thus apply the results of this section.

3 Constructing M_Q

In this section we give a recipe for producing M_Q from Q and prove the following result.

Proposition 3.1 *Let Q be an instance of* CNF-SAT. *Then there is a manifold M_Q with Heegaard genus $g(M_Q) \geq |Q| + 2$, with equality holding if and only if Q has a satisfying assignment.*

Recall that $|Q|$ is the length of Q without counting parentheses.

3.1 Constructing M_Q

The sentence Q will guide our construction of M_Q. To begin, rewrite Q by inserting parentheses, if necessary, to make it clear how each logical connective joins exactly two terms (i.e. Q is made fully parenthesized). The manifold M_Q is then constructed out of building blocks according to instructions provided by this modified version of Q. Each building block will have Heegaard genus 2 and some number of torus boundary components. Each such boundary component will be labelled with a subsentence of Q, and also be designated as either an *input* or an *output* to that block. We will depict such blocks so that the input boundary component is on top, and the outputs are on the bottom. See Fig. 5. Each block is chosen based on a desired bipartitioning of its boundary components by genus 2 Heegaard splittings as follows.

- VAR(iable) – For each distinct variable in Q let the block manifold M be a trefoil knot exterior (Fig. 8). Then M has one torus boundary component, $\partial M = T$, and any genus 2 Heegaard splitting induces the only boundary bipartition possible (up to ordering), $\{T|\emptyset\}$. We label the boundary component T with the corresponding variable, and consider it an output of the block.
- REP(licate) – To create multiple copies of a given variable, we use a block manifold M that is the exterior of the twisted torus link in Fig. 9. Then M has three torus boundary components, $\partial M = T_0 \cup T_1 \cup T_2$ where any genus two Heegaard splitting induces the boundary bipartition $\{T_0, T_1|T_2\}$ (Lemma 4.9). All three components will be labelled with the variable that is being duplicated.

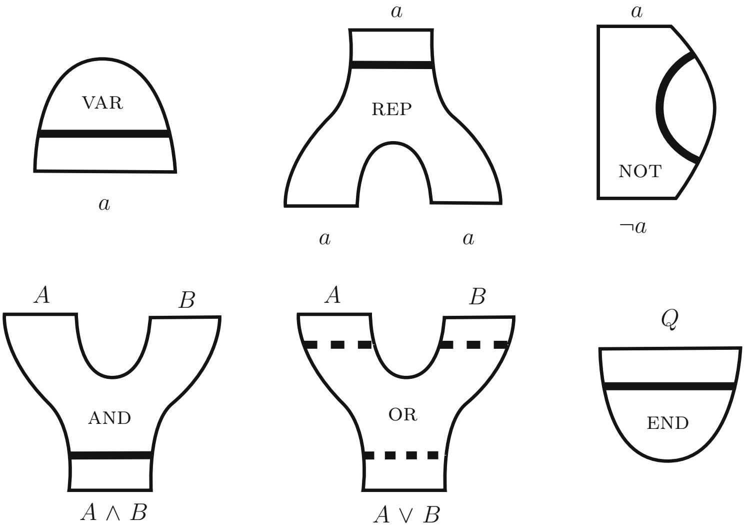


Fig. 5 Schematics indicating block types and their labelings. Input surfaces are depicted at the *top* of each block, and outputs at the *bottom*. Minimal genus Heegaard surfaces are depicted with *bold lines* (With the three possible such splittings of the OR block indicated with *bold dashed lines*)

We will say the boundary component T_2 is *preferred*, and will be the input. The other two boundary components are outputs.

- NOT – For each occurrence of "$\neg a$" in Q, the block manifold M will be a high distance filling on the twisted torus link as described in Lemma 4.10. Then $\partial M = T_0 \cup T_1$ and any genus two Heegaard splitting induces the bipartition $\{T_0, T_1|\emptyset\}$. Label one boundary component a, and consider it an input. The other boundary component is labelled $\neg a$ and is considered an output. Glue the input surface to the output of a REP block corresponding to a.

Once we have created one labeled output surface for each instance of each variable in Q, and each instance of its negation, we start gluing them to other kinds of blocks determined by the logical structure of Q, as follows:

- AND – For each conjunction $A \wedge B$ in Q, we let M be the exterior of the twisted torus link already used for REP. Then $\partial M = T_0 \cup T_1 \cup T_2$, and all genus two Heegaard splittings all induce the bipartition $\{T_0, T_1|T_2\}$ (Lemma 4.9). Label the preferred boundary component T_2 with the expression $A \wedge B$, and consider it an output. The other two boundary components are inputs, and are labelled with the expressions A and B respectively.
- OR – For each disjunction $A \vee B$ in Q we let M be the exterior of the three component chain indicated in Fig. 8. It is homeomorphic to *{pair of pants}* $\times S^1$, has three boundary components $\partial M = T_0 \cup T_1 \cup T_2$, and each of the three boundary bipartitions of the form $\{T_i, T_j|T_k\}$ is realized by some genus two

Heegaard splitting (Lemma 4.8). Choose one boundary component to label as $A \vee B$, and consider it an output. The remaining boundary components are inputs, and are labelled with the expressions A and B respectively.

- END – We end by capping the statement off with the same M, the trefoil knot exterior, used for VAR. The manifold M has one torus boundary component, $\partial M = T$, and a single boundary bipartition $\{T|\emptyset\}$. It is labelled with the entire expression Q, and is an input.

To glue the blocks, we choose "sufficiently complicated" maps so that every Heegaard splitting of M_Q of genus less than or equal to $|Q| + 2$ is an amalgamation of splittings of the blocks. (See the appendix.)

As an example, Fig. 1 gives the construction of the manifold M_Q from the expression

$$Q = ((a \vee c) \wedge (\neg a \vee b)) \wedge (b \vee c).$$

3.2 *Proof of Proposition 3.1*

Lemma 3.2 *The Heegaard genus of M_Q is at least $|Q| + 2$, and in the case of equality, any such minimal genus splitting is an amalgamation of minimal genus splittings of the building blocks.*

Proof Let S be a minimal genus Heegaard splitting of M_Q. If the genus of S is strictly greater than $|Q| + 2$ then the result follows. By way of contradiction, we assume the genus of S is at most $|Q| + 2$. By construction, S is then an amalgamation of Heegaard splittings of the building blocks. We now use Theorem 2.9 to compute the genus of S:

$$g(S) = \sum_{x \in \mathcal{G}} g(H_x) - \sum_{e \in \mathcal{G}} g(F_e) + 1 - \chi(\mathcal{G}).$$

Here $\mathcal{G}$ is the graph dual to the block structure. Let v be the number of vertices, one for each block, and e the number of edges, one for each gluing torus. Note that the number of variable occurrences in Q is the number of VAR and REP blocks. The operators in Q each have a corresponding NOT, OR, or AND block, and there is a final END block for the total statement Q. In particular, $v = |Q| + 1$. Since each block has genus 2, we have $g(H_x) \geq 2$ for each x, with equality holding only for those blocks with minimal splittings, and $g(F_e) = 1$ for each e. Thus,

$$g(S) \geq 2v - e + 1 - (-e + v) = v + 1 = |Q| + 2.$$

□

Lemma 3.3 *If the Heegaard genus of M_Q is equal to $|Q| + 2$ then there is a satisfying assignment of Q.*

Proof Suppose S is a minimal genus Heegaard surface of M_Q. If the genus of S is $|Q| + 2$, then by the previous lemma S is an amalgamation of minimal genus Heegaard surfaces $\{S_i\}$ in the building blocks.

Because S is an amalgamation, the surfaces $\{S_i\}$, together with the gluing surfaces, separate the manifold M_Q into compression bodies that can be colored "black" and "white" so that no two compression bodies with the same color are adjacent. Without loss of generality, we assume the compression body of the END block which contains its sole input surface is colored white.

We will now assign truth values to the gluing surfaces between blocks, according to this bicoloring. Let F be such a gluing surface. Then F is the input surface for some block. If the compression body in that block containing F is white, then we will say that F is *true*. Otherwise, we say it is *false*. Equivalently, we can say that F is *true* if it is the output of a block, and the compression body in that block that contains F is black. Thus, if the Heegaard surface in some block separates an input surface A of that block from an output surface, B, then A and B will have the same truth value. It follows immediately that the input and output surfaces of all REP blocks have the same truth value. Similarly, the truth value of the input of a NOT block labelled a will have the opposite truth value as the output labelled $\neg a$. Finally, note that the surface at the input of the END block (which we have labelled with the statement Q) is by choice assigned the truth value *true.*

In the next several claims, we show that our assignment of truth values respects the logical structure of the subsentences of Q that appear at the labels of (most of) the gluing surfaces.

Claim 3.4 *All surfaces at the inputs and outputs of the* AND *blocks are true.*

Proof The minimal genus Heegaard surface of an AND block separates the output surface from both inputs. Thus, the output and input surfaces all have the same truth value. The proof is complete by noting that since Q is in conjunctive normal form, the output of every AND block is glued to the input of the END block (a *true* surface), or the input of another AND block. □

We say an OR*-tree* is a component of the union of the OR blocks in M_Q.

Claim 3.5 *The output of every* OR*-tree is true, and at least one of the input surfaces of every* OR*-tree is true.*

Proof Let F_0 denote the output surface of an OR-tree. Since Q is in conjunctive normal form, F_0 is glued to the input of an AND block. By the previous claim, F_0 must be *true.* By construction, the Heegaard surface of the OR block that contains F_0 separates it from at least one of the input surfaces F_1 of that block. Thus, F_1 will also be *true*. Working up the tree, we now consider the OR block in the tree whose output is the surface F_1. By identical reasoning, one of its input surfaces F_2 must be *true* as well. Continuing in this way we eventually reach an input surface F_i of the entire OR-tree and conclude that it must be *true.* □

Note that some of the truth values of the sentences that label gluing surfaces interior to an OR-tree may not be correct, but the previous claim shows this does not disturb the logical structure of the OR-tree, taken as a whole.

To complete the lemma, note that we have assigned a truth value to the output surface of every VAR block. These surfaces correspond to the variables used in the sentence Q. We have shown above that our assignment of truth values to the input and output surfaces of REP, NOT, and AND blocks, as well as OR trees, respects the logical structure of the sentences that label them. Thus, we have produced an assignment of truth values for the variables that make the statement Q *true*. □

Lemma 3.6 *If there is a satisfying assignment of Q, then the Heegaard genus of M_Q is equal to $|Q| + 2$.*

Proof If there is a satisfying assignment of Q, then that assignment gives a truth value to each expression at the gluing surfaces. In this way, each boundary component of each building block gets assigned a truth value. We color the sides of each such surface black/white so that if F is a *true* surface at the output of a block, then the side of F facing into that block is black. Similarly, if F is a *true* surface at the input of a block, then the side facing in is colored white. Conversely, the side of a *false* surface at the output of a block is colored white, and the side of a *false* surface at an input is black.

Claim 3.7 *There is a minimal genus splitting of each block that separates all white surfaces on the inside of the block from all black surfaces facing in.*

Proof Consider first the END block. Since there is only one boundary component, any Heegaard splitting (and in particular the minimal genus one) has the desired separation property.

Next we consider the AND blocks. Since Q is in conjunctive normal form, the output of each such block is either attached to the END block, or another AND block. Hence, if there is a satisfying assignment for Q then the labels at every input and output surface of an AND block are *true* logical sentences. It follows that the side of the input surfaces that face into such a block are white, and the side of the output surface facing into the block is black. Such a block has the output as a preferred boundary component, meaning that a minimal genus splitting separates the output surface from both input surfaces. Hence, the minimal genus splitting has the desired separation property.

An OR block has no preferred boundary component. Thus, there is a minimal genus splitting for each non-trivial bipartitioning of the boundary components. It follows that the only way the separation property can fail is if the side of every boundary surface facing in to the block is the same color. If they are all white, then this corresponds to both inputs being *true*, and the output being *false*. If they are all black, then both inputs are *false*, and the output is *true*. Neither situation obeys the properties of the logical "or" operation, so we will not see these sets of truth values for the labels of the surfaces at the boundary of an OR block.

By construction, a REP block has the same logical value at each input and output. If they are all *true*, then the side of the input surface that faces into the block is white, and the side of the outputs that faces in is black. The input surface of this block is a preferred boundary component, so the minimal genus splitting separates black from white as desired. If all surfaces are *false*, the situation is reversed.

Finally, we consider the NOT blocks. The sentences at the boundary components of a NOT block will have opposite truth values. Thus, the side of the input surface facing into the block will have the same color as the side of the output surface facing in. Both surfaces are on the same side of a minimal genus splitting of a NOT block. □

Assume we have now chosen splittings of each block in accordance with the conclusion of Claim 3.7. Then the building blocks are separated into compression bodies by these splittings, and these compression bodies inherit the color black or white, according to the colors of their negative boundaries. Furthermore, because opposite sides of any single gluing surface are different colors, it follows that neighboring compression-bodies in M_Q are colored differently.

According to Theorem 2.8, to show that we can amalgamate our choice of splittings of the building blocks, it remains to show that the directed graph $\mathcal{G}$ that is dual to the gluing surfaces has no directed cycles. (Recall that each edge of this graph is oriented so that it passes from a black compression body into a white one.)

We have constructed M_Q vertically so that the output surface(s) of any given block is below its input surface(s). Any directed cycle must have a local maximum, x. Let e_1 and e_2 be the edges of the cycle that meet x, where e_1 is oriented toward x, and e_2 is oriented away. As x is a local maximum, both e_1 and e_2 correspond to output surfaces of the building block corresponding to x. It follows that this building block is a REP block, as this is the only type of block that has two output surfaces. However, according to our coloring scheme, both output surfaces of a REP block are on the boundary of the same compression body. If this compression body is black, then both e_1 and e_2 are oriented away from x. If the compression body is white, then both are oriented toward x. This contradiction establishes that there are no directed cycles in $\mathcal{G}$.

By Theorem 2.8 we can now amalgamate the chosen splittings of our building blocks, creating a splitting of M_Q. By the computation given in the proof of Lemma 3.2, the genus of this splitting is $|Q| + 2$. □

Finally, note that if one were to remove the VAR blocks from M_Q, we would obtain a manifold with a boundary component corresponding to each variable, and, for each satisfying assignment, a minimal genus Heegaard splitting that induces a $\{true \mid false\}$ bipartition of the corresponding boundary components. That is the basis for the following corollary.

Corollary 3.8 *Let $\mathcal{P}$ be a non-empty set of bipartitions of $1, 2, \ldots, n$. Then there is a 3-manifold X and a numbering of its boundary components, $1, 2, \ldots, n$, so that the set of bipartitions of ∂X induced by minimal genus Heegaard splittings of X is precisely $\mathcal{P}$.*

Proof Suppose that P is a bipartition of $1, .., n$. That is, $P = \{P_+|P_-\}$ so that $P_+ \cup P_- = 1, .., n$ and $P_+ \cap P_- = \emptyset$. Let $v_i, i = 1, .., n$ be variables and let the clause

$q(P)$ be a conjunction of each variable or its negation, depending on which side of the bipartition P its index belongs to:

$$q(P) = \bigwedge\{v_i | i \in P_+\} \bigwedge\{\neg v_i | i \in P_-\}.$$

Of course, $q(P)$ accepts exactly one satisfying assignment, and that corresponds (via the correspondence $i \in P_+ \iff v_i = \mathit{true}$) to the bipartition P. Now let $\mathcal{P}$ be a set of bipartitions of $1, \ldots, n$ and let $\mathcal{P}^C$ be its complement, i.e. the set of bipartitions not in $\mathcal{P}$. Let

$$Q(\mathcal{P}^C) = \bigvee\{q(P) | P \in \mathcal{P}^C\}$$

Now, let $Q = Q(\mathcal{P}) = \neg Q(\mathcal{P}^C)$ which, after applying De Morgan's laws, is an instance of CNF-SAT. Let M_Q be built according to the procedure above. Now it is easy to check that satisfying assignments are in 1-1 correspondence with bipartitions $P \in \mathcal{P}$, again by using the correspondence $i \in P_+ \iff v_i = \mathit{true}$.

Let M_Q be constructed as before. Note that since Q is satisfiable, M_Q has Heegaard genus $|Q| + 2$. Let M'_Q be the manifold obtained by removing each VAR block. Because each VAR block removed is a leaf in $\mathcal{G}$, the graph dual to the block structure, the proofs of Lemmas 3.3 and 3.6 apply to M'_Q as well as to M_Q. In particular, a minimal genus splitting of M'_Q determines a satisfying truth assignment to the v_i's, and vice-versa. Note that each v_i labels a boundary component of M'_Q, and each minimal genus splitting separates the *true* variables from the *false* variables, so bipartitions induced by minimal genus splittings are in 1-1 correspondence with satisfying assignments which in turn are in 1-1 correspondence with bipartitions $P \in \mathcal{P}$ (via $i \in P_+ \iff v_i = \mathit{true}$). □

4 Triangulating M_Q

In this section, we describe how to triangulate the manifold M_Q so that the number of tetrahedra used is at most quadratic in $|Q|$, the length of the statement Q. Our goal is the following:

Proposition 4.1 *A triangulated M_Q can be produced in quadratic time (and tetrahedra) in $|Q|$.*

We proceed in several steps. First, in Sects. 4.1 and 4.2 we give a method to perform high distance triangulated gluings via *layered triangulations*. For the most part, these are not new results. Our statements about distances in the Farey graph in Sect. 4.1 are certainly well known, and layered triangulations (Sect. 4.2) are described by Jaco and Rubinstein in [12]. We include these sections, instead of just citing earlier work, because they are both accessible to the non-expert and also make explicit the relationship between the distance of the gluing and the number of layers.

Next, in Sect. 4.3, we give a topological description of block manifolds whose boundary components are appropriately bipartitioned by minimal genus Heegaard splittings. We consolidate some well known results and substantially leverage the work of Morimoto, Sakuma, and Yokota on Heegaard splittings of twisted torus knots [27], and the work of Moriah, Rieck, Rubinstein and Sedgwick that characterizes how and when a Dehn filling creates new Heegaard splittings [22, 25, 28–30].

We conclude, in Sect. 4.4, with a proof of Proposition 4.1 that describes how the blocks can be triangulated and then glued together.

4.1 Slopes and the Farey Graph

A *slope* is the isotopy class of an essential simple closed curve on a torus. Fix a pair of basis elements for the homology, $\mathbb{Z} \times \mathbb{Z}$, of the torus. Then any slope can be written as a pair (a, b), and because it is realized by a simple (connected) curve, we have $\gcd(a, b) = 1$. The usual convention is thus to represent the slope by the extended rational $\frac{a}{b} \in \mathbb{Q} \cup \{\infty\}$, where $\infty = \frac{1}{0}$.

We say that a pair of slopes have *distance one* if there are a pair of curves representing the slopes that intersect transversely in a single point. It is well known that a pair of slopes have distance one if and only if their extended rationals (with respect to any basis), $\frac{a}{b}$ and $\frac{c}{d}$, satisfy $|ad - bc| = 1$.

Definition 4.2 Let T be a torus. The *Farey graph for T* is the graph whose vertex set is the set of slopes and whose edges join any pair of vertices whose underlying slopes have distance one. Of course, after choosing a basis for homology, we are able to label each vertex of the graph with an extended rational $\frac{a}{b} \in \mathbb{Q} \cup \{\infty\}$. Each edge then joins a pair of extended rationals, $\frac{a}{b}$ and $\frac{c}{d}$, which satisfies $|ad - bc| = 1$.

Definition 4.3 If α and β are slopes in a torus T, then the *Farey distance* between them $d_{\mathcal{F}}(\alpha, \beta)$ is their distance in the Farey graph. If $a \subset T$ and $b \subset T$ are closed essential curves, then we define their distance, $d_{\mathcal{F}}(a, b) = d_{\mathcal{F}}(\alpha, \beta)$, to be the distance between α and β, isotopy classes of single components of a and b, respectively.

Form a 2-complex, the *curve complex of the torus T*, by attaching to the Farey graph a triangular face for every triple of slopes that pairwise intersect once. Fixing a basis for T, every edge is specified by a pair $\left(\frac{a}{b}, \frac{c}{d}\right)$ satisfying $|ad-bc| = 1$. It is not hard to see that in the curve complex, there are precisely two triangles, $\left(\frac{a}{b}, \frac{c}{d}, \frac{a+c}{b+d}\right)$ and $\left(\frac{a}{b}, \frac{c}{d}, \frac{a-c}{b-d}\right)$ attached to the edge $\left(\frac{a}{b}, \frac{c}{d}\right)$. This is described by the well known *Farey tessellation* of the Poincaré disk model of $\mathbb{H}^2$, see Fig. 6.

Moreover, each triangular face identifies a triangulation of the torus T up to isotopy: The slopes $\frac{a}{b}$ and $\frac{c}{d}$ can be realized by a pair of curves in the torus meeting in a single point. Together, they cut the torus into a rectangle. This rectangle has exactly two choices for a diagonal curve, with slopes $\frac{a+c}{b+d}$ and $\frac{a-c}{b-d}$ when connected through the intersection point. Choose one, say $\frac{a+c}{b+d}$. Then the triple of curves $\left(\frac{a}{b}, \frac{c}{d}, \frac{a+c}{b+d}\right)$

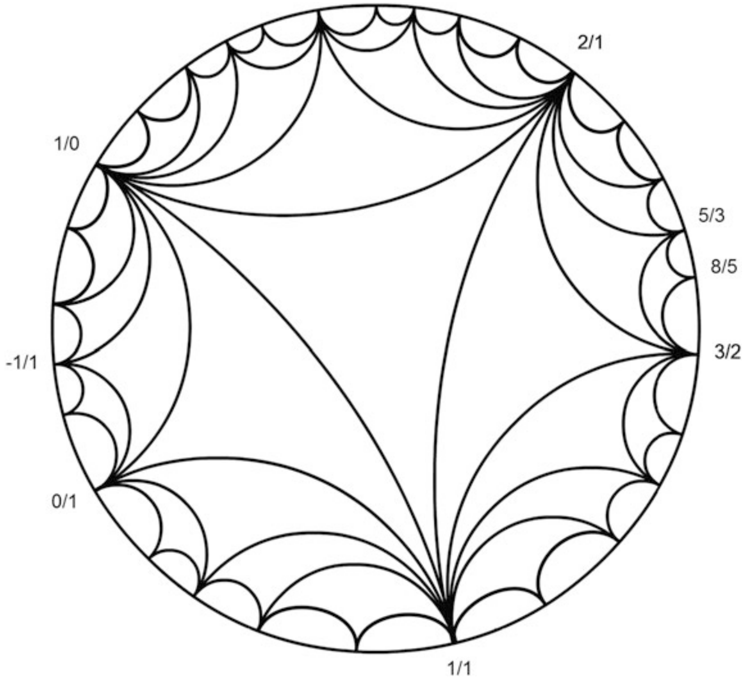


Fig. 6 The Farey tessellation of the Poincaré disk

intersect in a single common point. Treating that point as a vertex, we have formed a (non-simplicial) triangulation of the torus T with one vertex, three edges and two faces. We call this a *one-vertex triangulation of the torus*. Note that the two triangulations $\left(\frac{a}{b}, \frac{c}{d}, \frac{a+c}{b+d}\right)$ and $\left(\frac{a}{b}, \frac{c}{d}, \frac{a-c}{b-d}\right)$ meeting the edge $\left(\frac{a}{b}, \frac{c}{d}\right)$ are related by a *diagonal flip*, that exchanges the diagonal $\frac{a+c}{b+d}$ for the diagonal $\frac{a-c}{b-d}$, or vice-versa.

4.2 Layering

Later we will assume that our manifold X has been endowed with a triangulation that restricts to a one vertex triangulation of each of its torus boundary components [11].

Let e be an edge in the triangulation of the boundary torus $T \subset \partial X$. Then e can be regarded as the diagonal of a rectangle R bounded by the other two edges. Picture a new tetrahedron, Δ, as being a slightly thickened horizontal rectangle. Its bottom is a rectangle R_Δ with diagonal e_Δ and its top is a rectangle R'_Δ with diagonal e'_Δ. See Fig. 7. One can form a new triangulated manifold $X' = X \cup_{R=R_\Delta} \Delta$, by gluing R to R_Δ so that the diagonals e and e_Δ are identified. This

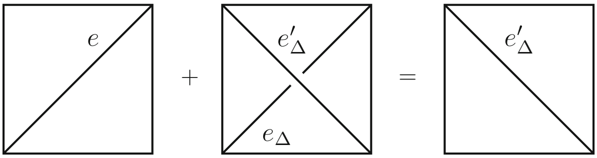


Fig. 7 Layering a tetrahedron on the boundary swaps a diagonal

process is called *layering at e* (see also [13]). It is not hard to see that the manifold X' is homeomorphic to X (as it retracts onto X) but that the boundary triangulation has changed. In particular, while e is no longer in the boundary torus, the boundary of R is still in the boundary torus, but its diagonal is now opposite and realized by e'_Δ. Thus, layering at e performs a diagonal flip on e in the boundary triangulation. The two triangulations are represented in the Farey tessellation by a pair of triangles that share a common edge.

Lemma 4.4 *Let $T \subset \partial X$ have a one-vertex triangulation with edge slopes $\left(\frac{0}{1}, \frac{1}{0}, \frac{1}{1}\right)$. Then, by layering on k tetrahedra, we can obtain a new triangulation of X with edge slopes $\left(\frac{F_{k-1}}{F_{k-2}}, \frac{F_k}{F_{k-1}}, \frac{F_{k+1}}{F_k}\right)$, where F_k is the kth Fibonacci number.*

Proof Consider the sequence $\frac{0}{1}, \frac{1}{0}, \frac{1}{1}, \frac{2}{1}, \frac{3}{2}, \frac{5}{3}, \ldots, \frac{F_{k-1}}{F_{k-2}}, \frac{F_k}{F_{k-1}}, \frac{F_{k+1}}{F_k}$. Note that each successive triple of terms determines a triangulation, and that each successive pair of triples share two slopes. Hence, the latter boundary triangulation can be obtained by layering on the edge of the former that they do not share. It takes k steps, hence k layers, to move from the first triple to the last. □

Furthermore, continued layering in this fashion increases the distance between the latest edge slopes and the original edge slopes:

Lemma 4.5 *Let F_k be the kth Fibonacci number. Then,*

$$d_{\mathcal{F}}\left(\frac{F_{k+1}}{F_k}, \infty\right) = \lfloor k/2 \rfloor + 1$$

Proof We will give an inductive proof. It is easy to verify that the statement holds for $k = 0, 1, 2$, where $\frac{F_{k+1}}{F_k} = \frac{1}{1}, \frac{2}{1}, \frac{3}{2}$, respectively, and the distances to $\infty = \frac{1}{0}$ are $1, 1, 2$, respectively. Let k be the least k for which the conclusion of the lemma does not hold. In the Poincaré disk, consider the triangle $\left(\frac{F_{k-1}}{F_{k-2}}, \frac{F_k}{F_{k-1}}, \frac{F_{k+1}}{F_k}\right)$ which is bounded by edges of the Farey Graph (see Fig. 6). This triangle separates the disk into 3 components.

First, we claim that the points $\frac{F_{k+1}}{F_k}$ and $\infty = \frac{1}{0}$ lie on opposite sides of the edge $\left(\frac{F_k}{F_{k-1}}, \frac{F_{k-1}}{F_{k-2}}\right)$. To see this, note that the point $\frac{F_{k-2}}{F_{k-3}}$ is the other corner of the second triangle that meets the edge $\left(\frac{F_k}{F_{k-1}}, \frac{F_{k-1}}{F_{k-2}}\right)$. The inductive hypothesis implies

$d_{\mathcal{F}}\left(\frac{F_{k-2}}{F_{k-3}}, \infty\right) < d_{\mathcal{F}}\left(\frac{F_k}{F_{k-1}}, \infty\right)$, so the second triangle must lie on the same side of the edge $\left(\frac{F_k}{F_{k-1}}, \frac{F_{k-1}}{F_{k-2}}\right)$ as ∞, hence the point $\frac{F_{k+1}}{F_k}$, lies on the other side.

Now, take a minimal path in the Farey Graph joining ∞ to $\frac{F_{k-1}}{F_{k-2}}$. By adjoining the edge $\left(\frac{F_{k-1}}{F_{k-2}}, \frac{F_{k+1}}{F_k}\right)$ to that path, we obtain a path from ∞ to $\frac{F_{k+1}}{F_k}$. It follows that $d_{\mathcal{F}}\left(\frac{F_{k+1}}{F_k}, \infty\right) \leq d_{\mathcal{F}}\left(\frac{F_{k-1}}{F_{k-2}}, \infty\right) + 1$.

Now, take a minimal path from ∞ to $\frac{F_{k+1}}{F_k}$. Because ∞ and $\frac{F_k}{F_{k-1}}$ lie on opposite sides of the edge $\left(\frac{F_{k-1}}{F_{k-2}}, \frac{F_k}{F_{k-1}}\right)$, this minimal path must pass through either the point $\frac{F_{k-1}}{F_{k-2}}$ or the point $\frac{F_k}{F_{k-1}}$. It follows that

$$\begin{aligned} d_{\mathcal{F}}\left(\frac{F_{k+1}}{F_k}, \infty\right) &\geq \min\left\{d_{\mathcal{F}}\left(\frac{F_{k-1}}{F_{k-2}}, \infty\right) + 1, d_{\mathcal{F}}\left(\frac{F_k}{F_{k-1}}, \infty\right) + 1\right\} \\ &= d_{\mathcal{F}}\left(\frac{F_{k-1}}{F_{k-2}}, \infty\right) + 1. \end{aligned}$$

Thus, $d_{\mathcal{F}}\left(\frac{F_{k+1}}{F_k}, \infty\right) = d_{\mathcal{F}}\left(\frac{F_{k-1}}{F_{k-2}}, \infty\right) + 1$ and the desired result follows. □

Lemma 4.6 *Let X be a (possibly disconnected) 3-manifold given via a triangulation that has a single vertex in each of two torus boundary components, T_0 and T_1. If $\alpha_0 \subset T_0$ and $\alpha_1 \subset T_1$ are slopes and $D \in \mathbb{N}$, then there is a triangulated manifold X' obtained from X by gluing T_0 to T_1 so that*

- *$d_{\mathcal{F}}(\alpha_0, \alpha_1) > D$, where distance is measured in the common image of T_0 and T_1 in X', and*
- *$t(X') = t(X) + 2D$, where $t(\cdot)$ is number of tetrahedra.*

Proof Fix an orientation on X and assume that the $T_i, i = 0, 1$, have the induced boundary orientation. For each $i = 0, 1$, we may choose a basis, $(0, \infty)$, for the homology of the boundary torus T_i so that the edges of the one-vertex triangulation have slopes $(0, \infty, 1)$, the basis $(0, \infty)$ induces the boundary orientation, and α_i has non-positive slope, $\alpha_i \leq 0$.

Applying Lemma 4.4, layer $2D$ tetrahedra on the boundary component T_0 so that the resulting triangulation has edges with slopes $\left(\frac{F_{2D-1}}{F_{2D-2}}, \frac{F_{2D}}{F_{2D-1}}, \frac{F_{2D+1}}{F_{2D}}\right)$.

Now, let X' be the manifold obtained by gluing the boundary triangulations together via an orientation reversing map that identifies the edge with slope $\frac{F_{2D+1}}{F_{2D}}$ in T_0 with the edge with slope 0 in T_1. This identifies the pair of edges with slopes $\left(\frac{F_{2D-1}}{F_{2D-2}}, \frac{F_{2D}}{F_{2D-1}}\right)$ in T_0, with the pair of edges with slopes $(1, \infty)$ in T_1, or its reverse. Note that the edge $\left(\frac{F_{2D-1}}{F_{2D-2}}, \frac{F_{2D}}{F_{2D-1}}\right)$ in the Farey graph for T_0 separates ∞ and the image of α_1.

Now compute the distance in the original basis for T_0 using Lemma 4.5. We have distance $d_{\mathcal{F}}(\alpha_0, \alpha_1) > d_{\mathcal{F}}\left(\infty, \frac{F_{2D-1}}{F_{2D-2}}\right) = \lfloor \frac{2D-2}{2} \rfloor + 1 = D$, as claimed. □

4.3 Blocks from Links

In this section we construct the required block manifolds. In each case, we prescribe a set of bipartitions of boundary components and then construct a manifold whose minimal genus Heegaard surfaces induce precisely that set of bipartitions of boundary components. All of our examples are Heegaard genus two. Three of the four are realized as the *exterior* of a knot or link in S^3, that is, each manifold is homeomorphic to $X(L) = S^3 - N(L)$ where L is a knot or link in S^3 and $N(\cdot)$ denotes an open regular neighborhood. The boundary of each manifold is a union of tori, and we often abuse notation by referring to components of the link, rather than to their corresponding boundary components. The fourth block manifold is obtained by Dehn filling on a torus boundary component of the third block manifold. Many of the results in this section are not new, and are collected for the sake of specificity.

For VAR blocks and the END block we need a genus two manifold with a single incompressible torus boundary component. The exterior of any tunnel number one knot will do, we choose a simple one:

Lemma 4.7 (VAR, END) *Let $K \subset S^3$ be the trefoil knot (see Fig. 8) and $X(K) = S^3 - N(K)$ be its exterior. Then $X(K)$ has Heegaard genus two.*

Proof It is well known that K is tunnel number one (genus two), see e.g. [16]. □

For OR blocks, we want a manifold whose minimal genus Heegaard surfaces realize every non-trivial bipartition of its three boundary components. The simplest such manifold seems to be the exterior of the three component chain, whose irreducible, and even non-irreducible, Heegaard splittings are quite well understood [24, 35]. Note that it is impossible for a genus two Heegaard surface to trivially bipartition the boundary components, $\{T_0, T_1, T_2|\emptyset\}$, as a genus two compression body V cannot have three torus boundary components in $\partial_- V$.

Lemma 4.8 (OR) *Let $C \subset S^3$ be the three component chain (see Fig. 8), and $X(C) = S^3 - N(C)$ its exterior. Then,*

(1) *$X(C)$ has Heegaard genus two,*
(2) *every non-trivial bipartition $\{T_i, T_j|T_k\}$ of the three boundary components of $\partial X(C)$ is induced by a genus two Heegaard surface for $X(C)$.*

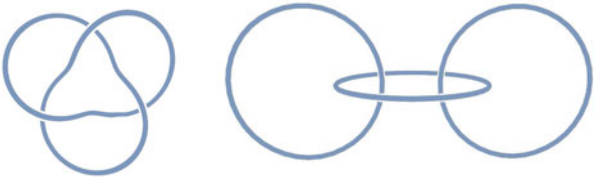


Fig. 8 Trefoil knot and three link chain

Proof Again, these facts are well known: it is easy to see that for each pair of link components, there is a handle and a short arc connecting them that induces a genus two Heegaard splitting that separates the pair from the other link component. □

For AND and REP blocks, we want a manifold whose minimal genus Heegaard surfaces all prefer the same bipartition of its three boundary components. This is a bit more challenging. Fortunately, Morimoto, Sakuma and Yokota showed that certain twisted torus knots are not 1-bridge with respect to an unknotted torus in S^3, providing the basis for the following.

Lemma 4.9 (AND, REP) *Let $L \subset S^3$ be the link indicated in Fig. 9. It is the union of the twisted torus knot $T(7, 17, 6)$ along with two unknotted components U_0 and U_1. Let $X(L)$ be its exterior. Then,*

(1) *$X(L)$ has Heegaard genus two,*
(2) *any genus two Heegaard splitting of $X(L)$ induces the same bipartition of boundary components, that is $\{U_0, U_1 | T(7, 17, 6)\}$,*
(3) *$X(L)$ does not contain a Möbius band with its boundary contained on the knotted boundary component.*

Note that conclusion (3) is not needed for the AND or REP blocks themselves. Rather, it is technical condition used for the construction of the NOT block via Lemma 4.10, which follows.

Proof (1) It is well known [27] and easy to see that a short arc joining the pair of twisted strands is a tunnel system for $T(7, 17, 6)$. The strands can be untwisted by sliding them over the tunnel, after which the tunnel appears to be the "middle tunnel" [26] for the torus knot $T(7, 17)$. Moreover, this gives a genus two splitting of the entire link as the indicated unknots U_0 and U_1 are cores for the complementary handlebody. Note that this genus two splitting induces the bipartition

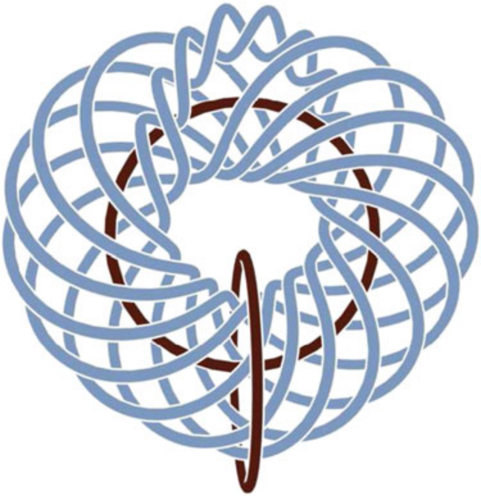


Fig. 9 Link with three components: $T(7, 17, 6)$ and two unknots U_0 and U_1

$\{T(7, 17, 6)|U_0, U_1\}$ of the boundary components. This is also a minimal genus splitting as no exterior of a link with 3 components has genus one.

(2) Suppose that a genus two Heegaard splitting induces a bipartition that isolates one of the two unknotted components, $\{U_i|U_j, T(7, 17, 6)\}$, for some $i \neq j$. In particular, this implies that the link $T(7, 17, 6) \cup U_j$ is tunnel number one. Lemma 4.13 of [26] states that any knot whose union with some unknot is a tunnel number one link must be $(1, 1)$. That is, it has a 1-bridge presentation with respect to an unknotted torus. However this is a contradiction, as Morimoto, Sakuma and Yokota [27] demonstrated that the knot $T(7, 17, 6)$ is not $(1, 1)$. It follows that any genus two Heegaard splitting of $X(L)$ induces the bipartition $\{U_0, U_1|T(7, 17, 6)\}$.

(3) Note that the exterior of the link $U_0 \cup U_1$ is a product, $T^2 \times [-1, 1]$. Draw the $(7, 17)$ torus knot as a curve on the level surface $T^2 \times \{0\}$ in this product. Choose two strands of the torus knot and give them 6 half twists to obtain the twisted torus knot $T(7, 17, 6)$. Its union with the pair of unknots is our twisted torus link L.

Now, note that the $(2, 5)$ curve drawn on the same level torus meets the $(7, 17)$ curve in a single point. Then the product $(2, 5) \times [-1, 1]$ is a properly embedded annulus in the product that meets the torus knot once, and the unknots in slopes $\frac{2}{5}$ and $\frac{5}{2}$, respectively. Moreover, the twisting needed to construct $T(7, 17, 6)$ can be performed in the complement of this annulus. Drill out the twisted torus knot. The annulus is punctured once (with slope $\infty = \frac{1}{0}$ on the knot) and becomes an essential pair of pants P in the link exterior.

Let $B \subset X(L)$ be a properly embedded Möbius band with its boundary in the knotted component and that meets P in the minimal number of components. Because both surfaces are essential, the intersection consists of a collection of arcs that are essential in both surfaces.

In fact, there is only a single arc of intersection: if there were two or more, then there would be a pair of arcs that are parallel and adjacent on P and that are also parallel on B. Then the union $B' = R_P \cup R_B$, where R_P and R_B are the rectangles the arcs bound in P and R, respectively, is a Möbius band (see for example [28]) that can be isotoped to meet P in a single arc.

However, it is also impossible for $P \cap B$ to consist of a single arc: this implies that the Möbius band has slope $\frac{n}{2}$ for some n as it meets the meridian $\frac{1}{0}$ twice. But, any $\frac{n}{2}$ curve also bounds a Möbius band in the solid torus that is attached to perform the meridional (S^3) filling on the knotted component. The union of the B and the Möbius band in the solid torus is a Klein bottle embedded in S^3, a contradiction. □

Finally, for NOT blocks we want a manifold for which no minimal genus Heegaard surface splits its two boundary components. Note that $X(L)$ is almost what we want; no minimal Heegaard surface splits the two unknotted boundary components. Nonetheless, there is an inconvenient third boundary component (the knotted one). Can we get rid of it?

There are many results that demonstrate that after a "sufficiently large" Dehn filling, the filled manifold inherits the qualities of the unfilled manifold. Fortunately, that is also true for Heegaard structure [22, 25, 28–30] and that is precisely what we use here:

Lemma 4.10 (NOT) *Let $L \subset S^3$ be the link indicated in Fig. 9, and let $X(L;\gamma)$ be the manifold obtained by Dehn filling the knotted component along the slope γ. If $d_{\mathcal{F}}(\gamma,\infty) > 10$, where $d_{\mathcal{F}}$ is the distance in the Farey graph, then*

(1) *$X(L;\gamma)$ has Heegaard genus two,*
(2) *every genus two Heegaard splitting of $X(L;\gamma)$ induces the trivial boundary bipartition $\{U_0, U_1|\emptyset\}$.*

Proof Heegaard surfaces survive Dehn fillings. That is, after filling any slope γ, a Heegaard surface for $X(L)$ is also a Heegaard surface for $X(L;\gamma)$. Thus the genus of $X(L;\gamma)$ is at most 2.

We now show that under the hypothesis $d_{\mathcal{F}}(\gamma,\infty) > 10$, every genus two Heegaard splitting of $X(L;\gamma)$ is isotopic (in $X(L;\gamma)$) to a Heegaard splitting of $X(L)$. It will follow that the genus of $X(L;\gamma)$ is exactly two, and any genus two splitting induces the desired bipartition of boundary components.

We will say that a filled manifold $X(L;\alpha)$ has a *new* Heegaard surface if there is a Heegaard surface $\Sigma \subset X(L;\alpha)$ for the filled manifold that is *not* isotopic in $X(L;\alpha)$ to a Heegaard surface for $X(L)$. Rieck and Sedgwick [30] have shown that there are two possibilities for a new Heegaard surface Σ, depending on whether the core of the attached solid torus is isotopic into Σ in the filled manifold. In either case, we can find a useful derived surface $\Sigma' \subset X(L)$ by isotoping Σ in $X(L;\alpha)$ and then drilling out the core: if the attached core is not isotopic into Σ, then Σ is isotopic to a "thick level" in some thin presentation of the core, which is a knot in $X(L;\alpha)$. After drilling out the core, we obtain a properly embedded surface $\Sigma \subset X(L)$ that meets the knotted boundary component in curves of slope α. If the core *is* isotopic into Σ, then drilling out the core and possibly compressing, we obtain a properly embedded essential surface $\Sigma' \subset X(L)$. Its genus is at most that of Σ and its boundary curves meet the knotted boundary component in a slope α', where $d_{\mathcal{F}}(\alpha',\alpha) = 1$.

If two different filled manifolds $X(L;\alpha)$ and $X(L;\beta)$ have new Heegaard surfaces, then the pair of bounded surfaces derived above, each either essential or "thick," can be isotoped to intersect essentially [8, 28]. Moreover, the previous lemma shows that there is no Möbius band in $X(L)$ with its boundary in the knotted component. In that case Rieck showed that the number of intersections between the slopes α and β is bounded by a quadratic function, $36g_1g_2 + 36g_1 + 18g_2 + 18$, where g_1 and g_2, $g_1 \geq g_2$, are the genera of the derived surfaces ([28] Theorem 5.2). (Theorem 5.2 is stated with a stronger hypothesis, that $X(L)$ is a-cylindrical, but the proof clearly states that either the bound holds or there is a Möbius band meeting the boundary component that was filled.)

Now, we know that the manifold $X(L,\infty)$ is the product $T^2 \times [-1, 1]$ and thus has a new Heegaard surface of genus 1. (As the knotted component is not a torus knot, in this case the derived surface is a thick level with genus 1 and slope ∞.)

Suppose then that $X(L,\gamma)$ has a new Heegaard surface of genus at most 2. Then the slopes of the derived surfaces intersect at most 180 times (applying the above quadratic function with $g_1 = 2 \geq g_2 = 1$) and thus have distance in the Farey graph $d_{\mathcal{F}} \leq \log_2 180 + 1 < 9$. As the derived surface in $X(L, \gamma)$ has distance 0 or 1 from γ, we have $d_{\mathcal{F}}(\gamma, \infty) < 10$, a contradiction.

It follows that $X(L, \gamma)$ has no new Heegaard surfaces with genus at most 2. Then the genus of $X(L, \gamma)$ is 2. Moreover, every genus two Heegaard surface of $X(L, \gamma)$ is isotopic in $X(L, \gamma)$ to a Heegaard surface for $X(L)$, and in particular induces the boundary bipartition $\{U_0, U_1 | \emptyset\}$. This completes the proof. □

Construct the NOT blocks by using Lemmas 4.6 and 4.10 to glue the triangulated twisted torus link exterior to a one-tetrahedron solid torus (see for example, [13]) so that μ, the curve bounding a meridional disk of the solid torus, and ∞ the meridian of the twisted torus link, satisfy $d_{\mathcal{F}}(\mu, \infty) > 11$.

4.4 *Proof of Proposition* 4.1

Proof The manifold M_Q is obtained by gluing a collection of blocks along pairs of torus boundary components via high distance maps. There is exactly one block for each term (VAR, AND, OR, NOT) in Q, plus the END block, for a total of $|Q| + 1$ blocks.

As a preprocessing step, we triangulate each of the block types so that each torus boundary component has a one-vertex triangulation. For each of the three link exteriors, use the method Weeks describes in [38] and implements in his *SnapPea* program, to convert the link diagrams given by Figs. 8 and 9 to ideal triangulations of the link exteriors. Then construct a (non-ideal) triangulation by subdividing and deleting tetrahedra meeting the ideal vertex. Use Jaco and Rubinstein's method to convert this triangulation to a 0-efficient triangulation [11], which has the desired property that it restricts to a one-vertex triangulation of each torus boundary component. For each torus boundary component of each block, use normal surface theory to identify, among essential surfaces meeting the boundary component, a surface maximizing Euler characteristic.

Let T be the maximal number of tetrahedra used by one of the four triangulated blocks types. Since there are $|Q| + 1$ blocks, we thus require at most $T(|Q| + 1)$ tetrahedra before gluing.

There is a computable constant K, depending only on the homeomorphism types of the blocks, so that if any set of blocks are glued with maps of distance at least Kg (relative to the boundaries), then any Heegaard surface whose genus is at most g is an amalgamation of splittings of the blocks. (The proof of this is given in the appendix; distance is measured between the surfaces chosen above.) As we want to guarantee that any splitting of genus at most $|Q| + 2$ is an amalgamation, it is thus sufficient to glue each pair of blocks with a map of distance $K(|Q| + 2)$, which by Lemma 4.6 requires $2K(|Q| + 2)$ tetrahedra per gluing. Since each of the $|Q| + 1$ blocks has at most 3 boundary components, there are at most $\frac{3}{2}(|Q| + 1)$ pairs of boundary components to glue. We conclude that we need at most $\frac{3}{2}(|Q| + 1)2K(|Q| + 2)$ tetrahedra to glue the blocks.

The total number of tetrahedra required to construct M_Q is then the sum of those for the blocks and those for gluings,

$$t(M_Q) \leq T(|Q| + 1) + 3K(|Q| + 1)(|Q| + 2)$$

which is clearly quadratically bounded in $|Q|$. □

5 Open Questions

We now discuss some questions that remain. The most obvious is:

Question 5.1 Is HEEGAARD GENUS $\leq g$ in NP?

Next, since the 3-sphere is, by definition, the 3-manifold with genus 0, 3-SPHERE RECOGNITION is precisely HEEGAARD GENUS ≤ 0, i.e., a special case of our general problem with fixed parameter $g = 0$. Schleimer showed that 3-SPHERE RECOGNITION is in NP [34]. And, using Kuperberg's work [17], Zentner showed that 3-SPHERE RECOGNITION is also in co-NP if we assume that the Generalized Riemann Hypothesis is true [39]. Thus, without disproving a major conjecture, we do not expect the special case HEEGAARD GENUS ≤ 0 to be NP-hard. Since Heegaard genus is such an important invariant, it is worth asking about the complexity of the problem for other small fixed values of g, in particular $g \leq 2$:

Question 5.2 What is the computational complexity of deciding HEEGAARD GENUS ≤ 1 and HEEGAARD GENUS ≤ 2?

Finally, note that our construction produces non-hyperbolic manifolds because the identified torus boundary components are incompressible after gluing. It seems probable that hyperbolic examples can be constructed by gluing together hyperbolic block manifolds that have higher genus boundary components. But, the resulting manifolds would most definitely be *Haken* (have embedded incompressible surfaces). Do embedded essential surfaces explain NP-hardness or,

Question 5.3 Is HEEGAARD GENUS $\leq g$ NP-hard when restricted to the class of non-Haken manifolds?

Appendix: Sufficiently Complicated Amalgamations

In this section we provide a proof of the following proposition, based on several well-known results.

Proposition 1 *There is a computable constant K, depending only on the homeomorphism types of the blocks, so that if any set of blocks are glued with maps of distance at least Kg (in the sense of Theorem 2 below), then any Heegaard surface whose genus is at most g is an amalgamation of splittings of the blocks.*

Proof Suppose H is a minimal genus Heegaard splitting of M_Q. It follows from the results of [33] that there is a DAG Γ such that H is an amalgamation of some generalized Heegaard splitting $\bigcup_{x \in \Gamma} M_x$ of M_Q, such that for each $x \in \Gamma$, $V_x \cap W_x$ is *strongly irreducible* in M_x, and for each $x \neq y$, $V_x \cap W_y$ is a (possibly empty) incompressible surface in M. In the parlance of [2], both kinds of surfaces are *topologically minimal* in M. Let $\mathcal{H}$ denote the union of all such topologically minimal surfaces.

For each boundary component F of each block used in the original construction of M_Q (see Sect. 3), choose a maximal Euler characteristic, properly embedded, incompressible, boundary incompressible surface in that block that is incident to F. Let $\mathcal{S}$ be the collection of these chosen surfaces. (Note that the surfaces in $\mathcal{S}$ need not be disjoint in each block).

Let M_- and M_+ denote blocks used in the construction of M_Q, such that $M_+ \cap M_- \neq \emptyset$. Let F be a component of $M_+ \cap M_-$. Then F can be identified with boundary components $F_- \subset \partial M_-$ and $F_+ \subset \partial M_+$. Let $\phi : F_- \to F_+$ denote the gluing map used to attach M_- to M_+ along F in the construction of M_Q. Let M_ϕ denote the manifold obtained from M_- and M_+ by gluing F_- to F_+ via the map ϕ. Note that M_ϕ may be different from $M_- \cup M_+$, as the latter manifold may be obtained from M_- and M_+ by gluing along multiple surfaces. However, if $\mathcal{F}$ denotes the collection of surfaces at the interfaces between all blocks in M_Q, then M_ϕ can be identified with a component of the complement of $\mathcal{F} \setminus F$.

By [4], we can isotope each surface in $\mathcal{H}$ so that it meets the complementary pieces of $\mathcal{F} \setminus F$ in a collection of surfaces that are topologically minimal (in particular, either incompressible or strongly irreducible). After such an isotopy, let H' denote a component of the intersection of such a surface with M_ϕ.

The first author, building on work of Tao Li [19], proved the following theorem, restated here with notation consistent with that of the present paper:

Theorem 2 (cf. [3], Theorem 5.4.) *Let S_- and S_+ denote the surfaces in $\mathcal{S}$ chosen to meet F_- and F_+ in M_- and M_+. Let $K = 24(1 - 3\chi(S_-) - 3\chi(S_+))$. If*

$$d(\phi(S_- \cap F_-), S_+ \cap F_+) \geq K \cdot \ genus(H)$$

then H' can be isotoped to be disjoint from F in M_ϕ.[1]

Note that H' is a component of $\mathcal{H} \cap M_\phi$. Applying this Theorem to every such component (noting that genus$(H') \leq$ genus (H)), we conclude $\mathcal{H}$ can be isotoped to be disjoint from F in M_Q. Each surface in the resulting collection is now topologically minimal in $M_Q - F$. Repeating this argument for every surface in $\mathcal{F}$ shows that every surface in $\mathcal{H}$ can be isotoped entirely into some block. It then follows from standard arguments that each surface of $\mathcal{F}$ can be identified with a component of ∂M_x, for some $x \in \Gamma$. Thus, for each block B in M_Q, there is a collection of vertices $\mathcal{V}$ of Γ such that $B = \bigcup_{x \in \mathcal{V}} M_x$. Amalgamating this

[1] The original theorem is stated so that H' is a closed surface, but this assumption is never used in the proof.

generalized Heegaard splitting of B then produces a Heegaard splitting of B. Our original Heegaard surface H is then an amalgamation of these Heegaard surfaces of the blocks. □

References

1. I. Agol, J. Hass, W. Thurston, 3-manifold knot genus is NP-complete, in *Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing* (ACM, New York, 2002), pp. 761–766
2. D. Bachman, Topological index theory for surfaces in 3-manifolds. Geom. Topol. **14**(1), 585–609 (2010)
3. D. Bachman, Stabilizing and destabilizing Heegaard splittings of sufficiently complicated 3-manifolds. Math. Ann. **355**(2), 697–728 (2013)
4. D. Bachman, S. Schleimer, E. Sedgwick, Sweepouts of amalgamated 3-manifolds. Algebr. Geom. Topol. **6**, 171–194 (2006) (electronic)
5. B.A. Burton, J. Spreer, The complexity of detecting taut angle structures on triangulations, in *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013* (New Orleans, 2013), 6–8 Jan 2013, pp. 68–183
6. B.A. Burton, É.C. de Verdière, A. de Mesmay, On the Complexity of Immersed Normal Surfaces. arXiv:1412.4988, Dec 2014, Preprint
7. B.A. Burton, A. de Mesmay, U. Wagner, Finding non-orientable surfaces in 3-manifolds. arXiv:0901.0208, Feb 2016, Preprint
8. D. Gabai, Foliations and the topology of 3-manifolds. III. J. Differ. Geom. **26**(3), 479–536 (1987)
9. W. Haken, Theorie der Normalflächen. Acta Math. **105**, 245–375 (1961)
10. J. Hass, J.C. Lagarias, N. Pippenger, The computational complexity of knot and link problems. J. ACM **46**(2), 185–211 (1999)
11. W. Jaco, J. Hyam Rubinstein, 0-efficient triangulations of 3-manifolds. J. Differ. Geom. **65**(1), 61–168 (2003)
12. W. Jaco, J. Hyam Rubinstein, Layered-triangulations of 3-manifolds. arXiv:math/0603601, Mar 2006, Preprint
13. W. Jaco, E. Sedgwick, Decision problems in the space of Dehn fillings. Topology **42**(4), 845–906 (2003)
14. K. Johannson, Heegaard surfaces in Haken 3-manifolds. Bull. Am. Math. Soc. (N.S.) **23**(1), 91–98 (1990)
15. K. Johannson, *Topology and Combinatorics of 3-Manifolds*. Lecture Notes in Mathematics, vol. 1599 (Springer, Berlin, 1995)
16. T. Kobayashi, Classification of unknotting tunnels for two bridge knots, in *Proceedings of the Kirbyfest*, Berkeley, 1998. Geology and Topology Monographs, vol. 2 (Geometry and Topology in collaboration Publication, Coventry, 1999), pp. 259–290 (electronic)
17. G. Kuperberg, Knottedness is in NP, modulo GRH. Adv. Math. **256**, 493–506 (2014)
18. M. Lackenby, Some conditionally hard problems on links and 3-manifolds. arXiv:1602.08427 (2016, Preprint)
19. T. Li, Heegaard surfaces and the distance of amalgamation. Geom. Topol. **14**(4), 1871–1919 (2010)
20. T. Li, An algorithm to determine the Heegaard genus of a 3-manifold. Geom. Topol. **15**(2), 1029–1106 (2011)
21. J. Matoušek, E. Sedgwick, M. Tancer, U. Wagner, Embeddability in the 3-sphere is decidable, in *Computational Geometry (SoCG'14)* (ACM, New York, 2014), pp. 78–84

22. Y. Moriah, H. Rubinstein, Heegaard structures of negatively curved 3-manifolds. Commun. Anal. Geom. **5**(3), 375–412 (1997)
23. Y. Moriah, J. Schultens, Irreducible Heegaard splittings of Seifert fibered spaces are either vertical or horizontal. Topology **37**(5), 1089–1112 (1998)
24. Y. Moriah, E. Sedgwick, Closed essential surfaces and weakly reducible Heegaard splittings in manifolds with boundary. J. Knot Theory Ramif. **13**(6), 829–843 (2004)
25. Y. Moriah, E. Sedgwick, The Heegaard structure of Dehn filled manifolds, in *Workshop on Heegaard Splittings*. Geology and Topology Monographs, vol. 12 (Geometry and Topology in collaboration Publication, Coventry, 2007), pp. 233–263
26. Y. Moriah, E. Sedgwick, Heegaard splittings of twisted torus knots. Topol. Appl. **156**(5), 885–896, (2009)
27. K. Morimoto, M. Sakuma, Y. Yokota, Examples of tunnel number one knots which have the property "$1 + 1 = 3$". Math. Proc. Camb. Philos. Soc. **119**(1), 113–118 (1996)
28. Y. Rieck, Heegaard structures of manifolds in the Dehn filling space. Topology **39**(3), 619–641 (2000)
29. Y. Rieck, E. Sedgwick, Finiteness results for Heegaard surfaces in surgered manifolds. Comm. Anal. Geom. **9**(2), 351–367 (2001)
30. Y. Rieck, E. Sedgwick, Persistence of Heegaard structures under Dehn filling. Topol. Appl. **109**(1), 41–53 (2001)
31. J.H. Rubinstein, An algorithm to recognize the 3-sphere, in *Proceedings of the International Congress of Mathematicians*, Zürich, 1994, vols. 1, 2 (Birkhäuser, Basel, 1995), pp. 601–611
32. M. Scharlemann, A. Thompson, Heegaard splittings of (surface) $\times I$ are standard. Math. Ann. **295**, 549–564 (1993)
33. M. Scharlemann, A. Thompson, Thin position for 3-manifolds, in *Geometric Topology (Haifa, 1992)*. Contemporary Mathematics, vol. 164 (American Mathematical Society, Providence, 1994), pp. 231–238
34. S. Schleimer, Sphere recognition lies in NP, in *Low-Dimensional and Symplectic Topology*. Proceedings of Symposia in Pure Mathematics, vol. 82 (American Mathematical Society, Providence, 2011), pp. 183–213
35. J. Schultens, The classification of Heegaard splittings for (compact orientable surface) $\times S^1$. Proc. Lond. Math. Soc. (3) **67**(2), 425–448 (1993)
36. J. Schultens, R. Weidmann, Destabilizing amalgamated Heegaard splittings. Geo. Topol. Monogr. **12**, 319–334 (2007)
37. A. Thompson, Thin position and the recognition problem for S^3. Math. Res. Lett. **1**(5), 613–630 (1994)
38. J. Weeks, Computation of hyperbolic structures in knot theory, in *Handbook of Knot Theory* (Elsevier B. V., Amsterdam, 2005), pp. 61–480
39. R. Zentner, Integer homology 3-spheres admit irreducible representations in SL(2,C). Preprint arXiv:1605.08530, May 2016

Approximation-Friendly Discrepancy Rounding

Nikhil Bansal and Viswanath Nagarajan

Abstract Rounding linear programs using techniques from discrepancy is a recent approach that has been very successful in certain settings. However this method also has some limitations when compared to approaches such as randomized and iterative rounding. We provide an extension of the discrepancy-based rounding algorithm due to Lovett–Meka that (i) combines the advantages of both randomized and iterated rounding, (ii) makes it applicable to settings with more general combinatorial structure such as matroids. As applications of this approach, we obtain new results for various classical problems such as linear system rounding, degree-bounded matroid basis and low congestion routing.

1 Introduction

A very common approach for solving discrete optimization problems is to solve some linear programming relaxation, and then round the fractional solution into an integral one, without (hopefully) incurring much loss in quality. Over the years several ingenious rounding techniques have been developed (see e.g. [24, 25]) based on ideas from optimization, probability, geometry, algebra and various other areas. Randomized rounding and iterative rounding are two of the most commonly used methods.

Recently, discrepancy-based rounding approaches have also been very successful; a particularly notable result is for bin packing due to Rothvoss [18]. Discrepancy is a well-studied area in combinatorics with several surprising results (see e.g. [16]), and as observed by Lovász et al. [14], has a natural connection to rounding. However, until the recent algorithmic developments [1, 9, 15, 17, 19], most of the results in discrepancy were non-constructive and hence not directly useful

N. Bansal (✉)
Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, Netherlands
e-mail: n.bansal@tue.nl

V. Nagarajan
Department of Industrial and Operations Engineering, University of Michigan, 48109 Ann Arbor, MI, USA

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_4

for rounding. These algorithmic approaches combine probabilistic approaches like randomized rounding with linear algebraic approaches such as iterated rounding [12], which makes them quite powerful.

Interestingly, given the connection between discrepancy and rounding, these discrepancy algorithms can in fact be viewed as meta-algorithms for rounding. We discuss this in Sect. 1.1 in the context of the Lovett–Meka (LM) algorithm [15]. This suggests the possibility of one single approach that generalizes both randomized and iterated rounding. This is our motivating goal in this paper.

While the LM algorithm is already an important step in this direction, it still has some important limitations. For example, it is designed for obtaining additive error bounds and it does not give good multiplicative error bounds (like those given by randomized rounding). This is not an issue for discrepancy applications, but crucial for many approximation algorithms. Similarly, iterated rounding can work well with exponentially sized LPs by exploiting their underlying combinatorial structure (e.g., degree-bounded spanning tree [21]), but the current discrepancy results [15, 19] give extremely weak bounds in such settings.

Our Results: We extend the LM algorithm to overcome the limitations stated above. In particular, we give a new variant that also gives Chernoff type multiplicative error bounds (sometimes with an additional logarithmic factor loss). We also show how to adapt the above algorithm to handle exponentially large LPs involving matroid constraints, as in iterated rounding.

This new discrepancy-based algorithm gives new results for problems such as linear system rounding with violations [4, 13], degree-bounded matroid basis [6, 11], low congestion routing [10, 13] and multi-budgeted matroid basis [8], These results simultaneously combine non-trivial guarantees from discrepancy, randomized rounding and iterated rounding and previously such bounds were not even known existentially.

Our results are described formally in Sect. 1.2. To place them in the proper context, we first need to describe some existing rounding approaches (Sect. 1.1). The reader familiar with the LM algorithm can directly go to Sect. 1.2.

1.1 Preliminaries

We begin by describing LM rounding [15], randomized rounding and iterated rounding in a similar form, and then discuss their strengths and weaknesses.

LM Rounding: Let A be a $m \times n$ matrix with 0–1 entries,[1] $x \in [0, 1]^n$ a fractional vector and let $b = Ax$. Lovett and Meka [15] showed the following rounding result.

[1] The results below generalize to arbitrary real matrices A and vectors x in natural ways, but we consider 0–1 case for simplicity.

Theorem 1 (LM Rounding [15]) *Given A and x as above, For $j = 1, \ldots, m$, pick any λ_j satisfying*

$$\sum_j \exp(-\lambda_j^2/4) \leq n/16. \tag{1}$$

Then there is an efficient randomized algorithm to find a solution x' such that: (i) at most $n/2$ variables of x' are fractional (strictly between 0 and 1) and, (ii) $|\langle a_j, x' - x\rangle| \leq \lambda_j \|a_j\|_2$ for each $j = 1, \ldots, m$, where a_j denotes the j-th row of A.

Remark The right hand side of (1) can be set to $(1 - \epsilon)n$ for any fixed constant $\epsilon > 0$, at the expense of $O_\epsilon(1)$ factor loss in other parameters of the theorem; see e.g. [2].

Randomized Rounding: Chernoff bounds state that if $X_1, \ldots, X_n$ are independent Bernoulli random variables, and $X = \sum_i X_i$ and $\mu = \mathbb{E}[X]$, then

$$\Pr[|X - \mu| \geq \epsilon\mu] \leq 2\exp(-\epsilon^2\mu/4) \qquad \text{for } \epsilon \leq 1.$$

Then independent randomized rounding can be viewed as the following (by using Chernoff bounds and union bound, and denoting $\lambda_j = \epsilon_j\sqrt{b_j}$).

Theorem 2 (Randomized Rounding) *For $j = 1, \ldots, m$, pick any λ_j satisfying $\lambda_j \leq \sqrt{b_j}$, and*

$$\sum_j \exp(-\lambda_j^2/4) < 0.5 \tag{2}$$

Then independent randomized rounding gives a solution x' such that: (i) All variables are 0–1, and (ii) $|\langle a_j, x' - x\rangle| \leq \lambda_j\sqrt{b_j}$ for each $j = 1, \ldots, m$.

Iterated Rounding [12]: This is based on the following linear-algebraic fact.

Theorem 3 *If $m < n$, then there is a solution $x' \in [0, 1]^n$ such that (i) x' has at least $n - m$ variables set to 0 or 1 and, (ii) $A(x' - x) = 0$ (i.e., $b = Ax'$).*
In iterated rounding applications, if $m > n$ then some cleverly chosen constraints are dropped until $m < n$ and integral variables are obtained. This is done repeatedly.

Strengths of LM rounding: Note that if we set $\lambda_j \in \{0, \infty\}$ in LM rounding, then it gives a very similar statement to Theorem 3. E.g., if we only care about some $m = n/2$ constraints then Theorem 3 gives an x' with at least $n/2$ integral variables and $a_j x = a_j x'$ for all these m constraints. Theorem 1 (and the remark below it) give the same guarantee if we set $\lambda_j = 0$ for all constraints. In general, LM rounding can be much more flexible as it allows arbitrary λ_j.

Second, LM rounding is also related to randomized rounding. Note that (2) and (1) have the same left-hand-side. However, the right-hand-side of (1) is $\Omega(n)$, while that of (2) is $O(1)$. This actually makes a huge difference. In particular, in (2) one cannot set $\lambda_j = 1$ for more than a couple of constraints (to get an $o(\sqrt{b_j})$ error bound on constraints), while in (1), one can even set $\lambda_j = 0$ for $O(n)$ constraints. In fact, almost all non-trivial results in discrepancy [16, 22, 23] are based on this ability.

Weaknesses of LM rounding: First, Theorem 1 only gives a partially integral solution instead of a fully integral one as in Theorem 2.

Second, and more importantly, it only gives additive error bounds instead of multiplicative ones. In particular, note the $\lambda_j\|a_j\|_2$ vs $\lambda_j\sqrt{b_j}$ error in Theorems 1 and 2. E.g., for a constraint $\sum_i x_i = \log n$, Theorem 2 gives $\lambda\sqrt{\log n}$ error but Theorem 1 gives a much higher $\lambda\sqrt{n}$ error. So, while randomized rounding can give a good multiplicative error like $a_j x' \leq (1 \pm \epsilon_j)b_j$, LM rounding is completely insensitive to b_j.

Finally, iterated rounding works extremely well in many settings where Theorem 1 does not give anything useful. E.g., in problems involving exponentially many constraints such as the degree bounded spanning tree problem. The problem is that if m is exponentially large, then the λ_j's in Theorem 1 need to be very large to satisfy (2).

1.2 Our Results and Techniques

Our first result is the following improvement over Theorem 1.

Theorem 4 *There is a constant $K_0 > 0$ and randomized polynomial time algorithm that given any $n > K_0$, fractional solution $y \in [0,1]^n$, $m \leq 2^n$ linear constraints $a_1, \ldots, a_m \in \mathbb{R}^n$ and $\lambda_1, \cdots, \lambda_m \geq 0$ with $\sum_{j=1}^m e^{-\lambda_j^2/K_0} < \frac{n}{16}$, finds a solution $y' \in [0,1]^n$ such that:*

$$|\langle y' - y, a_j\rangle| \leq \lambda_j \cdot \sqrt{W_j(y)} + \frac{1}{n^2}\cdot\|a_j\|, \quad \forall j = 1, \cdots m \tag{3}$$

$$y'_i \in \{0,1\}, \quad \textit{for } \Omega(n) \textit{ indices } i \in \{1, \cdots, n\} \tag{4}$$

Here $W_j(y) := \sum_{i=1}^n a_{ji}^2 \cdot \min\{y_i, 1 - y_i\}^2$ for each $j = 1, \cdots m$.

Remarks

(1) The error $\lambda_j\sqrt{W_j(y)}$ is always smaller than $\lambda_j\|a_j\|$ in LM-rounding and $\lambda_j(\sum_{i=1}^n a_{ji}^2 \cdot y_i(1 - y_i))^{1/2}$ in randomized rounding. In fact it could even be much less if the y_i are very close to 0 or 1.

(2) The term $n/16$ above can be made $(1 - \epsilon)n$ for any fixed constant $\epsilon > 0$, at the expense of worsening other constants (just as in LM rounding).

(3) The additional error term $\frac{1}{n^2}\cdot\|a_j\|$ above is negligible and can be reduced to $\frac{1}{n^c}\cdot\|a_j\|$ for any constant c, at the expense of a larger running time $n^{O(c)}$.

We note that Theorem 4 can also be obtained in a "black box" manner from LM-rounding (Theorem 1) by rescaling the polytope and using its symmetry.[2] However, such an approach does not work in the setting of matroid polytopes (Theorem 5). In the matroid case, we need to modify LM-rounding as outlined below.

Applications: We focus on *linear system rounding* as the prime example. Here, given matrix $A \in [0,1]^{m\times n}$ and vector $b \in \mathbb{Z}_+^m$, the goal is to find a vector $z \in \{0,1\}^n$ satisfying $Az = b$. As this is NP-hard, the focus has been on finding a $z \in \{0,1\}^n$ where $Az \approx b$.

Given any fractional solution $y \in [0,1]^n$ satisfying $Ay = b$, using Theorem 4 iteratively we can obtain an integral vector $z \in \{0,1\}^n$ with

$$|a_j z - b_j| \le \min\left\{O(\sqrt{n\log(2+m/n)})\,,\ \sqrt{L\cdot b_j}+L\right\}, \quad \forall j \in [m], \tag{5}$$

where $L = O(\log n \log m)$ and $[m] := \{1,2,\ldots,m\}$.[3] Previously known algorithms could provide a bound of either $O(\sqrt{n\log(m/n)})$ for all constraints [15] or $O(\sqrt{\log m}\cdot\sqrt{b_j}+\log m)$ for all constraints (Theorem 2). Note that this does not imply a $\min\{\sqrt{n\log(m/n)}, \sqrt{\log m}\cdot\sqrt{b_j}+\log m\}$ violation per constraint, as in general it is not possible to combine two integral solutions and achieve the better of their violation bounds on all constraints. To the best of our knowledge, even the existence of an integral solution satisfying the bounds in (5) was not known prior to our work.

In the setting where the matrix A is "column sparse", i.e. each variable appears in at most Δ constraints, we obtain a more refined error of

$$|a_j y - b_j| \le \min\left\{O(\sqrt{\Delta}\log n)\,,\ \sqrt{L\cdot b_j}+L\right\}, \quad \forall j \in [m], \tag{6}$$

where $L = O(\log n \cdot \log m)$. Previous algorithms could separately achieve bounds of $\Delta - 1$ [4], $O(\sqrt{\Delta}\log n)$ [15] or $O(\sqrt{\log\Delta}\cdot\sqrt{b_j}+\log\Delta)$ [13]. For clarity, Fig. 1 plots the violation bounds achieved by these different algorithms as a function of the right-hand-side b when $m = n$ (we assume $b, \Delta \ge \log^2 n$). Note again that since there are multiple constraints we can not simply combine algorithms to achieve the smaller of their violation bounds.

One can also combine the bounds in (5) and (6), and use some additional ideas from discrepancy to obtain:

$$|a_j y - b_j| \le O(1)\cdot\min\left\{\sqrt{j},\ \sqrt{n\log(2+\frac{m}{n})},\ \sqrt{L\cdot b_j}+L,\ \sqrt{\Delta}\log n\right\}, \quad \forall j \in [m]. \tag{7}$$

[2] We thank an anonymous reviewer for pointing this out.

[3] For any integer $t \ge 1$, we use the notation $[t] := \{1,2,\ldots,t\}$.

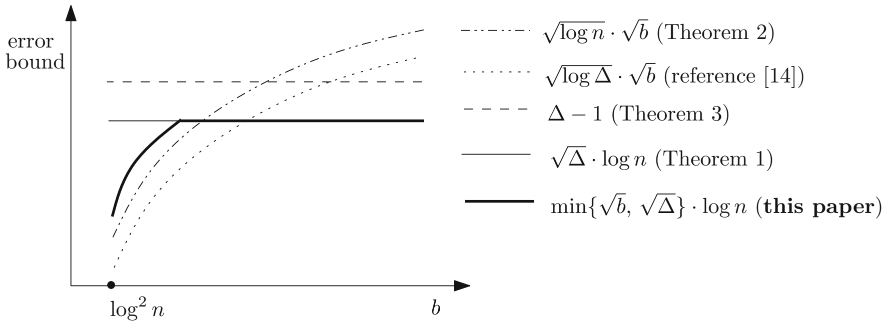


Fig. 1 Additive violation bounds for linear system rounding when $\Delta \geq \log^2 n$ and $b \geq \log^2 n$

Matroid Polytopes: Our main result is an extension of Theorem 4 where the fractional solution lies in a matroid polytope in addition to satisfying the linear constraints $\{a_j\}_{j=1}^m$. Recall that a matroid $\mathcal{M}$ is a tuple $(V, \mathcal{I})$ where V is the groundset of elements and $\mathcal{I} \subseteq 2^V$ is a collection of independent sets satisfying the hereditary and exchange properties [20]. The rank function $r : 2^V \to \mathbb{Z}$ of a matroid is defined as $r(S) = \max_{I \in \mathcal{I}, I \subseteq S} |I|$. The matroid polytope (i.e. convex hull of all independent sets) is given by the following linear inequalities:

$$P(\mathcal{M}) \quad := \quad \left\{ x \in \mathbb{R}^n \ : \ \sum_{i \in S} x_i \leq r(S) \ \forall S \subseteq V, \ \ x \geq 0 \right\}.$$

As is usual when dealing with matroids, we assume access to an "independent set oracle" for $\mathcal{M}$ that given any subset $S \subseteq V$ returns whether/not $S \in \mathcal{I}$ in polynomial time.

Theorem 5 *There is a randomized polynomial time algorithm that given matroid $\mathcal{M}$, fractional solution $y \in P(\mathcal{M})$, linear constraints $\{a_j \in \mathbb{R}^n\}_{j=1}^m$ and values $\{\lambda_j\}_{j=1}^m$ satisfying the conditions in Theorem 4, finds a solution $y' \in P(\mathcal{M})$ satisfying* (3) *and* (4).

We note that the same result can be obtained even if we want to compute a base (maximal independent set) in the matroid: the only difference here is to add the equality $\sum_{i \in V} x_i = r(V)$ to $P(\mathcal{M})$ which corresponds to the base polytope of $\mathcal{M}$.

The fact that we can *exactly* preserve matroid constraints leads to a number of improvements:

*Degree-bounded matroid basis (*DEGMAT*).* Given a matroid on elements $[n] := \{1, 2, \ldots, n\}$ with costs $d : [n] \to \mathbb{Z}_+$ and m "degree constraints" $\{S_j, b_j\}_{j=1}^m$ where each $S_j \subseteq [n]$ and $b_j \in \mathbb{Z}_+$, the goal is to find a minimum-cost basis I in the matroid that satisfies $|I \cap S_j| \leq b_j$ for all $j \in [m]$. Since even the feasibility problem is NP-hard, we consider *bicriteria* approximation algorithms that violate the degree bounds. We obtain an algorithm where the solution costs at most the optimal and the degree bound violation is as in (7); here Δ denotes the maximum number of sets $\{S_j\}_{j=1}^m$ containing any element.

Previous algorithms achieved approximation ratios of $(1, b + O(\sqrt{b \log n}))$ [6], based on randomized swap rounding, and $(1, b + \Delta - 1)$ [11] based on iterated rounding. Again, these bounds could not be combined together as they used different algorithms. We note that in general the $(1, b + O(\sqrt{n \log(m/n)}))$ approximation is the best possible (unless P=NP) for this problem [3, 5].

Multi-criteria matroid basis. Given a matroid on elements $[n]$ with k different cost functions $d_i : [n] \to \mathbb{Z}_+$ (for $i = 1, \cdots, k$) and budgets $\{B_i\}_{i=1}^k$, the goal is to find (if possible) a basis I with $d_i(I) \le B_i$ for each $i \in [k]$. We obtain an algorithm that for any $\epsilon > 0$ finds in $n^{O(k^{1.5}/\epsilon)}$ time, a basis I with $d_i(I) \le (1+\epsilon)B_i$ for all $i \in [k]$. Previously, [8] obtained such an algorithm with $n^{O(k^2/\epsilon)}$ running time.

Low congestion routing. Given a directed graph $G = (V, E)$ with edge capacities $b : E \to \mathbb{Z}_+$, k source-sink pairs $\{(s_i, t_i)\}_{i=1}^k$ and a length bound Δ, the goal is to find an $s_i - t_i$ path P_i of length at most Δ for each pair $i \in [k]$ such that the number N_e of paths using any edge e is at most b_e. Using an LP-based reduction [6] this can be cast as an instance of DegMat. So we obtain violation bounds as in (7) which implies:

$$N_e \le b_e + \min\left\{O(\sqrt{\Delta}\log n),\ O(\sqrt{b_e}\log n + \log^2 n)\right\}, \quad \forall e \in E.$$

Here $n = |V|$ is the number of vertices. Previous algorithms achieved bounds of $\Delta - 1$ [10] or $O(\sqrt{\log \Delta} \cdot \sqrt{b_j} + \log \Delta)$ [13] separately. We can also handle a richer set of routing requirements: given a laminar family $\mathcal{L}$ on the k pairs, with a requirement r_T on each set $T \in \mathcal{L}$, we want to find a multiset of paths so that there are at least r_T paths between the pairs in each $T \in \mathcal{L}$. Although this is not an instance of DegMat, the same approach works.

Overview of techniques: Our algorithm in Theorem 4 is similar to the Lovett–Meka algorithm, and is also based on performing a Gaussian random walk at each step in a suitably chosen subspace. However, there some crucial differences. First, instead of updating each variable by the standard Gaussian $N(0, 1)$, the variance for variable i is chosen proportional to $\min(y_i, 1 - y_i)$, i.e. proportional to how close it is to the boundary 0 or 1. This is crucial for getting the multiplicative error instead of the additive error in the constraints. However, this slows down the "progress" of variables toward reaching 0 or 1. To get around this, we add $O(\log n)$ additional constraints to define the subspace where the walk is performed: these restrict the total fractional value of variables in a particular "scale" to remain fixed. Using these we can ensure that enough variables eventually reach 0 or 1.

In order to handle the matroid constraints (Theorem 5) we need to incorporate them (although they are exponentially many) in defining the subspace where the random walk is performed. One difficulty that arises here is that we can no longer implement the random walk using "near tight" constraints as in [15] since we are unable to bound the dimension of near-tight matroid constraints. However, as is well known, the dimension of *exactly* tight matroid constraints is at most $n/2$ at any (strictly) fractional solution, and so we implement the random walk using exactly tight constraints. This requires us to truncate certain steps in the random walk (when we move out of the polytope), but we show that the effect of such truncations is negligible.

2 Matroid Partial Rounding

In this section we will prove Theorem 5 which also implies Theorem 4.

We may assume, without loss of generality, that $\max_{j=1}^{m} \lambda_j \le n$. This is because setting $\mu_j = \min\{\lambda_j, n\}$ we have $\sum_{j=1}^{m} e^{-\mu_j^2/K_0} \le \sum_{j=1}^{m} e^{-\lambda_j^2/K_0} + m \cdot e^{-n} < \frac{n}{16} + 1$ (we used the assumption $m \le 2^n$). So we can apply Theorem 5 with μ_js instead of λ_js to obtain a stronger result.

Let $y \in \mathbb{R}^n$ denote the initial solution. The algorithm will start with $X_0 = y$ and update this vector over time. Let X_t denote the vector at time t for $t = 1, \ldots, T$. The value of T will be defined later. Let $\ell = 3\lceil \log_2 n \rceil$. We classify the n elements into 2ℓ classes based on their initial values $y(i)$ as follows.

$$U_k := \begin{cases} \left\{i \in [n] : 2^{-k-1} < y(i) \le 2^{-k}\right\} & \text{if } 1 \le k \le \ell - 1 \\ \left\{i \in [n] : y(i) \le 2^{-\ell}\right\} & \text{if } k = \ell. \end{cases}$$

$$V_k := \begin{cases} \left\{i \in [n] : 2^{-k-1} < 1 - y(i) \le 2^{-k}\right\} & \text{if } 1 \le k \le \ell - 1 \\ \left\{i \in [n] : 1 - y(i) \le 2^{-\ell}\right\} & \text{if } k = \ell. \end{cases}$$

Note that the U_k's partition elements of value (in y) between 0 and $\frac{1}{2}$ and the V_k's form a symmetric partition of elements valued between $\frac{1}{2}$ and 1. This partition does not change over time, even though the value of variables might change. We define the “scale” of each element as:

$$s_i \;:=\; 2^{-k}, \qquad \forall i \in U_k \cup V_k, \quad \forall k \in [\ell].$$

Define $W_j(s) = \sum_{i=1}^{n} a_{ji}^2 \cdot s_i^2$ for each $j \in [m]$. Note that $W_j(s) \ge W_j(y)$ and

$$W_j(s) - 4 \cdot W_j(y) \le \sum_{i=1}^{n} a_{ji}^2 \cdot \frac{1}{n^6} = \frac{\|a_j\|^2}{n^6}.$$

So $\sqrt{W_j(y)} \le \sqrt{W_j(s)} \le 2\sqrt{W_j(y)} + \frac{\|a_j\|}{n^3}$. Our algorithm will find a solution y' with $\Omega(n)$ integral variables such that:

$$|\langle y' - y, a_j \rangle| \le \lambda_j \cdot \sqrt{W_j(s)} + \frac{1}{n^3} \cdot \|a_j\|, \quad \forall j \in [m].$$

This suffices to prove Theorem 5 as

$$\lambda_j \cdot \sqrt{W_j(s)} + \frac{1}{n^3} \cdot \|a_j\| \le 2\lambda_j \cdot \sqrt{W_j(y)} + \left(\frac{1}{n^3} + \frac{\lambda_j}{n^3}\right) \cdot \|a_j\| \le 2\lambda_j \cdot \sqrt{W_j(y)} + \frac{\|a_j\|}{n^2}.$$

Consider the polytope $\mathcal{Q}$ of points $x \in \mathbb{R}^n$ satisfying the following constraints.

$$x \in P(\mathcal{M}), \tag{8}$$

$$|\langle x - y, a_j \rangle| \le \lambda_j \cdot \sqrt{W_j(s)} + \frac{1}{n^3} \cdot \|a_j\| \qquad \forall j \in [m], \tag{9}$$

$$\sum_{i \in U_k} x_i = \sum_{i \in U_k} y_i \qquad \forall k \in [\ell], \tag{10}$$

$$\sum_{i \in V_k} x_i = \sum_{i \in V_k} y_i \qquad \forall k \in [\ell], \tag{11}$$

$$0 \le x_i \le \min\{\alpha \cdot 2^{-k}, 1\} \qquad \forall i \in U_k,\ \forall k \in [\ell], \tag{12}$$

$$0 \le 1 - x_i \le \min\{\alpha \cdot 2^{-k}, 1\} \qquad \forall i \in V_k,\ \forall k \in [\ell]. \tag{13}$$

Above $\alpha = 40$ is a constant whose choice will be clear later. The algorithm will maintain the invariant that at any time $t \in [T]$, the solution X_t lies in $\mathcal{Q}$. In particular the constraint (8) requires that X_t stays in the matroid polytope. Constraint (9) controls the violation of the side constraints over all time steps. The last two constraints (12) and (13) enforce that variables in U_k (and symmetrically V_k) do not deviate far beyond their original scale of 2^{-k}. The constraints (10) and (11) ensure that throughout the algorithm, the total value of elements in U_k (and V_k) stay equal to their initial sum (in y). These constraints will play a crucial role in arguing that the algorithm finds a partial coloring. Note that there are only 2ℓ such constraints.

In order to deal with complexity issues, we will assume (without loss of generality, by scaling) that all entries in the constraints describing $\mathcal{Q}$ are integers bounded by some value B. Our algorithm will then run in time polynomial in n, m and $\log_2 B$, given an independent set oracle for the matroid $\mathcal{M}$. Also, our algorithm will only deal with points having rational entries of small "size". Recall that the size of a rational number is the number of bits needed to represent it, i.e. the size of p/q (where $p, q \in \mathbb{Z}$) is $\log_2 |p| + \log_2 |q|$.

The Algorithm: Let $\gamma = n^{-6}$ and $T = K/\gamma^2$ where $K := 10\alpha^2$. The algorithm starts with solution $X_0 = y \in \mathcal{Q}$, and does the following at each time step $t = 0, 1, \ldots, T$:

1. Consider the set of constraints of $\mathcal{Q}$ that are tight at the point $x = X_t$, and define the following sets based on this.

(a) Let $\mathcal{C}_t^{var}$ be the set of tight variable constraints among (12) and (13). This consists of:

i. $i \in U_k$ (for any k) with $X_t(i) = 0$ or $X_t(i) = \min\{\alpha \cdot 2^{-k},\ 1\}$; and
ii. $i \in V_k$ (for any k) with $X_t(i) = 1$ or $X_t(i) = \max\{1 - \alpha \cdot 2^{-k},\ 0\}$.

(b) Let $\mathcal{C}_t^{side}$ be the set of tight side constraints from (9), i.e. those $j \in [m]$ with

$$|\langle X^t - y, a_j\rangle| = \lambda_j \cdot \sqrt{W_j(s)} + \frac{1}{n^3}\,\|a_j\|.$$

(c) Let $\mathcal{C}_t^{part}$ denote the set of the 2ℓ equality constraints (10) and (11).
(d) Let $\mathcal{C}_t^{rank}$ be a maximal linearly independent set of tight rank constraints from (8). As usual, a set of constraints is said to be linearly independent if the corresponding coefficient vectors are linearly independent. Since $\mathcal{C}_t^{rank}$ is maximal, every tight rank constraint is a linear combination of constraints in $\mathcal{C}_t^{rank}$. By Claim 2, $|\mathcal{C}_t^{rank}| \leq n/2$.

2. Let $\mathcal{V}_t$ denote the subspace orthogonal to all the constraints in $\mathcal{C}_t^{var}$, $\mathcal{C}_t^{side}$, $\mathcal{C}_t^{part}$ and $\mathcal{C}_t^{rank}$. Let D be an $n \times n$ diagonal matrix with entries $d_{ii} = 1/s_i$, and let $\mathcal{V}_t'$ be the subspace $\mathcal{V}_t' = \{Dv : v \in \mathcal{V}_t\}$. As D is invertible, $\dim(\mathcal{V}_t') = \dim(\mathcal{V}_t)$.
3. Let $\{b_1, \dots, b_k\}$ be an almost orthonormal basis of $\mathcal{V}_t'$ given by Fact 2. Note that all entries in these vectors are rationals of size $O(n^2 \log B)$.
4. Let G_t be a random direction defined as $G_t := \sum_{h=1}^{k} g_h b_h$ where the g_h are independent $\{-1, +1\}$ Bernoulli random variables.
5. Let $\overline{G}_t := D^{-1} G_t$. As $G_t \in \mathcal{V}_t'$, it must be that $G_t = Dv$ for some $v \in \mathcal{V}_t$ and thus $\overline{G}_t = D^{-1} G_t \in \mathcal{V}_t$. Note that all entries in $\overline{G}_t$ are rationals of size $O(n^3 \log B)$.
6. Set $Y_t = X_t + \gamma \cdot \overline{G}_t$.

(a) If $Y_t \in \mathcal{Q}$ then $X_{t+1} \leftarrow Y_t$ and continue to the next iteration.
(b) Else $X_{t+1} \leftarrow$ the point in $\mathcal{Q}$ that lies on the line segment (X_t, Y_t) and is closest to Y_t. This can be found by binary search and testing membership in the matroid polytope. By Claim 1 it follows that the number of steps in the binary search is at most $O(n \log B)$.

This completes the description of the algorithm. We actually do not need to compute the tight constraints from scratch in each iteration. We start the algorithm off with a strictly feasible solution $y \in \mathcal{Q}$ which does not have any tight constraint other than (10) and (11). Then, the only place a new constraint gets tight is Step 6b: at this point, we add the new constraint to the appropriate set among $\mathcal{C}_t^{var}$, $\mathcal{C}_t^{side}$ and $\mathcal{C}_t^{var}$ and continue.

In order to keep the analysis clean and convey the main ideas, we will assume that the basis $\{b_1, \dots, b_k\}$ in Step 3 is exactly orthonormal. When the basis is "almost orthonormal" as given in Fact 2, the additional error incurred is negligible.

Running Time Since the number of iterations is polynomial, we only need to show that each of the steps in any single iteration can be implemented in polynomial time. The only step that requires justification is 6b, which is shown in Claim 1.

Moreover, we need to ensure that all points considered in the algorithm have rational coefficients of polynomial size. This is done by a rounding procedure (see Fact 1) that given an arbitrary point, finds a nearby rational point of size $O(n^2 \log B)$. Since the number of steps in the algorithm is polynomial, the total error incurred by such rounding steps is small.

Claim 1 *The number of binary search iterations performed in Step 6b is* $O(n^4 \log B)$.

Proof To reduce notation let $a = X_t$, $d = \gamma \overline{G}_t$ and $Y(\mu) := a + \mu \cdot d \in \mathbb{R}^n$ for any $\mu \in \mathbb{R}$. Recall that Step 6b involves finding the maximum value of μ such that point $Y(\mu) \in \mathcal{Q}$.

By the rounding procedure (Fact 1) we know that a has rational entries of size $O(n^2 \log B)$.

We now show that the direction d has rational entries of size $O(n^3 \log B)$. This is because (i) the basis vectors $\{b_1, \cdots, b_k\}$ have rational entries of size $O(n^2 \log B)$ by Fact 2, (ii) $G_t = \sum_{h=1}^{k} g_h \cdot b_h$ (where each $g_h = \pm 1$) has rational entries of size $O(n^3 \log B)$ and (iii) $\overline{G}_t = D^{-1} G_t$ where D^{-1} is a diagonal matrix with rational entries of size $O(n \log B)$.

Next, observe that for any constraint $\langle a', x \rangle \leq \beta$ in $\mathcal{Q}$, the point of intersection of the hyperplane $\langle a', x \rangle = \beta$ with line $\{Y(\mu) : \mu \in \mathbb{R}\}$ is $\mu = \frac{\beta - \langle a', a \rangle}{\langle a', d \rangle}$ which is a rational of size at most $\sigma = O(n^4 \log B)$ as a', a, d, β all have rational entries of size $O(n^3 \log B)$. Let $\epsilon = 2^{-2\sigma}$ be a value such that the difference between any two distinct rationals of size at most σ is more than ϵ.

In Step 6b, we start the binary search with the interval $[0, 1]$ for μ where $Y(0) \in \mathcal{Q}$ and $Y(1) \notin \mathcal{Q}$. We perform this binary search until the interval width falls below ϵ, which requires $\log_2 \frac{1}{\epsilon} = O(n^4 \log B)$ iterations. At the end, we have two values $\mu_0 < \mu_1$ with $\mu_1 - \mu_0 < \epsilon$ such that $Y(\mu_0) \in \mathcal{Q}$ and $Y(\mu_1) \notin \mathcal{Q}$. Moreover, we obtain a constraint $\langle a', x \rangle \leq \beta$ in $\mathcal{Q}$ that is not satisfied by $Y(\mu_1)$. We set μ' to be the (unique) value such that $Y(\mu')$ satisfies this constraint at equality, and set $X_{t+1} = Y(\mu')$. Note that $\mu_0 \leq \mu' < \mu_1$. To see that $Y(\mu') \in \mathcal{Q}$, suppose (for contradiction) that some constraint in $\mathcal{Q}$ is not satisfied at $Y(\mu')$; then the point of intersection of line $\{Y(\mu) : \mu \in \mathbb{R}\}$ with this constraint must be at $\mu \in [\mu_0, \mu')$ which (by the choice of ϵ) can not be a rational of size at most σ— a contradiction. □

Analysis The analysis involves proving the following main lemma.

Lemma 1 *With constant probability, the final solution* X_T *has* $|\mathcal{C}_T^{var}| \geq \frac{n}{20}$.

We first show how this implies Theorem 5.

Proof of Theorem 5 from Lemma 1 The algorithm outputs the solution $y' := X_T$. By design the algorithm ensures that $X_T \in \mathcal{Q}$, and thus $X_T \in P(\mathcal{M})$ and it satisfies the error bounds (9) on the side constraints. It remains to show that $\Omega(n)$ variables in X_T must be integer valued whenever $|\mathcal{C}_T^{var}| \geq \frac{n}{20}$. For each $k \in [\ell]$ define $u_k := |\{i \in U_k : X_T(i) = \alpha \cdot 2^{-k}\}|$ and $v_k := |\{i \in V_k : X_T(i) = 1 - \alpha \cdot 2^{-k}\}|$. By the equality constraints (10) for U_k, it follows that

$$u_k \cdot \alpha \cdot 2^{-k} \le \sum_{i \in U_k} X_T(i) = X_T(U_k) = y(U_k) \le |U_k| \cdot 2^{-k}.$$

This gives that $u_k \le \frac{1}{\alpha}|U_k|$. Similarly, $v_k \le \frac{1}{\alpha}|V_k|$. This implies that $\sum_{k=1}^{\ell}(u_k + v_k) \le n/\alpha$. As the tight variables in $\mathcal{C}_t^{var}$ have values either 0 or 1 or $\alpha \cdot 2^{-k}$ or $1 - \alpha \cdot 2^{-k}$, it follows that the number of $\{0, 1\}$ variables is at least

$$|\mathcal{C}_t^{var}| - \sum_{k=1}^{\ell}(u_k + v_k) \ge \left(|\mathcal{C}_t^{var}| - \frac{n}{\alpha}\right) \ge \left(\frac{1}{20} - \frac{1}{\alpha}\right) n$$

which is at least $n/40$ by choosing $\alpha = 40$. □

In the rest of this section we prove Lemma 1.

Claim 2 *Given any $x \in P(\mathcal{M})$ with $\mathbf{0} < x < \mathbf{1}$, the maximum number of tight linearly independent rank constraints is $n/2$.*

Proof Recall that a tight constraint in $P(\mathcal{M})$ is any subset $T \subseteq V$ with $\sum_{i \in T} x_i = r(T)$. The claim follows from the known property (see e.g. [20]) that for any $x \in P(\mathcal{M})$ there is a linearly independent collection $\mathcal{C}$ of tight constraints such that (i) $\mathcal{C}$ spans all tight constraints and (ii) $\mathcal{C}$ forms a chain family. Since all right-hand-sides are integer and each variable is strictly between 0 and 1, it follows that $|\mathcal{C}| \le \frac{n}{2}$. □

Claim 3 *The truncation Step 6b occurs at most n times.*

Proof We will show that whenever Step 6b occurs (i.e. the random move gets truncated) the dimension $dim(\mathcal{V}_{t+1})$ decreases by at least 1, i.e. $dim(\mathcal{V}_{t+1}) \le dim(\mathcal{V}_t) - 1$. As the maximum dimension is n this would imply the claim.

Let $\mathcal{E}_t$ denote the subspace spanned by all the tight constraints of $X_t \in \mathcal{Q}$; Recall that $\mathcal{V}_t = \mathcal{E}_t^{\perp}$ is the subspace orthogonal to $\mathcal{E}_t$, and thus $dim(\mathcal{E}_t) = n - dim(\mathcal{V}_t)$. We also have $\mathcal{E}_0 \subseteq \mathcal{E}_1 \subseteq \cdots \mathcal{E}_T$. Suppose that Step 6b occurs in iteration t. Then we have $X_t \in \mathcal{Q}$, $Y_t \notin \mathcal{Q}$ and $Y_t - X_t \in \mathcal{V}_t$. Moreover $X_{t+1} = X_t + \epsilon(Y_t - X_t) \in \mathcal{Q}$ where $\epsilon \in [0, 1)$ is such that $X_t + \epsilon'(Y_t - X_t) \notin \mathcal{Q}$ for all $\epsilon' > \epsilon$. So there is some constraint $\langle a', x\rangle \le \beta$ in $\mathcal{Q}$ with:

$$\langle a', X_t\rangle \le \beta, \quad \langle a', X_{t+1}\rangle = \beta \quad \text{and} \quad \langle a', Y_t\rangle > \beta.$$

Since this constraint satisfies $\langle a', Y_t - X_t\rangle > 0$ and $Y_t - X_t \in \mathcal{V}_t$, we have $a' \notin \mathcal{E}_t$. As a' is added to $\mathcal{E}_{t+1}$, we have $dim(\mathcal{E}_{t+1}) \ge 1 + dim(\mathcal{E}_t)$. This proves the desired property and the claim. □

The statements of the following two lemmas are similar to those in [15], but the proofs require additional work since our random walk is different. The first lemma shows that the expected number of tight side constraints at the end of the algorithm is not too high, and the second lemma shows that the expected number of tight variable constraints is large.

Lemma 2 $\mathbb{E}[|\mathcal{C}_T^{side}|] < \frac{n}{4}$.

Proof Note that $X_T - y = \gamma \sum_{t=0}^{T} \overline{G}_t + \sum_{q=1}^{n} \Delta_{t(q)}$ where Δs correspond to the truncation incurred during the iterations $t = t(1), \cdots, t(n)$ for which Step 6b applies (by Claim 3 there are at most n such iterations). Moreover for each q, $\Delta_{t(q)} = \delta \cdot \overline{G}_{t(q)}$ for some δ with $0 < |\delta| < \gamma$.

If $j \in \mathcal{C}_T^{side}$, then $|\langle X_T - y, a_j \rangle| = \lambda_j \sqrt{W_j(s)} + \frac{1}{n^3} \cdot \|a_j\|$. We have

$$|\langle X_T - y, a_j\rangle| \leq |\gamma \sum_{t=0}^{T} \langle \overline{G}_t, a_j\rangle| + \sum_{q=1}^{n} \gamma |\langle \overline{G}_{a(q)}, a_j\rangle| \leq |\gamma \sum_{t=0}^{T} \langle \overline{G}_t, a_j\rangle|$$
$$+ n\gamma \cdot \max_{t=0}^{T} |\langle \overline{G}_t, a_j\rangle|.$$

Note that at any iteration t,

$$|\langle \overline{G}_t, a_j\rangle| = |\langle D^{-1} G_t, a_j\rangle| \leq |\langle G_t, a_j\rangle| \leq \sum_{h=1}^{k} |\langle b_h, a_j\rangle| \leq n\|a_j\|.$$

The first inequality above uses that D^{-1} is a diagonal matrix with entries at most one, the second inequality is by definition of G_t where $\{b_h\}$ is an orthonormal basis of $\mathcal{V}_t'$, and the last inequality uses that each b_h is a unit vector. As $\gamma = n^{-6}$, we have $n\gamma \cdot \max_{t=0}^{T} |\langle \overline{G}_t, a_j\rangle| \leq \|a_j\|/n^4$. So it follows that if $j \in \mathcal{C}_T^{side}$, then we must have:

$$|\gamma \sum_{t=0}^{T} \langle \overline{G}_t, a_j\rangle| \;\geq\; \lambda_j \sqrt{W_j(s)}.$$

In order to bound the probability of this event, we consider the sequence $\{Z_t\}$ where $Z_t = \langle \overline{G}_t, a_j\rangle$, and note the following useful facts.

Observation 1 *The sequence $\{Z_t\}$ forms a martingale satisfying:*

1. $\mathbb{E}\left[Z_t \mid Z_{t-1}, \ldots, Z_0\right] = 0$ *for all t.*
2. $|Z_t| \leq n\|a_j\|$ *whp for all t.*
3. $\mathbb{E}\left[Z_t^2 \mid Z_{t-1}, \ldots, Z_0\right] \leq \sum_{i=1}^{n} s_i^2 \cdot a_{ji}^2 = W_j(s)$ *for all t.*

Proof As $\overline{G}_t = \sum_{h=1}^{k} g_h \cdot b_h$ where each $\mathbb{E}[g_h] = 0$, we have $\mathbb{E}[\overline{G}_t | \overline{G}_0, \ldots, \overline{G}_{t-1}] = \mathbf{0}$. Note that $\overline{G}_t$ is not independent of $\overline{G}_0, \ldots, \overline{G}_{t-1}$, as these choices determine the subspace where $\overline{G}_t$ lies. So $\{Z_t\}$ forms a martingale sequence with the first property.

For the remaining two properties, we fix $j \in [m]$ and t and condition on $Z_0, \ldots, Z_{t-1}$. To reduce notation we drop all subscripts: so $a = a_j$, $G = G_t$, $\mathcal{V}' = \mathcal{V}_t'$ and $Z = Z_t$.

Let $\{b_r\}$ denote an orthonormal basis for the linear subspace $\mathcal{V}'$. Then $G = \sum_r g_r \cdot b_r$ where each g_r is iid ± 1 with probability half. As $\overline{G} = D^{-1}G$, we have $Z = \langle \overline{G}, a\rangle = \sum_r \langle D^{-1}b_r, a\rangle\, g_r = \sum_r \langle D^{-1}a, b_r\rangle\, g_r$. So, we can bound

$$|Z| \;\le\; \sum_r |\langle D^{-1}a, b_r\rangle| \cdot |g_r| \;\le\; \|D^{-1}a\| \sum_r |g_r| \;\le\; n\|a\|.$$

The first inequality follows from the triangle inequality, the second by Cauchy–Schwartz and as b_r is a unit-vector, and the third follows as D^{-1} is a diagonal matrix with entries at most one. This proves property 2.

Finally, $\mathbb{E}[Z^2] = \sum_r \langle D^{-1}a, b_r\rangle^2\, \mathbb{E}[g_r^2] = \sum_r \langle D^{-1}a, b_r\rangle^2 \le \|D^{-1}a\|^2$, where the last step follows as $\{b_r\}$ is an orthonormal basis for a subspace of $\mathbb{R}^n$. This proves property 3. □

Using a martingale concentration inequality, we obtain:

Claim 4 $\Pr\left[|\gamma \sum_{t=0}^{T} \langle \overline{G}_t, a_j\rangle| \;\ge\; \lambda_j \sqrt{W_j(s)}\right] = \Pr\left[|\sum_{t=0}^{T} Z_t| \;\ge\; \frac{\lambda_j}{\gamma}\sqrt{W_j(s)}\right] \le 2\cdot \exp(-\lambda_j^2/3K)$.

Proof The first equality is by definition of the Z_ts. We now use the following concentration inequality:

Theorem 6 (Freedman [7] (Theorem 1.6)) *Consider a real-valued martingale sequence $\{Z_t\}_{t\ge 0}$ such that $Z_0 = 0$, $\mathbb{E}[Z_t \mid Z_{t-1}, \ldots, Z_0] = 0$ for all t, and $|Z_t| \le M$ almost surely for all t. Let $W_t = \sum_{j=0}^{t} \mathbb{E}\left[Z_j^2 \mid Z_{j-1}, Z_{j-2}, \ldots Z_0\right]$ for all $t \ge 1$. Then for all $\ell \ge 0$ and $\sigma^2 > 0$, and any stopping time τ we have*

$$\Pr\left[|\sum_{j=0}^{\tau} Z_j| \ge \ell \textit{ and } W_\tau \le \sigma^2\right] \quad\le\quad 2\exp\left(-\frac{\ell^2/2}{\sigma^2 + M\ell/3}\right)$$

We apply this with $M = n\|a_j\|$, $\ell = \frac{\lambda_j}{\gamma}\sqrt{W_j(s)}$, $\sigma^2 = T \cdot W_j(s)$ and $\tau = T$. Note that

$$\frac{\ell^2}{2\sigma^2 + \frac{2}{3}M\ell} \quad=\quad \frac{\lambda_j^2}{2\gamma^2 T + \frac{2}{3}\gamma n\|a_j\|\lambda_j/\sqrt{W_j(s)}} \quad\ge\quad \frac{\lambda_j^2}{2\gamma^2 T + 1},$$

where the last inequality uses $W_j(s) \ge \|a_j\|^2/n^6$, $\lambda_j \le n$ and $\gamma = n^{-6}$. Thus

$$\Pr\left[|\gamma \sum_{t=0}^{T} \langle \overline{G}_t, a_j\rangle| \;\ge\; \lambda_j \sqrt{W_j(s)}\right] \le 2\exp\left(\frac{-\lambda_j^2}{2\gamma^2 T + 1}\right) \le 2\cdot\exp(-\lambda_j^2/3K).$$

The last inequality uses $T = K/\gamma^2$ and $K \ge 1$. This completes the proof of the claim. □

By the above claim, we have $\mathbb{E}[|\mathcal{C}_T^{side}|] < 2\sum_{j=1}^m \exp(-\lambda_j^2/(30\alpha^2)) < 0.25n$. This completes the proof of Lemma 2. □

We now prove that in expectation, at least $0.1n$ variables become tight at the end of the algorithm. This immediately implies Lemma 1.

Lemma 3 $\mathbb{E}[|\mathcal{C}_T^{var}|] \geq 0.1n$.

Proof Define the following potential function, which will measure the progress of the algorithm toward the variables becoming tight.

$$\Phi(x) \quad := \quad \sum_{k=1}^{\ell} 2^{2k} \cdot \left(\sum_{i \in U_k} x(i)^2 + \sum_{i \in V_k} (1 - x(i))^2 \right), \qquad \forall x \in \mathcal{Q}.$$

Note that since $X_T \in \mathcal{Q}$, we have $X_T(i) \leq \alpha \cdot 2^{-k}$ for $i \in U_k$ and $1 - X_T(i) \leq \alpha \cdot 2^{-k}$ for $i \in V_k$. So $\Phi(X_T) \leq \alpha^2 \cdot n$. We also define the "incremental function" for any $x \in \mathcal{Q}$ and $g \in \mathbb{R}^n$, $f(x, g) := \Phi(x + \gamma D^{-1} g) - \Phi(x)$. Recall that D^{-1} is the $n \times n$ diagonal matrix with entries $(s_1, \ldots, s_n)$ where $s_i = 2^{-k}$ for $i \in U_k \cup V_k$. So

$$\begin{aligned} f(x, g) &= \gamma^2 \sum_{i=1}^n g(i)^2 + 2 \sum_{k=1}^{\ell} 2^{2k} \cdot \left(\sum_{i \in U_k} x(i) \gamma s_i \cdot g(i) - \sum_{i \in V_k} (1 - x(i)) \gamma s_i \cdot g(i) \right) \\ &= \gamma^2 \sum_{i=1}^n g(i)^2 + 2\gamma \sum_{k=1}^{\ell} \left(\sum_{i \in U_k} \frac{x(i) g(i)}{s_i} - \sum_{i \in V_k} \frac{(1 - x(i)) g(i)}{s_i} \right) \end{aligned}$$

Suppose the algorithm was modified to never have the truncation step 6b, then in any iteration t, the increase $\Phi(Y_t) - \Phi(X_t) = f(X_t, G_t)$ where G_t is the random direction chosen in $\mathcal{V}_t'$. The following is by simple calculation.

$$\begin{aligned} f(X_t, G_t) - f(X_t, \delta G_t) &= \gamma^2 (1 - \delta^2) \|G_t\|_2^2 + 2\gamma (1 - \delta) \\ &\quad \sum_{k=1}^{\ell} \left(\sum_{i \in U_k} \frac{X_t(i)}{s_i} \cdot G_t(i) - \sum_{i \in V_k} \frac{1 - X_t(i)}{s_i} \cdot G_t(i) \right) \\ &\leq \gamma^2 (1 - \delta^2) \|G_t\|_2^2 + 2\alpha\gamma (1 - \delta) \sum_{i=1}^n |G_t(i)| \\ &\leq \gamma^2 \|G_t\|_2^2 + 2\gamma\alpha \|G_t\|_1 \\ &\leq \gamma^2 n + 2\gamma\alpha n^{3/2} \quad \leq \quad \frac{1}{n} \end{aligned} \tag{14}$$

The first inequality in (14) uses the fact that G_t is the sum of orthogonal unit vectors, and the second inequality uses $\gamma = n^{-6}$ and $\alpha = O(1)$.

This implies that

$$\begin{aligned}\Phi(X_T) - \Phi(X_0) = \sum_{t=0}^{T} f(X_t, \delta_t G_t) \;\geq\; &\sum_{t=0}^{T} f(X_t, G_t)\\ &-\frac{1}{n}\sum_{t=0}^{T} \mathbf{1}[\text{step 6b occurs in iteration } t]\\ \geq &\sum_{t=0}^{T} f(X_t, G_t) - 1 \qquad \text{(by Claim 3)} \qquad (15)\end{aligned}$$

Claim 5 $\mathbb{E}[\Phi(X_T)] - \Phi(y) \geq \gamma^2 T \cdot \mathbb{E}[\dim(\mathcal{V}_T)] - 1$.

Proof From (15) we have:

$$\mathbb{E}[\Phi(X_T)] - \Phi(X_0) \geq \sum_{t=0}^{T} \mathbb{E}[f(X_t, G_t)] - 1. \qquad (16)$$

In any iteration t, as $G_t = \sum_{h=1}^{k} g_h b_h$ where $\{b_h\}$ is an orthonormal basis for $\mathcal{V}_t'$ and $g_h = \pm 1$,

$$\begin{aligned}\mathbb{E}[f(X_t, G_t)] &= \gamma^2 \sum_{i=1}^{n} \mathbb{E}[G_t(i)^2] = \gamma^2 \sum_{h=1}^{k} \|b_h\|^2 = \gamma^2 k = \gamma^2 \mathbb{E}[dim(\mathcal{V}_t')]\\ &= \gamma^2 \mathbb{E}[dim(\mathcal{V}_t)].\end{aligned}$$

Moreover, because $\mathcal{V}_0 \supseteq \mathcal{V}_1 \supseteq \cdots \mathcal{V}_T$, we have $\mathbb{E}[dim(\mathcal{V}_t)] \geq \mathbb{E}[dim(\mathcal{V}_T)]$. So

$$\sum_{t=0}^{T} \mathbb{E}[f(X_t, G_t)] \quad \geq \quad \gamma^2 T \cdot \mathbb{E}[dim(\mathcal{V}_T)]. \qquad (17)$$

Combining (16) and (17), we complete the proof of Claim 5. □

By Claim 2 and the fact that $|\mathcal{C}_T^{part}| = 2\ell$, we have

$$\begin{aligned}\dim(\mathcal{V}_T) &\geq n - \dim(\mathcal{C}_T^{var}) - \dim(\mathcal{C}_T^{side}) - \dim(\mathcal{C}_T^{rank}) - \dim(\mathcal{C}_T^{part})\\ &\geq \frac{n}{2} - 2\ell - \dim(\mathcal{C}_T^{var}) - \dim(\mathcal{C}_T^{side})\end{aligned}$$

Taking expectations and by Claim 2, this gives

$$\mathbb{E}[\dim(\mathcal{V}_T)] \geq \frac{n}{4} - 2\ell - \dim(\mathcal{C}_T^{var}) \qquad (18)$$

Using $\Phi(X_T) \leq \alpha^2 n$ and Claim 5, we obtain:

$$\alpha^2 n \geq \mathbb{E}[\Phi_T] \geq \gamma^2 T \cdot \left(\frac{n}{4} - 2\ell - \mathbb{E}[\dim(\mathcal{C}_T^{var})]\right) - 1.$$

Rearranging and using $T = K/\gamma^2$, $K = 10\alpha^2$ and $\ell = \log n$ gives that

$$\mathbb{E}[\dim(\mathcal{C}_T^{var})] \geq \frac{n}{4} - \frac{\alpha^2 n}{K} - 2\ell - \frac{1}{K} \geq 0.1n,$$

where we used $K = 10\alpha^2$, $\alpha = 40$ and $\ell = O(\log n)$. This completes the proof of Lemma 3. □

3 Applications

3.1 Linear System Rounding with Violations

Consider a 0–1 integer program on n variables where each constraint $j \in [m]$ corresponds to some subset $S_j \subseteq [n]$ of the variables having total value $b_j \in \mathbb{Z}_+$. That is,

$$P \quad = \quad \left\{ x \in \{0,1\}^n \quad : \quad \sum_{i \in S_j} x_i = b_j, \;\; \forall j \in [m] \right\}.$$

Theorem 7 *There is a randomized polynomial time algorithm that given any fractional solution satisfying the constraints in P, finds an integer solution $x \in \{0,1\}^n$ where for each $j \in [m]$,*

$$|x(S_j) - b_j| \quad \leq \quad O(1) \cdot \min\left\{ \sqrt{j},\; \sqrt{n \log(m/n)},\; \sqrt{\log m \log n \cdot b_j} + \log m \log n,\; \sqrt{\Delta} \log n \right\}.$$

Above $\Delta = \max_{i=1}^{n} |\{j \in [m] : i \in S_j\}|$ is the maximum number of constraints that any variable appears in.

Proof Let $y \in [0,1]^n$ be a fractional solution with $\sum_{i \in S_j} y_i = b_j$ for all $j \in [m]$. The algorithm in Theorem 7 uses Theorem 4 iteratively to obtain the integral solution x.

In each iteration, we start with a fractional solution y' with $f \leq n$ *fractional* variables and set the parameters λ_j suitably so that $\sum_{j=1}^{m} e^{-\lambda_j^2/K_0} \leq \frac{f}{16}$. That is, the condition in Theorem 4 is satisfied. Note that $W_j(y') = \sum_{i \in S_j} (y'_i)^2 \leq y'(S_j)$ and

$W_j(y') \le f$. Now, by applying Theorem 4, we would obtain a new fractional solution y'' such that:

- For each $j \in [m]$, $|y''(S_j) - y'(S_j)| \le \lambda_j \sqrt{W_j(y')} + \frac{1}{n} \le O(\lambda_j) \cdot \sqrt{f}$.
- The number of fractional variables in y'' is at most $\frac{f}{K}$ for some constant $K > 1$.

Therefore, after $\frac{\log n}{\log K} = O(\log n)$ iterations we obtain a solution with $O(1)$ fractional variables. Setting these fractional variables arbitrarily to 0–1 values, we obtain an integral solution x.

Let us partition the constraints into sets M_1, M_2, M_3 and M_4 based on which of the four terms in Theorem 7 is minimum. That is, $M_1 \subseteq [m]$ consists of constraints $j \in [m]$ where $\sqrt{j}$ is smaller than the other three terms; M_2, M_3, M_4 are defined similarly. Below we show how to set the parameters λ_j and bound the constraint violations for these parts separately.

Error bound of $\min\{\sqrt{j},\ \sqrt{n\log(m/n)}\}$ for $j \in M_1 \cup M_2$ In any iteration with $f \le n$ fractional variables, we set the parameters λ_js in Theorem 4 as follows:

$$\lambda_j = \begin{cases} 0 & \text{if } j < c_1 f \\ \sqrt{c_2 \log \frac{j}{c_1 f}} & \text{if } j \ge c_1 f \end{cases}$$

Here c_1 and c_2 are constants that will be fixed later. Note that

$$\sum_{j \in M_1 \cup M_2}^{m} e^{-\lambda_j^2/K_0} \le c_1 f + \sum_{j \ge c_1 f} e^{-\frac{c_2}{K_0}\log\frac{j}{c_1 f}} \le c_1 f + \sum_{i \ge 0} 2^i c_1 f \cdot e^{-ic_2/K_0}$$
$$\le c_1 f + c_1 f \sum_{i \ge 0} 2^{-i} \le 3c_1 f,$$

which is at most $f/48$ for $c_1 < 1/150$. The second inequality above is obtained by bucketing the js into intervals of the form $[2^i \cdot c_1 f,\ 2^{i+1} \cdot c_1 f]$. The third inequality uses $c_2 \ge 2K_0$.

We now bound the error incurred.

1. Consider first a constraint $j \le n$. Note that λ_j stays zero until the number of fractional variables f drops below j/c_1. So we can bound $|x(S_j) - b_j|$ by:

$$\sum_{i \ge 0} \sqrt{c_2 \frac{j}{c_1 K^i} \cdot \log K^i} \le O(\sqrt{j}) \sum_{i \ge 0} \sqrt{i} K^{-i/2} = O(\sqrt{j}),$$

where i indexes the iterations of the algorithm after f drops below j/c_1 for the first time.

2. Now consider a constraint $j > n$. Similarly, we bound $|x(S_j) - b_j|$ by:

$$\sum_{i\geq 0}\sqrt{c_2\frac{n}{K^i}\cdot\log(\frac{j}{c_1 n}K^i)} \leq O(\sqrt{n\log(j/n)})\sum_{i\geq 0}\sqrt{i}K^{-i/2} = O(\sqrt{n\log(j/n)}).$$

Here i indexes the number of iterations of the algorithm from its start.

Error bound of $\sqrt{L\cdot b_j}+L$ for $j\in M_3$, where $L=\Theta(\log m\log n)$ Note that the additive term in this expression is at least L. If any $b_j < L$ then we increase it to L (and add dummy elements to S_j and ensure $y(S_j) = L$); this only affects the error term by a constant factor as $L \leq \sqrt{L\cdot b_j}+L \leq 2L$. So in the following we assume that $\min_j b_j \geq L$.

Here we set $\lambda_j = \infty$ in all iterations, which satisfies $\sum_{j\in M_3} e^{-\lambda_j^2/K_0} = 0$.

The analysis of the error incurred is similar to that in Lemma 2 and we only sketch the details; the main difference is that we analyze the deviation in a combined manner over all $O(\log n)$ iterations. Fix any constraint $j \in [m]$. If we ignore the error due to the truncation steps over all iterations[4] then we can write $|x(S_j) - b_j| = |\sum_{t=0}^{P}\gamma Z_t|$ where $\gamma = n^{-6}$ and $Z_t = \langle\overline{G}_t, \mathbf{1}_{S_j}\rangle$; recall that each $\overline{G}_t = D^{-1}G_t$ for random direction G_t as in Step 4 of the algorithm in Sect. 2. Here $P = O(\log n/\gamma^2)$ since there are $O(\log n)$ iterations and $O(1/\gamma^2)$ steps in each iteration. We will use the concentration inequality in Theorem 6 with martingale $\{Z_t\}_{t\geq 0}$ and stopping time τ being the first time t' where $|\sum_{t=0}^{t'} Z_t| > \frac{1}{\gamma}\sqrt{Lb_j}$. Then it follows that at any step t' before stopping, the current solution y' satisfies $y'(S_j) - y(S_j) = \gamma\sum_{t=0}^{t'} Z_t \leq \sqrt{Lb_j} \leq b_j$ (using the assumption $b_j \geq L$), i.e. $y'(S_j) \leq 2b_j$. Now we can bound $W_\tau \leq P\cdot O(b_j) = O(\log n/\gamma^2)\cdot b_j$. Using Theorem 6 with $\ell = \sqrt{Lb_j}/\gamma$, we obtain:

$$\Pr\left[|\gamma\sum_{t=0}^{\tau} Z_t| \;\geq\; \sqrt{Lb_j}\right] \leq 2\exp\left(\frac{-Lb_j}{O(\log n)b_j}\right) \leq \frac{1}{m^2},$$

by choosing a large enough constant in $L = O(\log m\log n)$. It follows that with probability at least $1-m^{-2}$, we have $\tau = P$ and $|x(S_j)-b_j| = |\sum_{t=0}^{P}\gamma Z_t| \leq \sqrt{L\cdot b_j}$. Finally, taking a union bound over $|M_3| \leq m$ such events, we obtain that with high probability, $|x(S_j) - b_j| \leq \sqrt{L\cdot b_j}$ for all $j \in M_3$.

Error bound of $\sqrt{\Delta}\log n$ for $j\in M_4$ Here we set $\lambda_j = \sqrt{K_1\Delta}/\sqrt{|S_j|}$ in all iterations, where K_1 is a constant to be fixed later. We first bound $\sum_{j\in M_4} e^{-\lambda_j^2/K_0}$. Note that when restricted to the f fractional variables in any iteration, $\sum_{j=1}^{m}|S_j| \leq \Delta f$ since each variable appears in at most Δ constraints. So the number of constraints with $|S_j| > 64\Delta$ is at most $\frac{f}{64}$. For $h \geq 0$, the number of constraints

[4] This can be bounded by $o(1)$ exactly as in Event 2 of Lemma 2.

with $|S_j| \in [2^{-h-1}64\Delta, 2^{-h}64\Delta)$ is at most $2^{h+1}\frac{f}{64}$. So,

$$\sum_{j\in M_4} e^{-\lambda_j^2/K_0} \leq \frac{f}{64} + \sum_{h=0}^{\infty} 2^{h+1}\frac{f}{64}\exp\left(\frac{-K_1\Delta}{2^{-h}64\Delta\cdot K_0}\right)$$

$$\leq \frac{f}{64} + \frac{f}{64}\sum_{h=0}^{\infty} 2^{h+1}e^{-2^{h+2}} \leq \frac{f}{48}.$$

The second inequality is by choosing large enough constant K_1.

We now bound the error incurred for any constraint $j \in M_4$. The error in a single iteration is at most $O(\sqrt{\Delta}) + \frac{1}{n}$. So the overall error $|x(S_j) - b_j| = O(\sqrt{\Delta}\log n)$.

Overall iteration By setting the λ_j parameters for the different parts M_1, M_2, M_3, M_4 as above, it follows that in any iteration with f fractional variables, we have $\sum_{j=1}^{m} e^{-\lambda_j^2/K_0} \leq \frac{f}{24}$ which satisfies the condition in Theorem 4. □

Remark The above result also extends to the following "group sparse" setting. Suppose the constraints in M_4 are further partitioned into g groups $\{G_k\}_{k=1}^{g}$ where the column sparsity restricted to constraints in each group G_k is Δ_k. Then we obtain an integral solution with $|x(S_j) - b_j| = O(\sqrt{g\cdot\Delta_k}\,\log n)$ for all $j \in G_k$. The only modification required in the above proof is to set $\lambda_j = \sqrt{K_1\cdot g\cdot\Delta_k}/\sqrt{|S_j|}$ for $j \in G_k$.

3.2 Minimum Cost Degree Bounded Matroid Basis

The input to the *minimum cost degree bounded matroid* problem (DegMat) is a matroid defined on elements $V = [n]$ with costs $d : V \to \mathbb{Z}_+$ and m "degree constraints" $\{S_j, b_j\}_{j=1}^{m}$ where each $S_j \subseteq [n]$ and $b_j \in \mathbb{Z}_+$. The objective is to find a minimum-cost base I in the matroid that obeys all the degree bounds, i.e. $|I \cap S_j| \leq b_j$ for all $j \in [m]$. Here we make a minor technical assumption that all costs are polynomially bounded integers.

An algorithm for DegMat is said to be an $(\alpha, \beta\cdot b+\gamma)$-bicriteria approximation algorithm if for any instance, it finds a base I satisfying $|I \cap S_j| \leq \beta\cdot b_j + \gamma$ for all $j \in [m]$ and having cost at most α times the optimum (which satisfies all degree bounds).

Theorem 8 *There is a randomized algorithm for* DegMat*, that on any instance, finds a base I^* of cost at most the optimum where for each $j \in [m]$:*

$$|I^*\cap S_j| \leq O(1)\cdot\min\left\{\sqrt{j},\ \sqrt{n\log(m/n)},\ \sqrt{\log m\log n\cdot b_j} + \log m\log n,\ \sqrt{\Delta}\log n\right\}.$$

Proof Let $y \in [0,1]^n$ be an optimal solution to the natural LP relaxation of DEGMAT. We now describe the rounding algorithm: this is based on iterative applications of Theorem 5. First, we incorporate the cost as a special degree constraint $v_0 = d$ indexed zero. We will require zero violation in the cost during each iteration, i.e. $\lambda_0 = 0$ always. We partition the degree constraints $[m]$ as in Theorem 7: recall the definitions of M_1, M_2, M_3, M_4, and the setting of their λ_j parameters in each iteration.

In each iteration, we start with a fractional solution y' with $f \le n$ *fractional* variables. Using the same calculations as Theorem 7, we have $\sum_{j=0}^{m} e^{-\lambda_j^2/K_0} \le 1 + \frac{f}{24} \le \frac{f}{16}$ assuming $f \ge 48$. For now assume $f \ge \max\{K_0, 48\}$; applying Theorem 5, we obtain a new fractional solution y'' that has:

- $|\langle v_0, y'' - y' \rangle| \le \|d\|/n^{O(1)} \le \frac{1}{n}$.
- For each $j \in [m]$, $|y''(S_j) - y'(S_j)| \le \lambda_j \sqrt{W_j(y')} + \frac{1}{n}$.
- The number of fractional variables in y'' is at most $\frac{f}{K'}$ for some constant $K' > 1$.

The first condition uses the fact that the error term $\|a_j\|/n^2$ in Theorem 5 can be reduced to $\|a_j\|/n^c$ for any constant c, and that $\|d\| \le poly(n)$ as we assumed all costs to be polynomially bounded.

We repeat these iterations as long as $f \ge \max\{K_0, 48\}$: this takes $T \le \frac{\log n}{\log K'} = O(\log n)$ iterations. The violation in the cost (i.e. constraint $j = 0$) is at most $\frac{T}{n} < 1$. For any degree constraint $j \in [m]$, the violation is exactly as in Theorem 7.

At the end of the above iterations, we are left with an almost integral solution x: it has $O(1)$ fractional variables. Notice that x lies in the matroid base polytope: so it can be expressed as a convex combination of (integral) matroid bases. We output the minimum cost base I^* in this convex decomposition of x. Note that the cost of solution I^* is at most that of x which is less than $\langle d, y \rangle + 1$. Moreover, I^* agrees with x on all integral variables of x: so the worst case additional violation of any degree constraint is just $O(1)$. □

We state two special cases of this result, which improve on prior work.

Corollary 1 *There are randomized bicriteria approximation algorithms for* DEGMAT *with ratios* $(1, b + O(\sqrt{n}\log(m/n)))$ *and* $(1, O(\sqrt{\Delta}\log n))$.

Previously, [6] obtained a $(1, b + O(\sqrt{n\log(m)}))$ bicriteria approximation and [11] obtained a $(1, \Delta - 1)$ bicriteria approximation for DEGMAT.

3.3 Multi-criteria Matroid Basis

The input to the *multi-criteria matroid basis* is a matroid $\mathcal{M}$ defined on elements $V = [n]$ with k different cost functions $d_j : [n] \to \mathbb{Z}_+$ (for $j = 1, \ldots, k$) and budgets

$\{B_j\}_{j=1}^k$. The goal is to find (if possible) a basis I with $d_j(I) \leq B_j$ for each $j \in [k]$. We obtain:

Theorem 9 *There is a randomized algorithm for multi-criteria matroid basis, that given any $\epsilon > 0$ finds in $n^{O(k^{1.5}/\epsilon)}$ time, a basis I with $d_j(I) \leq (1+\epsilon)B_j$ for all $j \in [k]$.*

Previously, [8] obtained a deterministic algorithm for MCM that required $n^{O(k^2/\epsilon)}$ time. One could also use the algorithm of [6] to obtain a randomized PTAS for MCM, but this approach requires at least $n^{\Omega(k/\epsilon^2)}$ time. Our running time is better when $\epsilon < 1/\sqrt{k}$.

We now describe the algorithm in Theorem 9. An element e is said to be *heavy* if its jth cost $d_j(e) > \frac{\epsilon}{\sqrt{k}} B_j$ for any $j \in [k]$. Note that the optimal solution contains at most $\frac{k^{1.5}}{\epsilon}$ heavy elements. The algorithm first guesses by enumeration all heavy elements in the optimal solution. Let $\mathcal{M}'$ denote the matroid obtained by contracting these heavy elements. Let B'_j denote the residual budget for each $j \in [k]$. The algorithm now solves the natural LP relaxation:

$$x \in P(\mathcal{M}'), \quad \langle d_j, x \rangle \leq B'_j, \ \forall j \in [k].$$

The rounding algorithm is an iterative application of Theorem 5: the number of fractional variables decreases by a factor of $K > 1$ in each iteration.

As long as the number of fractional variables $n' < 16k$, we use $\lambda_j = 0$ for all $j \in [k]$; note that this satisfies the condition $\sum_{j=1}^k e^{-\lambda_j^2/K_0} \leq n'/16$. Note that there is no loss in any of the budget constraints in this first phase of the rounding.

Once $n' \leq N := 16k$, we choose each $\lambda_j = \sqrt{K_0 \log(N/n')}$ which satisfies the condition on λs. The loss in the jth budget constraint in such an iteration is at most $\lambda_j \sqrt{n'} \cdot d_j^{max}$ where $d_j^{max} \leq \frac{\epsilon}{\sqrt{k}} B_j$ is the maximum cost of any element. So the increase in the jth budget constraint over all iterations is at most:

$$d_j^{max} \cdot \sum_{i=0}^{t-1} \sqrt{K_0 \frac{N}{K^i} \log(K^i)} \ \leq \ O(\sqrt{N}) \cdot d_j^{max} \ = \ O(\epsilon) B_j.$$

Above i indexes iterations in the second phase of rounding.

3.4 Low Congestion Routing on Short Paths

The routing on short paths (RSP) problem is defined on an n-vertex directed graph $G = (V, E)$ with edge capacities $b : E \to \mathbb{Z}_+$. There are k source-sink pairs $\{(s_i, t_i)\}_{i=1}^k$ and a length bound Δ. The goal in RSP is to find an $s_i - t_i$ path P_i of length at most Δ for each pair $i \in [k]$ such that the number of paths using any edge e is at most b_e.

The decision problem of determining whether there exist such paths is NP-complete. Hence we focus on bicriteria approximation algorithms, where we attempt to find paths P_is that violate the edge capacities by a small amount. As noted in [6], we can use any LP-based algorithm for DegMat to obtain one for RSP: for completeness we describe this briefly below.

Let $\mathcal{P}_i$ denote the set of all $s_i - t_i$ paths of length at most Δ. Consider the following LP relaxation for RSP.

$$\sum_{P \in \mathcal{P}_i} x_{i,P} \geq 1, \qquad \forall i \in [k]$$

$$\sum_{i=1}^{k} \sum_{P \in \mathcal{P}_i : e \in P} x_{i,P} \leq b_e, \qquad \forall e \in E$$

$$x \geq 0.$$

Although this LP has an exponential number of variables, it can be solved in polynomial time by an equivalent polynomial-size formulation using a "time-expanded network".

Given any feasible instance of RSP, we obtain a fractional solution to the above LP. Moreover, the number of non-zero variables $x_{i,P}$ is at most $k + |E| = poly(n)$. Let $\mathcal{P}'_i$ denote the set of $s_i - t_i$ paths with non-zero value in this fractional solution. Consider now an instance of DegMat on groundset $U = \cup_{i=1}^{k} \mathcal{P}'_i$ where the matroid is a *partition matroid* that requires one element from each $\mathcal{P}'_i$. The degree constraints correspond to edges $e \in E$, i.e. $S_e = \{P \in U : e \in P\}$. The goal is to find a base I in the partition matroid such that $|S_e \cap I| \leq b_e$ for all $e \in E$. Note that the column sparsity of the degree constraints is Δ since each path in U has length at most Δ. Moreover $\{x_{i,P} : P \in \mathcal{P}'_i, i \in [k]\}$ is a feasible fractional solution to the LP relaxation of this DegMat instance. So we obtain:

Corollary 2 *There is an algorithm that given any feasible instance of* RSP*, computes an $s_i - t_i$ path of length at most Δ for each $i \in [k]$ where the number of paths using any edge e is at most $b_e + \min\left\{O(\sqrt{\Delta}\log n),\ O(\sqrt{b_e}\log n + \log^2 n)\right\}$.*

Multipath routing with laminar requirements Our techniques can also handle a richer set of requirements in the RSP problem. In addition to the graph G, pairs $\{(s_i, t_i)\}_{i=1}^{k}$ and length bound Δ, there is a laminar family $\mathcal{L}$ defined on the pairs $[k]$ with an integer requirement r_T on each set $T \in \mathcal{L}$. The goal in the *laminar* RSP problem is to find a multiset of $s_i - t_i$ paths (for $i \in [k]$) such that:

1. each path has length at most Δ,
2. for each $T \in \mathcal{L}$, there are at least r_T paths between pairs of T, and
3. the number of paths using any edge e is at most b_e.

Consider the following LP relaxation for this problem.

$$\sum_{i\in T}\sum_{P\in\mathcal{P}_i} x_{i,P} \geq r_T, \qquad \forall T \in \mathcal{L}$$

$$\sum_{i=1}^{k}\sum_{P\in\mathcal{P}_i:e\in P} x_{i,P} \leq b_e, \qquad \forall e \in E$$

$$x \geq 0.$$

This LP can again be solved using an equivalent polynomial-sized LP. Let $\mathcal{P}_i'$ denote the set of $s_i - t_i$ paths with non-zero value in this fractional solution, and define groundset $U = \cup_{i=1}^{k}\mathcal{P}_i'$. As before, we also define "degree constraints" corresponding to edges $e \in E$, i.e. at most b_e elements can be chosen from $S_e = \{P \in U : e \in P\}$. Unlike the usual RSP problem we can not directly cast these laminar requirements as a matroid constraint, but a slight modification of the DegMat algorithm works.

The main idea is that the partial rounding result (Theorem 5) also holds if we want to exactly preserve any laminar family $\mathcal{L}$ of constraints (instead of a matroid). Note that a laminar family on $|U|$ elements might have $2|U|$ sets. However, it is easy to see that the number of *tight* constraints of $\mathcal{L}$ at any strictly fractional solution is at most $|U|/2$. Using this observation in place of Claim 2, we obtain the partial rounding result also for laminar constraints.

Finally using this partial rounding as in Theorem 8, we obtain:

Theorem 10 *There is an algorithm that given any feasible instance of laminar* RSP, *computes a multiset $\mathcal{Q}$ of $s_i - t_i$ paths such that:*

1. *each path in $\mathcal{Q}$ has length at most Δ,*
2. *for each $T \in \mathcal{L}$, there are at least r_T paths in $\mathcal{Q}$ between pairs of T, and*
3. *the number of paths in $\mathcal{Q}$ using any edge e is at most:*

$$b_e + \min\left\{O(\sqrt{\Delta}\log n),\ O(\sqrt{b_e}\log n + \log^2 n)\right\}.$$

Appendix: Useful Linear Programming Facts

Fact 1 *Consider any polyhedron given by $P = \{x : Ax \leq b\}$ where all entries in A, b are integers of size at most $\log_2 B$. Then there is a polynomial (in $\log B$ and size of A) time algorithm that given any point $u \in P$, finds another point $v^* \in P$ where (i) $\|u - v^*\|_1 \leq \frac{1}{n^7 B}$ and (ii) all entries in v^* are rationals of size $O(n^2 \log B)$.*

Proof Let $L := 2n^8 B$ and u' denote the point with coordinates $u_i' = \frac{1}{L}\lfloor L \cdot u_i \rfloor$ for all $i \in [n]$. We now write a linear program that computes the point $v \in P$ with minimum ℓ_1 distance from u'.

$$\begin{aligned}
\min\ & \textstyle\sum_{i=1}^{n} d_i \\
\text{s.t.}\ & Av \le b \\
& |Lv_i - Lu'_i| \le Ld_i \ \forall i \in [n] \\
& \textstyle\sum_{i=1}^{n} d_i \le 1 \\
& \mathbf{d}, \mathbf{v} \in \mathbb{R}^n.
\end{aligned}$$

Note that the feasible region of this LP is a polytope (bounded polyhedron) due to the last two constraints. So there is an optimal extreme point solution v^* that can be found in polynomial time. Since all constraint coefficients in this LP are integers bounded by L, the entries in v^* must be rationals bounded by $(2nL)^{2n}$. Finally, $u \in P$ corresponds to a feasible solution to this LP with $v = u, d_i = |v_i - u_i|$ (for $i \in [n]$) and objective $\|u - u'\|_1 \le \frac{n}{L}$. It now follows that $\|v^* - u'\|_1 \le \frac{n}{L}$ and so $\|v^* - u\|_1 \le \|v^* - u'\|_1 + \|u' - u\|_1 \le \frac{2n}{L}$. □

Fact 2 *Consider any linear subspace given by $\{x : Ax = 0\}$ where all entries in A are integers of size at most $\log_2 B$. Then there is a polynomial (in $\log B$ and size of A) time algorithm that computes a basis $\{b_j\}_{j=1}^{k}$ of this subspace where (i) all entries are rationals of size $O(n^2 \log B)$, (ii) $|\langle b_j, b_j\rangle - 1| \le \frac{1}{n^4 B}$ for all $j \in [k]$, and (iii) $|\langle b_j, b_\ell\rangle| \le \frac{1}{n^4 B}$ for all $j \ne \ell, j, \ell \in [k]$.*

Proof We can obtain an orthonormal basis $\{b'_j\}_{j=1}^{k}$ of this subspace using Gaussian elimination and Gram-Schmidt orthogonalization. This clearly satisfies the last two conditions. But some more work is needed since we require the entries in the basis vectors to be bounded integers.

To ensure this, we modify each vector b'_j into b_j separately by applying Fact 1 with polyhedron $P = \{x : Ax = 0\}$, $u = b'_j$, and then set $b_j = v^*$. Now the last condition follows from Fact 1(ii). The first and second conditions follow from Fact 1(i) since $\{b'_j\}_{j=1}^{k}$ is orthonormal. □

References

1. N. Bansal, Constructive algorithms for discrepancy minimization, in *Foundations of Computer Science (FOCS)*, 2010, pp. 3–10
2. N. Bansal, M. Charikar, R. Krishnaswamy, S. Li, Better algorithms and hardness for broadcast scheduling via a discrepancy approach, in *SODA*, 2014, pp, 55–71
3. N. Bansal, R. Khandekar, J. Könemann, V. Nagarajan, B. Peis, On generalizations of network design problems with degree bounds. Math. Program. **141**(1–2), 479–506 (2013)
4. J. Beck, T. Fiala, Integer-making theorems. Discret. Appl. Math. **3**, 1–8 (1981)
5. M. Charikar, A. Newman, A. Nikolov, Tight hardness results for minimizing discrepancy, in *SODA*, 2011, pp. 1607–1614
6. C. Chekuri, J. Vondrak, R. Zenklusen, Dependent randomized rounding via exchange properties of combinatorial structures, in *FOCS*, 2010, pp. 575–584
7. D.A. Freedman, On tail probabilities for martingales. Ann. Probab. **3**, 100–118 (1975)
8. F. Grandoni, R. Ravi, M. Singh, R. Zenklusen, New approaches to multi-objective optimization. Math. Program. **146**(1–2), 525–554 (2014)

9. N.J.A. Harvey, R. Schwartz, M. Singh, Discrepancy without partial colorings, in *APPROX/RANDOM*, 2014, pp. 258–273
10. R.M. Karp, F.T. Leighton, R.L. Rivest, C.D. Thompson, U.V. Vazirani, V.V. Vazirani, Global wire routing in two-dimensional arrays. Algorithmica **2**, 113–129 (1987)
11. T. Király, L.C. Lau, M. Singh, Degree bounded matroids and submodular flows, in *IPCO*, 2008, pp. 259–272
12. L.-C. Lau, R. Ravi, M. Singh, *Iterative Methods in Combinatorial Optimization* (Cambridge University Press, Cambridge, 2011)
13. F.T. Leighton, C.J. Lu, S. Rao, A. Srinivasan, New algorithmic aspects of the local lemma with applications to routing and partitioning. SIAM J. Comput. **31**(2), 626–641 (2001)
14. L. Lovasz, J. Spencer, K. Vesztergombi, Discrepancy of set-systems and matrices. Eur. J. Comb. **7**, 151–160 (1986)
15. S. Lovett, R. Meka, Constructive discrepancy minimization by walking on the edges, in *FOCS*, 2012, pp. 61–67
16. J. Matoušek, *Geometric Discrepancy: An Illustrated Guide* (Springer, Berlin/Heidelberg, 2010)
17. A. Nikolov, K. Talwar, Approximating hereditary discrepancy via small width ellipsoids, in *Symposium on Discrete Algorithms, SODA*, 2015, pp. 324–336
18. T. Rothvoss, Approximating bin packing within o(log OPT * log log OPT) bins, in *FOCS*, 2013, pp. 20–29
19. T. Rothvoss, Constructive discrepancy minimization for convex sets, in *IEEE Symposium on Foundations of Computer Science, FOCS*, 2014, pp. 140–145
20. A. Schrijver, *Combinatorial Optimization* (Springer, Berlin, 2003)
21. M. Singh, L.C. Lau, Approximating minimum bounded degree spanning trees to within one of optimal, in *STOC*, 2007, pp. 661–670
22. J. Spencer, Six standard deviations suffice. Trans. Am. Math. Soc. **289**(2), 679–706 (1985)
23. A. Srinivasan, Improving the discrepancy bound for sparse matrices: better approximations for sparse lattice approximation problems, in *Symposium on Discrete Algorithms (SODA)*, 1997, pp. 692–701
24. V.V. Vazirani, *Approximation Algorithms* (Springer, New York, 2001)
25. D. Williamson, D. Shmoys, *The Design of Approximation Algorithms* (Cambridge University Press, Cambridge, 2011)

A Tverberg Type Theorem for Matroids

Imre Bárány, Gil Kalai, and Roy Meshulam

In memory of Jirka Matousek

Abstract Let $b(M)$ denote the maximal number of disjoint bases in a matroid M. It is shown that if M is a matroid of rank $d+1$, then for any continuous map f from the matroidal complex M into $\mathbb{R}^d$ there exist $t \geq \sqrt{b(M)}/4$ disjoint independent sets $\sigma_1, \ldots, \sigma_t \in M$ such that $\bigcap_{i=1}^{t} f(\sigma_i) \neq \emptyset$.

1 Introduction

Tverberg's theorem [15] asserts that if $V \subset \mathbb{R}^d$ satisfies $|V| \geq (k-1)(d+1)+1$, then there exists a partition $V = V_1 \cup \cdots \cup V_k$ such that $\bigcap_{i=1}^{k} \operatorname{conv}(V_i) \neq \emptyset$. Tverberg's theorem and some of its extensions may be viewed in the following general context. For a simplicial complex X and $d \geq 1$, let the *affine Tverberg number* $T(X,d)$ be the maximal t such that for any affine map $f : X \to \mathbb{R}^d$, there exist disjoint simplices $\sigma_1, \ldots, \sigma_t \in X$ such that $\bigcap_{i=1}^{t} f(\sigma_i) \neq \emptyset$. The *topological Tverberg number* $TT(X,d)$ is defined similarly where now $f : X \to \mathbb{R}^d$ can be an arbitrary continuous map.

Let Δ_n denote the n-simplex and let $\Delta_n^{(d)}$ be its d-skeleton. Using the above terminology, Tverberg's theorem is equivalent to $T(\Delta_{(k-1)(d+1)}, d) = k$ which

I. Bárány (✉)
Rényi Institute, Hungarian Academy of Sciences, POB 127, 1364 Budapest, Hungary

Department of Mathematics, University College London, Gower Street, London, UK
e-mail: barany@renyi.hu

G. Kalai
Einstein Institute of Mathematics, Hebrew University, 9190 Jerusalem, Israel
e-mail: kalai@math.huji.ac.il

R. Meshulam
Department of Mathematics, Technion, 32000 Haifa, Israel
e-mail: meshulam@math.technion.ac.il

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_5

is clearly the same as $T(\Delta^{(d)}_{(k-1)(d+1)}, d) = k$. Similarly, the topological Tverberg theorem of Bárány, Shlosman and Szűcs [2] states that if p is prime then $TT(\Delta_{(p-1)(d+1)}, d) = p$. Schöneborn and Ziegler [14] proved that this implies the stronger statement $TT(\Delta^{(d)}_{(p-1)(d+1)}, d) = p$. This result was extended by Özaydin [13] for the case when p is a prime power. The question whether the topological Tverberg theorem holds for every p that is not a prime power had been open for long. Very recently, and quite surprisingly, Frick [7] has constructed a counterexample for every non-prime power p. His construction is built on work by Mabillard and Wagner [10]. See also [4] and [1] for further counterexamples.

There is a colourful version of Tverberg theorem. To state it let $n = r(d+1) - 1$ and assume that the vertex set V of Δ_n is partitioned into $d+1$ classes (called colours) and that each colour class contains exactly r vertices. We define $Y_{r,d}$ as the subcomplex of Δ_n (or $\Delta^{(d)}_n$) consisting of those $\sigma \subset V$ that contain at most one vertex from each colour class. The colourful Tverberg theorem of Živaljević and Vrećica [16] asserts that $TT(Y_{2p-1,d}, d) \geq p$ for prime p which implies that $TT(Y_{4k-1,d}, d) \geq k$ for arbitrary k. A neat and more recent theorem of Blagojević, Matschke, and Ziegler [5] says that $TT(Y_{r,d}, d) = r$ if $r + 1$ is a prime, which is clearly best possible. Further information on Tverberg's theorem can be found in Matoušek's excellent book [11].

Let M be a matroid (possibly with loops) with rank function ρ on the set V. We identify M with the simplicial complex on V whose simplices are the independent sets of M. It is well known (see e.g. Theorem 7.8.1 in [3]) that M is $(\rho(V) - 2)$-connected. Note that both $\Delta^{(d)}_n$ and $Y_{r,d}$ are matroids of rank $d + 1$. In this note we are interested in bounding $TT(M, d)$ for a general matroidal complex M. Let $b(M)$ denote the maximal number of pairwise disjoint bases in M. Our main result is the following

Theorem 1 *Let M be a matroid of rank $d + 1$. Then*

$$TT(M, d) \geq \sqrt{b(M)}/4 \ .$$

In Sect. 2 we give a lower bound on the topological connectivity of the deleted join of matroids. In Sect. 3 we use this bound and the approach of [2, 16] to prove Theorem 1.

2 Connectivity of Deleted Joins of Matroids

We recall some definitions. For a simplicial complex Y on a set V and an element $v \in V$ such that $\{v\} \in Y$, denote the *star* and *link* of v in Y by

$$\mathrm{st}(Y, v) = \{\sigma \subset V : \{v\} \cup \sigma \in Y\}$$
$$\mathrm{lk}(Y, v) = \{\sigma \in \mathrm{st}(Y, v) : v \notin \sigma\}.$$

For a subset $V' \subset V$ let $Y[V'] = \{\sigma \subset V' : \sigma \in Y\}$ be the induced complex on V'. We regard $\mathrm{st}(Y, v)$, $\mathrm{lk}(Y, v)$ and $Y[V']$ as complexes on the original set V (keeping in mind that not all elements of V have to be vertices of these complexes). Let $f_i(Y)$ denote the number of i-simplices in Y. Let $X_1, \ldots, X_k$ be simplicial complexes on the same set V and let $V_1, \ldots, V_k$ be k disjoint copies of V with bijections $\pi_i : V \to V_i$. The *join* $X_1 * \cdots * X_k$ is the simplicial complex on $\bigcup_{i=1}^k V_i$ with simplices $\bigcup_{i=1}^k \pi_i(\sigma_i)$ where $\sigma_i \in X_i$. The *deleted join* $(X_1 * \cdots * X_k)_\Delta$ is the subcomplex of the join consisting of all simplices $\bigcup_{i=1}^k \pi_i(\sigma_i)$ such that $\sigma_i \cap \sigma_j = \emptyset$ for $1 \leq i \neq j \leq k$. When all X_i are equal to X, we denote their deleted join by X_Δ^{*k}. Note that $\mathbb{Z}_k$ acts freely on X_Δ^{*k} by cyclic shifts.

Claim 2 *Let $M_1, \ldots, M_k$ be matroids on the same set V, with rank functions $\rho_1, \ldots, \rho_k$. Suppose $A_1, \ldots, A_k$ are disjoint subsets of V such that A_i is a union of at most m independent sets in M_i. Then $Y = (M_1 * \cdots * M_k)_\Delta$ is $(\lceil \frac{1}{m+1} \sum_{i=1}^k |A_i| \rceil - 2)$-connected.*

Proof Let $c = \lceil \frac{1}{m+1} \sum_{i=1}^k |A_i| \rceil - 2$. If $k = 1$ then $\rho_1(V) \geq \left\lceil \frac{|A_1|}{m} \right\rceil$ and hence $Y = M_1$ is $(\left\lceil \frac{|A_1|}{m} \right\rceil - 2)$-connected. For $k \geq 2$ we establish the Claim by induction on $f_0(Y) = \sum_{i=1}^k f_0(M_i)$. If $f_0(Y) = 0$ then all A_i's are empty and the Claim holds. We henceforth assume that $f_0(Y) > 0$ and consider two cases:

(a) If $M_i = M_i[A_i]$ for all $1 \leq i \leq k$ then $Y = M_1 * \cdots * M_k$ is a matroid of rank

$$\sum_{i=1}^k \rho_i(V) \geq \sum_{i=1}^k \left\lceil \frac{|A_i|}{m} \right\rceil \geq \left\lceil \frac{\sum_{i=1}^k |A_i|}{m} \right\rceil.$$

Hence Y is $(\left\lceil \frac{\sum_{i=1}^k |A_i|}{m} \right\rceil - 2)$-connected.

(b) Otherwise there exists an $1 \leq i_0 \leq k$ such that $M_{i_0} \neq M_{i_0}[A_{i_0}]$. Choose an element $v \in V - A_{i_0}$ such that $\{v\} \in M_{i_0}$. Without loss of generality we may assume that $i_0 = 1$ and that $v \notin \bigcup_{i=1}^{k-1} A_i$. Let $S = \bigcup_{i=1}^k V_i$ and let $Y_1 = Y[S - \{\pi_1(v)\}]$, $Y_2 = \mathrm{st}(Y, \pi_1(v))$. Then

$$Y_1 = (M_1[V - \{v\}] * M_2 * \cdots * M_k)_\Delta.$$

Noting that $f_0(Y_1) = f_0(Y) - 1$ and applying the induction hypothesis to the matroids $M_1[V - \{v\}], M_2, \ldots, M_k$ and the sets $A_1, \ldots, A_k$, it follows that Y_1 is c-connected. We next consider the connectivity of $Y_1 \cap Y_2$. Write $A_1 = \bigcup_{j=1}^t C_j$ where $t \leq m$, $C_j \in M_1$ for all $1 \leq j \leq t$, and the C_j's are pairwise disjoint. Since $\{v\} \in M_1$, it follows that there exist $\{C_j'\}_{j=1}^t$ such that $C_j' \subset C_j$, $|C_j'| \geq |C_j| - 1$, and $C_j' \in \mathrm{lk}(M_1, v)$ for all $1 \leq j \leq t$. Let

$$M_i' = \begin{cases} \mathrm{lk}(M_1, v) & i = 1, \\ M_i[V - \{v\}] & 2 \leq i \leq k, \end{cases}$$

and

$$A'_i = \begin{cases} \bigcup_{j=1}^t C'_j & i = 1, \\ A_i & 2 \le i \le k-1, \\ A_k - \{v\} & i = k. \end{cases}$$

Observe that

$$Y_1 \cap Y_2 = \mathrm{lk}(Y, \pi_1(v)) = (M'_1 * \cdots * M'_k)_\Delta$$

and that A'_i is a union of at most m independent sets in M'_i for all $1 \le i \le k$. Noting that $f_0(Y_1 \cap Y_2) \le f_0(Y) - 1$ and applying the induction hypothesis to the matroids $M'_1, \dots, M'_k$ and the sets $A'_1, \dots, A'_k$, it follows that $Y_1 \cap Y_2$ is c'-connected where

$$\begin{aligned} c' &= \left\lceil \frac{1}{m+1} \sum_{i=1}^{k} |A'_i| \right\rceil - 2 \\ &= \left\lceil \frac{1}{m+1} \left(\sum_{j=1}^{t} |C'_j| + \sum_{i=2}^{k-1} |A_i| + |A_k - \{v\}| \right) \right\rceil - 2 \\ &\ge \left\lceil \frac{1}{m+1} \left(|A_1| - m + \sum_{i=2}^{k-1} |A_i| + |A_k| - 1 \right) \right\rceil - 2 = c - 1. \end{aligned}$$

As Y_1 is c-connected, Y_2 is contractible and $Y_1 \cap Y_2$ is $(c-1)$-connected, it follows that $Y = Y_1 \cup Y_2$ is c-connected. □

Let M be a matroid on V with $b(M) = b$ disjoint bases $B_1, \dots, B_b$. Let $I_1 \cup \cdots \cup I_k$ be a partition of $[b]$ into almost equal parts $\lfloor \frac{b}{k} \rfloor \le |I_i| \le \lceil \frac{b}{k} \rceil$. Applying Claim 2 with $M_1 = \cdots = M_k = M$ and $A_i = \cup_{j \in I_i} B_j$, we obtain:

Corollary 3 *The connectivity of M_Δ^{*k} is at least*

$$\frac{b\rho(V)}{\lceil \frac{b}{k} \rceil + 1} - 2\ .$$

We suggest the following:

Conjecture 4 *For any $k \ge 1$ there exists an $f(k)$ such that if $b(M) \ge f(k)$ then M_Δ^{*k} is $(k\rho(V) - 2)$-connected.*

Remark Let M be the rank 1 matroid on m points $M = \Delta_{m-1}^{(0)}$. The chessboard complex $C(k, m)$ is the k-fold deleted join M_Δ^{*k}. Chessboard complexes play a key role in the works of Živaljević and Vrećica [16] and Blagojević, Matschke, and Ziegler [5] on the colourful Tverberg theorem. Let $k \ge 2$. Garst [9] and Živaljević and Vrećica [16] proved that $C(k, 2k-1)$ is $(k-2)$-connected. On the other hand, Friedman and Hanlon [8] showed that $\tilde{H}_{k-2}(C(k,2k-2);\mathbb{Q}) \neq 0$, so $C(k,2k-2)$ is not $(k-2)$-connected. This implies that the function $f(k)$ in Conjecture 4 must satisfy $f(k) \geq 2k-1$.

3 A Tverberg Type Theorem for Matroids

We recall some well-known topological facts (see [2]). For $m \geq 1, k \geq 2$ we identify the sphere $S^{m(k-1)-1}$ with the space

$$\left\{ (y_1,\dots,y_k) \in (\mathbb{R}^m)^k : \sum_{i=1}^{k} |y_i|^2 = 1\ ,\ \sum_{i=1}^{k} y_i = 0 \in \mathbb{R}^m \right\} \quad .$$

The cyclic shift on this space defines a $\mathbb{Z}_k$ action on $S^{m(k-1)-1}$. The action is free for prime k.

The *k-fold deleted product* of a space X is the $\mathbb{Z}_k$-space given by

$$X_D^k = X^k - \{(x,\dots,x) \in X^k : x \in X\}\ .$$

For $m \geq 1$ define a $\mathbb{Z}_k$-map

$$\phi_{m,k} : (\mathbb{R}^m)_D^k \to S^{m(k-1)-1}$$

by

$$\phi_{m,k}(x_1,\dots,x_k) = \frac{(x_1 - \frac{1}{k}\sum_{i=1}^{k} x_i, \dots, x_k - \frac{1}{k}\sum_{i=1}^{k} x_i)}{(\sum_{j=1}^{k} |x_j - \frac{1}{k}\sum_{i=1}^{k} x_i|^2)^{1/2}}\ .$$

We'll also need the following result of Dold [6] (see also Theorem 6.2.6 in [12]):

Theorem 5 (Dold) *Let p be a prime and suppose X and Y are free $\mathbb{Z}_p$-spaces such that* $\dim Y = k$ *and X is k-connected. Then there does not exist a $\mathbb{Z}_p$-map from X to Y.*

Proof of Theorem 1 Let M be a matroid on the vertex set V, and let $f : M \to \mathbb{R}^d$ be a continuous map. Let $b = b(M)$ and choose a prime $\sqrt{b}/4 \leq p \leq \sqrt{b}/2$. We'll show that there exist disjoint simplices (i.e. independent sets) $\sigma_1,\dots,\sigma_p \in M$ such that $\bigcap_{i=1}^{p} f(\sigma_i) \neq \emptyset$. Suppose for contradiction that $\bigcap_{i=1}^{p} f(\sigma_i) = \emptyset$ for all such choices of σ_i's. Then f induces a continuous $\mathbb{Z}_p$-map

$$f_* : M_\Delta^{*p} \to (\mathbb{R}^{d+1})_D^p$$

as follows. If $x_1, \ldots, x_p$ have pairwise disjoint supports in M and $(t_1, \ldots, t_p) \in \mathbb{R}_+^p$ satisfies $\sum_{i=1}^p t_i = 1$ then

$$f_*(t_1\pi_1(x_1) + \cdots + t_p\pi_p(x_p)) = (t_1, t_1 f(x_1), \ldots, t_p, t_p f(x_p)) \in (\mathbb{R}^{d+1})_D^p \ .$$

Hence $\phi_{d+1,p} f_*$ is a $\mathbb{Z}_p$-map between the free $\mathbb{Z}_p$-spaces M_Δ^{*p} and $S^{(d+1)(p-1)-1}$. This however contradicts Dold's Theorem since by Corollary 3 the connectivity of M_Δ^{*p} is at least

$$\frac{b(d+1)}{\lceil \frac{b}{p} \rceil + 1} - 2 \geq (d+1)(p-1) - 1$$

by the choice of p.

□

Acknowledgements Research of Imre Bárány was partially supported by ERC advanced grant 267165, and by Hungarian National grant K 83767. Research of Gil Kalai was supported by ERC advanced grant 320924. Research of Roy Meshulam is supported by ISF and GIF grants.

References

1. S. Avvakumov, I. Mabillard, A. Skopenkov, U. Wagner, Eliminating higher-multiplicity intersections, III. Codimension 2 (2015), 16 pp., arXiv:1511.03501
2. I. Bárány, S. Shlosman, A. Szűcs, On a topological generalization of a theorem of Tverberg. J. Lond. Math. Soc. **23**, 158–164 (1981)
3. A. Björner, Topological methods, in *Handbook of Combinatorics, 1819–1872*, ed. by R. Graham, M. Grötschel, L. Lovász (North-Holland, Amsterdam, 1995)
4. P.V.M. Blagojević, F. Frick, G.M. Ziegler, Barycenters of polytope Skeleta and counterexamples to the topological Tverberg conjecture, via constraints (2015), 6 pp., arXiv:1508.02349
5. P.V.M. Blagojević, B. Matschke, G.M. Ziegler, Optimal bounds for the colored Tverberg problem. J. European Math. Soc. **17**, 739–754 (2015)
6. A. Dold, Simple proofs of some Borsuk-Ulam results. Contemp. Math. **19**, 65–69 (1983)
7. F. Frick, Counterexamples to the topological Tverberg conjecture (2015), 3 pp., arXiv: 1502.00947
8. J. Friedman, P. Hanlon, On the Betti numbers of chessboard complexes, J. Algebraic Combin. **8**, 193–203 (1998)
9. P. Garst, Cohen-Macaulay complexes and group actions, Ph.D.Thesis, The University of Wisconsin – Madison, 1979
10. I. Mabillard, U. Wagner, Eliminating higher-multiplicity intersections, I. A Whitney trick for Tverberg-type problems (2015), 46 pp., arXiv:1508.02349
11. J. Matoušek, *Lectures on Discrete Geometry* (Springer, New York, 2002)
12. J. Matoušek, *Using the Borsuk-Ulam Theorem* (Springer, Berlin, 2003)
13. M.Özaydin, Equivariant maps for the symmetric group, 1987. Available at http://minds.wisconsin.edu/handle/1793/63829

14. T. Schöneborn, G.M. Ziegler, The topological Tverberg theorem and winding numbers. J. Combin. Theory Ser. A **112**, 82–104 (2005)
15. H. Tverberg, A generalization of Radon's theorem. J. Lond. Math. Soc. **41**, 123–128 (1966)
16. R. Živaljević, S. Vrećica, The colored Tverberg's problem and complexes of injective functions. J. Combin. Theory Ser. A **61**, 309–318 (1992)

Gershgorin Disks for Multiple Eigenvalues of Non-negative Matrices

Imre Bárány and József Solymosi

Dedicated to the memory of Jiří Matoušek

Abstract Gershgorin's famous circle theorem states that all eigenvalues of a square matrix lie in disks (called Gershgorin disks) around the diagonal elements. Here we show that if the matrix entries are non-negative and an eigenvalue has geometric multiplicity at least two, then this eigenvalue lies in a smaller disk. The proof uses geometric rearrangement inequalities on sums of higher dimensional real vectors which is another new result of this paper.

1 Introduction and Main Result

Gershgorin's circle theorem [4] is a fundamental and widely used result on localizing the eigenvalues of square matrices. It states that all eigenvalues are in disks (called Gershgorin disks) around the diagonal elements.

The main goal of this paper is to improve Gershgorin's theorem under special conditions, namely, when the matrix is non-negative and has a multiple eigenvalue. We show that such an eigenvalue lies in disks of smaller radius around a diagonal element. For the proof we establish various geometric inequalities concerning rearrangements of vector sums. This is an interesting connection between convex geometry and matrix theory. The geometric point of view in eigenvalue problems is certainly not new but this particular connection seems to be new.

I. Bárány
Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences, P.O. Box 127,
1364 Budapest, Hungary

Department of Mathematics, University College London, Gower Street, WC1E 6BT London, UK

J. Solymosi (✉)
Department of Mathematics, University of British Columbia, 1984 Mathematics Road,
V6T 1Z2 Vancouver, BC, Canada
e-mail: Solymosi@math.ubc.ca

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_6

Here we show that if the matrix entries are non-negative and an eigenvalue has geometric multiplicity at least two, then this eigenvalue lies in a smaller disk.

Let $D(a, r)$ denote the disk with center a and radius r on the complex plane:

$$D(a, r) = \{x \in \mathbb{C} : |x - a| \leqslant r\} .$$

For an $n \times n$ complex matrix, $A = [a_{ij}]$, the Gershgorin disks are $D(a_{ii}, R_i)$ where $R_i = \sum_{j:i \neq j} |a_{ij}|$. The most commonly cited form of Gershgorin's theorem says that every eigenvalue of A lies in some $D(a_{ii}, R_i)$. Varga's nice book *Gershgorin and His Circles* [15] surveys various applications and extensions of this important theorem. An interesting and recent theorem of Marsli and Hall [7] states that if an eigenvalue of a matrix A has geometric multiplicity k, then it lies in at least k of the Gershgorin disks of A. They have extended this result in subsequent papers [3, 8–10]. Here we focus on the $k = 2$ case for non-negative matrices.

Understanding the spectra of a matrix is a central question both in applied and pure mathematics. Here are some facts and results. There are two particular eigenvalues for which the multiplicity is of great importance; the largest eigenvalue which determines the spectral radius of the matrix and the multiplicity of the eigenvalue "0" since it determines the rank of the matrix. There are also applications using the smallest eigenvalue. For example Roy shows in [12] that the Euclidean representation number of a graph is closely related to the multiplicity of the smallest eigenvalue. The multiplicity of the largest and the second largest eigenvalues play a key role in some numerical methods. Del Corso [2] considers the problem of approximating an eigenvector belonging to the largest eigenvalue by the so called power method. It is proved that the rate of convergence depends on the ratio of the two largest eigenvalues and on their multiplicities. The rate increases with the multiplicity of the largest eigenvalue and decreases with the multiplicity of the second eigenvalue. In graph theory the Colin de Verdière number is the multiplicity of the second largest eigenvalue of the adjacency matrix, maximized by weighting the edges and nodes. For more details and the exact definition we refer to the papers [6] and [14].

Gershgorin's circle theorem is intertwined with the Perron–Frobenius theory. It is one of the tools used to bound the spectral radius of a matrix. It follows from the Perron–Frobenius theorem that the largest magnitude eigenvalue of any non-negative matrix is a positive real number, see in e.g. [1].

Let us define the *half Gershgorin* disks, $D(a_{ii}, r_i)$, which are subsets of the original. Instead of $R_i = \sum_{j:i \neq j} |a_{ij}|$ we take the partial sum of the $\lfloor n/2 \rfloor$ largest terms. This sum is denoted by r_i.

Recall that the *geometric multiplicity* of an eigenvalue λ of A is the dimension of the corresponding eigenspace of A, that is, the kernel of $A - \lambda I$. (Its algebraic multiplicity is the multiplicity of the root λ of the polynomial $\det(A - xI)$.)

We are going to show that multiple geometric eigenvalues are in smaller Gershgorin disks when the matrix is non-negative.

Theorem 1 *Let $A = \{a_{ij}\}$ be an $n \times n$ non-negative (real) matrix and λ an eigenvalue of A with geometric multiplicity at least two. Then λ is in a half Gershgorin disk, $D(a_{ii}, r_i)$, for some i.*

Actually we are going to prove that such an eigenvalue lies in the disk $D(a_{ii}, r)$ and various values of r for some suitable i. The proofs are based on geometric estimates that are of independent interest. They are given in the next section.

2 Rearrangement Inequalities for Vectors

Assume $V = \{v_1, \dots, v_n\} \subset \mathbb{R}^d$ and $\sum_1^n v_i = 0$. Further, let $\alpha_1 \geq \dots \geq \alpha_n \geq 0$ be real numbers. We write $[n]$ for the set $\{1, \dots, n\}$.

Theorem 2 *Under the above conditions set $\beta = \alpha_{\lfloor n/2 \rfloor + 1}$. Then for every permutation σ of $[n]$*

$$\left\| \sum_1^n \alpha_i v_{\sigma(i)} \right\| \leq \max_{i \in [n]} \|v_i\| \sum_1^n |\alpha_i - \beta|.$$

Corollary 1 *Under the above conditions, for every permutation σ of $[n]$*

$$\left\| \sum_1^n \alpha_i v_{\sigma(i)} \right\| \leq \max_{i \in [n]} \|v_i\| \sum_1^{\lfloor n/2 \rfloor} \alpha_i.$$

In the second geometric estimate we need a technical assumption.

Theorem 3 *Let $V = \{v_1, \dots, v_n\} \subset \mathbb{R}^d$ satisfy the previous assumption. Suppose further that the v_i are ordered with decreasing (Euclidean) length, that is, $\|v_1\| \geq \dots \geq \|v_n\|$. Let $\gamma \in [\alpha_{j+1}, \alpha_j]$ for some $j \in [n-1]$. Then for every permutation σ of $[n]$*

$$\left\| \sum_1^n \alpha_i v_{\sigma(i)} \right\| \leq \sum_1^j \alpha_i \|v_i\| - \frac{\gamma}{2} \left[\sum_1^j \|v_i\| - \sum_{j+1}^n \|v_i\| \right].$$

Here of course one wants to choose j and γ so that the right hand side is as small as possible. When $j = \lceil n/2 \rceil$, the sum between the brackets is non-negative. Choosing any γ from the interval $[\alpha_{j+1}, \alpha_j]$ gives the following.

Corollary 2 *Under the above conditions for every permutation σ of $[n]$*

$$\left\| \sum_1^n \alpha_i v_{\sigma(i)} \right\| \leq \sum_1^{\lceil n/2 \rceil} \alpha_i \|v_i\|.$$

We mention that the estimates in Theorems 2 and 3 are incomparable; sometimes the first, other times the second gives the better bound.

3 Proof of the Rearrangement Inequalities

Proof of Theorem 2 First fix some $\gamma \geq 0$. Then

$$\sum_1^n \alpha_i v_{\sigma(i)} = \sum_1^n \alpha_i v_{\sigma(i)} - \sum_1^n \gamma v_{\sigma(i)} = \sum_1^n (\alpha_i - \gamma) v_{\sigma(i)}.$$

By the triangle inequality

$$\left\| \sum_1^n \alpha_i v_{\sigma(i)} \right\| \leq \max_{i \in [n]} \|v_i\| \sum_1^n |\alpha_i - \gamma|.$$

Set $k = \lfloor n/2 \rfloor$ and define $\beta = \alpha_{k+1}$. It can be proven that the function $\gamma \to \sum_1^n |\alpha_i - \gamma|$ takes its minimum at $\gamma = \beta$ when n is odd, and at every γ from the interval $[\alpha_{k+1}, \alpha_k]$ when n is even. □

Corollary 1 follows immediately since with the above k and β

$$\begin{aligned}\sum_1^n |\alpha_i - \beta| &= \sum_1^k (\alpha_i - \beta) + \sum_{k+1}^n (\beta - \alpha_i) \\ &= \sum_1^k \alpha_i - \sum_\ell^n \alpha_i \leq \sum_1^k \alpha_i\end{aligned}$$

where ℓ equals $k+1$ for even n and $k+2$ for odd n.

Proof of Theorem 3 The zonotope $Z(V)$ spanned by V is, by definition, the set

$$Z(V) = \left\{ \sum_{i \in [n]} \xi_i v_i : 0 \leq \xi_i \leq 1 \ (\forall i) \right\}.$$

Let B denote the Euclidean unit ball of $\mathbb{R}^d$. We claim first that

$$Z(V) \subset \frac{1}{2}\Big(\|v_1\| + \cdots + \|v_n\|\Big)B. \tag{1}$$

It is well-known [11] and easy to check that $Z(V)$ is the convex hull of the points $s(W) = \sum_{v \in W} v$ where $W \subset V$. Thus it suffices to show that for every $W \subset V$, $\|s(W)\| \leq \frac{1}{2}(\|v_1\| + \ldots + \|v_n\|)$. Fix $U \subset V$ such that $s(U)$ has maximal length among all $s(W)$. Set $z = s(U)$ and observe that $-z = s(V \setminus U)$ as $s(V) = 0$. Since $\|z\| = \|-z\|$ evidently, we have

$$2\|z\| = \|z\| + \|-z\| = \|s(U)\| + \|s(V \setminus U)\| \le \sum_1^n \|v_i\|$$

by the triangle inequality. This implies that $\|z\| \le \frac{1}{2}\sum_1^n \|v_i\|$.

We observe next that

$$\sum_1^n \alpha_i v_{\sigma(i)} = \sum_1^j (\alpha_i - \gamma) v_{\sigma(i)} + \sum_1^j \gamma v_{\sigma(i)} + \sum_{j+1}^n \alpha_i v_{\sigma(i)}$$
$$= \sum_1^j (\alpha_i - \gamma) v_{\sigma(i)} + \gamma \left[\sum_1^j v_{\sigma(i)} + \sum_{j+1}^n \frac{\alpha_i}{\gamma} v_{\sigma(i)} \right].$$

The expression between the brackets is a vector u in $Z(V)$ so $\|u\| \le \frac{1}{2}\sum_1^n \|v_i\|$. By the triangle inequality the norm of $\sum_1^n \alpha_i v_{\sigma(i)}$ is at most

$$\sum_1^j (\alpha_i - \gamma)\|v_{\sigma(i)}\| + \gamma\|u\| \le \sum_1^j (\alpha_i - \gamma)\|v_i\| + \frac{\gamma}{2}\sum_1^n \|v_i\|$$
$$= \sum_1^j \alpha_i \|v_i\| - \frac{\gamma}{2}\left[\sum_1^j \|v_i\| - \sum_{j+1}^n \|v_i\| \right].$$

□

4 Proof of Theorem 1

We first recall the simple proof of Gershgorin's original theorem. Let $v = (v_1, \dots, v_n)$ be an eigenvector with eigenvalue λ where v_i are complex numbers. Assume $|v_i| = \max_{j \in [n]} |v_j|$. Then $\sum_{j=1}^n a_{ij} v_j = \lambda v_i$ implying

$$(\lambda - a_{ii}) v_i = \sum_{j:j \ne i} a_{ij} v_j. \tag{2}$$

Taking absolute value on both sides and using $|v_i| \ge |v_j|$ shows that $\lambda \in D(a_{ii}, R_i)$ with $R_i = \sum_{j:j \ne i} a_{ij}$ indeed.

When the eigenvalue λ has geometric multiplicity at least two, then its eigenspace contains a nonzero vector $v = (v_1, \dots, v_n)$ whose components sum to zero: $\sum_1^n v_i = 0$. Indeed, let u and w be two linearly independent eigenvectors from the eigenspace of λ. If $\sum_1^n u_i = 0$, then $v = u$ is a suitable eigenvector. If not, then $v = \left(\sum_1^n w_i\right) u - \left(\sum_1^n u_i\right) w$ has the required property.

As any multiplier of v is still an eigenvector, we can suppose that the largest magnitude component of v, v_i, is a positive real number. Actually we can and do assume that $v_i = 1$. Then the other components, v_j, are complex numbers with $|v_j| \le 1$.

The **proof** of Theorem 1 is based on equation (2) plus the condition that $\sum_1^n v_j = 0$. As $\mathbb{C}$ is a vector space of dimension 2 over $\mathbb{R}$, we can consider the components v_j of v as vectors in $\mathbb{R}^2$. Then Theorem 2 with $d = 2$ applies to the $v_j \in \mathbb{R}^2$, we just have to imagine that on the right hand side of (2) v_i is added with coefficient zero. So define $b_{ii} = 0$ and $b_{ij} = a_{ij}$ if $i \neq j$. Let b^* be the median of the sequence $b_{i1}, \dots, b_{in}$. Theorem 2 gives then that λ lies in the disk $D(a_{ii}, r)$ where

$$r = \sum_{j \neq i} |b_{ij} - b^*|. \tag{3}$$

The proof of Theorem 1 uses Corollary 1: λ lies in the disk $D(a_{ii}, r)$ where r is the sum of the largest $\lfloor n/2 \rfloor$ entries in the ith row of A (disregarding a_{ii}). Note that in general the estimate in (3) is gives a better bound on r than Theorem 1. □

We can also apply Corollary 2 to the components of v, considered again as vectors in $\mathbb{R}^2$. This gives that λ lies in the disk $D(a_{ii}, r)$ where r is the sum of the $k = \lceil n/2 \rceil$ largest entries in row i of A (disregarding a_{ii} again). In any special case a better estimate may come from the more general Theorem 3.

Remark 1 One could hope that an eigenvalue with (geometric) multiplicity 3 or higher should lie strictly inside the half Gershgorin disk. The simple example below shows that this is not the case.

Let A be an $n \times n$ matrix with $n = 3k$, consisting of three $k \times k$ blocks along the main diagonal, with each block being a doubly stochastic matrix. Then $\lambda = 1$ is an eigenvalue with multiplicity 3, which lies on the boundary of each half Gershgorin disk $D(a_{ii}, r_i)$. Indeed r_i is the sum of the largest $\lfloor n/2 \rfloor$ entries of the ith row (disregarding a_{ii}) which equals $1 - a_{ii}$.

This example shows, however, that λ lies in the "third Gershgorin disk". This is the disk centred at a_{ii} and of radius r which is the sum of the largest $n/3$ entries in the ith row (disregarding again a_{ii}). We return to this question at the end of the paper.

5 Examples

In what follows we show examples illustrating the limits of possible extensions of the results above. Note that one can not expect in general that a multiple eigenvalue is strictly inside the half Gershgorin disk. The simplest illustration to this is the matrix A below where 1 is an eigenvalue with (geometric) multiplicity two.

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

Next we are going to give further examples. The first two show that Theorem 1 does not extend to real matrices that have both positive and negative entries. The second is a positive semidefinite Hermitian matrix (with complex entries) where the triple eigenvalue "0" lies on the boundary of the half Gershgorin disk. Perhaps some form of Theorem 1 can be extended to such matrices.

5.1 Real Matrices with Both Positive and Negative Entries

The matrices in Theorem 1 have non-negative entries. This condition cannot be deleted as the following symmetric circulant matrix with 0, ±1 entries shows:

$$B = \begin{bmatrix} 0 & 1 & -1 & -1 & 1 \\ 1 & 0 & 1 & -1 & -1 \\ -1 & 1 & 0 & 1 & -1 \\ -1 & -1 & 1 & 0 & 1 \\ 1 & -1 & -1 & 1 & 0 \end{bmatrix}$$

Like every 5×5 symmetric circulant matrix, B has two multiple eigenvalues. They are $\sqrt{5} \approx 2.236$ and $-\sqrt{5}$ and both lie outside the half Gershgorin disk.

The following 7×7 matrix is again circulant and has 0, ±1 entries. Its multiple eigenvalue ≈ -3.494 is even further from the half Gershgorin disk which has radius 3 around the origin.

$$C = \begin{bmatrix} 0 & 1 & -1 & 1 & 1 & -1 & 1 \\ 1 & 0 & 1 & -1 & 1 & 1 & -1 \\ -1 & 1 & 0 & 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 0 & 1 & -1 & 1 \\ 1 & 1 & -1 & 1 & 0 & 1 & -1 \\ -1 & 1 & 1 & -1 & 1 & 0 & 1 \\ 1 & -1 & 1 & 1 & -1 & 1 & 0 \end{bmatrix}$$

5.2 A Positive Semidefinite Matrix

The next construction gives a 9×9 positive semidefinite Hermitian matrix H with the triple eigenvalue "0" lying on the boundary of the half Gershgorin disk. (This is very different from the example in Remark 1 where the half disk and the third disk were the same.) The other eigenvalue is 6 and it lies on the boundary of the "quarter disk". This example comes from the Hesse configuration of 9 points and 12 lines in $\mathbb{CP}^2$ [5]. The matrix H looks interesting on its own right. It shows further that strengthening Theorem 1 to more general matrices (with high multiplicity eigenvalues) might be difficult.

One possible realization of the Hesse configuration is given by the following 9 points on the complex projective plane

$$\begin{array}{lll} p_1 = (0,1,-1) & p_2 = (0,1,-\omega) & p_3 = (0,1,-\omega^2) \\ p_4 = (1,0,-1) & p_5 = (1,0,-\omega^2) & p_6 = (1,0,-\omega) \\ p_7 = (1,-1,0) & p_8 = (1,-\omega,0) & p_9 = (1,-\omega^2,0) \end{array}$$

where $\omega = \frac{-1+i\sqrt{3}}{2}$ is a third root of unity. In this arrangement each point lies on four lines and each line contains three points. Our first matrix, A, records the linear dependencies of the points. It has 9 columns, one for each point, and 12 rows, one for each line. If p_i, p_j and p_k are collinear, then there are nonzero complex multipliers α, β, γ such that $\alpha p_i + \beta p_j + \gamma p_k = 0$. For example the sixth (highlighted) row in the matrix A below represents the equation

$$-\omega^2(0,1,-1) - (0,1,-\omega) - \omega(0,1,-\omega^2) = (0,0,0).$$

Thus the matrix A encodes the linear dependencies of collinear triples in the point-line arrangement of the Hesse configuration.

$$A = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\omega^2 & -1 & -\omega \\ 0 & 0 & 0 & -\omega^2 & -\omega & -1 & 0 & 0 & 0 \\ \boldsymbol{-\omega^2} & \mathbf{-1} & \boldsymbol{-\omega} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 0 & \omega & 0 & 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & -\omega^2 & 0 & 0 & 1 & 0 & 0 & -1 \\ -\omega^2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ \omega & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 1 & 0 & -\omega & 0 & 0 & 0 & 0 & \omega \\ 0 & 0 & \omega & -1 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

The points of the Hesse configuration satisfy the homogeneous system of equations $A\mathbf{x} = \mathbf{0}$ where $x_i \in \mathbb{CP}^2$. An affine image of a solution is also a solution, implying that the rank of A is at most 6. It is easy to see that the rank is exactly 6: the rank remains the same if one multiplies a matrix with its Hermitian transpose (complex conjugate transpose). So consider the 9×9 matrix $H = \overline{A^T}A$.

$$H = \begin{bmatrix} 4 & \omega & \omega^2 & -1 & -\omega & -\omega^2 & 1 & \omega^2 & \omega \\ \omega^2 & 4 & \omega & -\omega & -\omega^2 & -1 & 1 & \omega^2 & \omega \\ \omega & \omega^2 & 4 & -\omega^2 & -1 & -\omega & 1 & \omega^2 & \omega \\ -1 & -\omega^2 & -\omega & 4 & \omega^2 & \omega & -1 & -1 & -1 \\ -\omega^2 & -\omega & -1 & \omega & 4 & \omega^2 & -1 & -1 & -1 \\ -\omega & -1 & -\omega^2 & \omega^2 & \omega & 4 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 & 4 & \omega & \omega^2 \\ \omega & \omega & \omega & -1 & -1 & -1 & \omega^2 & 4 & \omega \\ \omega^2 & \omega^2 & \omega^2 & -1 & -1 & -1 & \omega & \omega^2 & 4 \end{bmatrix}$$

Matrix H is a positive semidefinite Hermitian matrix that has two eigenvalues: 0 with multiplicity 3 (so the rank of A is indeed 6) and 6 with multiplicity 6. All non-diagonal entries have norm one and the diagonal entries are 4. Thus $\lambda = 0$ is on the boundary of the half Gershgorin disk $D(4, 4)$ and $\lambda = 6$ on the boundary of $D(4, 2)$, the “quarter disk” (Fig. 1).

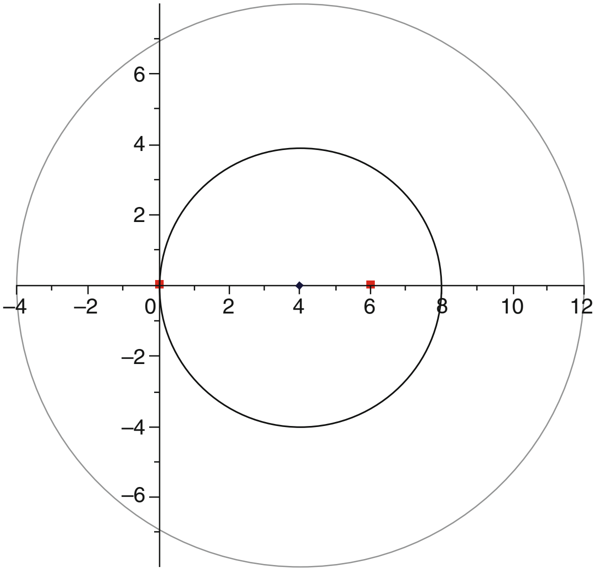


Fig. 1 The Gershgorin disk and half-disk of H

6 Remarks

There are several questions that remain open.

- What can be said about the location of an eigenvalue with larger multiplicity? Our method, using the zonotope $Z(V)$ in the proof of Theorem 3 has its limitations. Perhaps inequality (1) can be improved. For instance, for an eigenvalue with multiplicity at least k one would like to use an eigenvector $v = (v_1, \ldots, v_n)$ such that the corresponding zonotope $Z(V)$ satisfies

$$Z(V) \subset c\,(\|v_1\| + \ldots + \|v_n\|)\,B$$

  where c decreases as k grows. Unfortunately one can not expect c to go below $\frac{1}{\pi}$, (see Exercise 14.9 in [13])
- How about other matrices? What is the radius of the shrunken Gershgorin disk which contains a multiple eigenvalue of a general complex matrix? Are there better bounds for special matrices, like real or positive semidefinite Hermitian matrices?

Acknowledgements This research was supported by ERC Advanced Research Grant no 267165 (DISCONV). Imre Bárány is partially supported by Hungarian National Research Grant K 111827. József Solymosi is partially supported by Hungarian National Research Grant NK 104183 and by an NSERC Discovery Grant. We are indebted to three anonymous referees for very useful comments and information that have improved the presentation of this paper.

References

1. A. Berman, R.J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences* (SIAM, Philadelphia, 1994)
2. G.M. Del Corso, Estimating an eigenvector by the power method with a random start. SIAM. J. Matrix Anal. Appl. **18**, 913–937 (1997)
3. M. Fiedler, F.J. Hall, R. Marsli, Gershgorin discs revisited. J. Linear Algebra Appl. **438**, 598–603 (2013)
4. S. Gerschgorin, Über die Abgrenzung der Eigenwerte einer Matrix. Izv. Akad. Nauk. USSR Otd. Fiz.-Mat. Nauk **6**, 749–754 (1931)
5. O. Hesse, Über die Wendepunkte der Curven dritter Ordnung. J. Reine Angew. Math. **28**, 97–102 (1844)
6. L. Lovász, Steinitz Representations of Polyhedra and the Colin de Verdière Number. J. Combin. Theory B **82**, 223–236 (2001)
7. R. Marsli, F.J. Hall, Geometric multiplicities and Gershgorin discs. Am. Math. Mon. **120**, 452–455 (2013)
8. R. Marsli, F.J. Hall, Some refinements of Gershgorin discs. Int. J. Algebra **7**, 573–580 (2013)
9. R. Marsli, F.J. Hall, Further results on Gershgorin discs. J. Linear Algebra Appl. **439**, 189–195 (2013)
10. R. Marsli, F.J. Hall, Some new inequalities on geometric multiplicities and Gershgorin discs. Int. J. Algebra **8**, 135–147 (2014)

11. P. McMullen, Transforms, diagrams and representations, in *Contributions to Geometry. Proceedings of the Geometry-Symposium*, Siegen, 1978 (Birkhäuser, Basel/Boston, 1979), pp. 92–130
12. A. Roy, Minimal Euclidean representations of graphs. Discret. Math. **310**, 727–733 (2010)
13. J.M. Steele, *The Cauchy-Schwarz Master Class. An Introduction to the Art of Mathematical Inequalities* (Cambridge University Press, New York, 2004)
14. H. van der Holst, L. Lovász, A. Schrijver, The Colin de Verdière graph parameter, in *Graph Theory and Combinatorial Biology*. Bolyai Society Mathematical Studies, vol. 7 (János Bolyai Mathematical Society, Budapest, 1999), pp. 29–85
15. R.S. Varga, *Gershgorin and His Circles* (Springer, Berlin, 2004)

Computing the Partition Function of a Polynomial on the Boolean Cube

Alexander Barvinok

Abstract For a polynomial $f : \{-1, 1\}^n \longrightarrow \mathbb{C}$, we define the partition function as the average of $e^{\lambda f(x)}$ over all points $x \in \{-1, 1\}^n$, where $\lambda \in \mathbb{C}$ is a parameter. We present a quasi-polynomial algorithm, which, given such f, λ and $\epsilon > 0$ approximates the partition function within a relative error of ϵ in $N^{O(\ln n - \ln \epsilon)}$ time provided $|\lambda| \leq (2L\sqrt{\deg f})^{-1}$, where $L = L(f)$ is a parameter bounding the Lipschitz constant of f from above and N is the number of monomials in f. As a corollary, we obtain a quasi-polynomial algorithm, which, given such an f with coefficients ± 1 and such that every variable enters not more than 4 monomials, approximates the maximum of f on $\{-1, 1\}^n$ within a factor of $O\left(\delta^{-1}\sqrt{\deg f}\right)$, provided the maximum is $N\delta$ for some $0 < \delta \leq 1$. If every variable enters not more than k monomials for some fixed $k > 4$, we are able to establish a similar result when $\delta \geq (k-1)/k$.

1991 *Mathematics Subject Classification*. 90C09, 68C25, 68W25, 68R05.

1 Introduction and Main Results

1.1 Polynomials and Partition Functions

Let $\{-1, 1\}^n$ be the n-dimensional Boolean cube, that is, the set of all 2^n n-vectors $x = (\pm 1, \ldots, \pm 1)$ and let $f : \{-1, 1\}^n \longrightarrow \mathbb{C}$ be a polynomial with complex coefficients. We assume that f is defined as a linear combination of square-free

This research was partially supported by NSF Grant DMS 1361541.

A. Barvinok (✉)
Department of Mathematics, University of Michigan, 530 Church street, 48109-1043 Ann Arbor, MI, USA
e-mail: barvinok@umich.edu

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_7

monomials:

$$f(x) = \sum_{I \subset \{1,\dots,n\}} \alpha_I \mathbf{x}^I \quad \text{where} \quad \alpha_I \in \mathbb{C} \quad \text{for all} \quad I$$
$$\text{and} \quad \mathbf{x}^I = \prod_{i \in I} x_i \quad \text{for} \quad x = (x_1, \dots, x_n), \tag{1}$$

where we agree that $\mathbf{x}^{\emptyset} = 1$. As is known, the monomials $\mathbf{x}^I$ for $I \subset \{1, \dots, n\}$ constitute a basis of the vector space of functions $f : \{-1, 1\}^n \longrightarrow \mathbb{C}$.

We introduce two parameters measuring the complexity of the polynomial f in (1). The *degree* of f is the largest degree of a monomial $\mathbf{x}^I$ appearing in (1) with a non-zero coefficient, that is, the maximum cardinality $|I|$ such that $\alpha_I \neq 0$:

$$\deg f = \max_{I:\ \alpha_I \neq 0} |I|.$$

We also introduce a parameter which controls the Lipschitz constant of f:

$$L(f) = \max_{i=1,\dots,n} \sum_{\substack{I \subset \{1,\dots,n\} \\ i \in I}} |\alpha_I|.$$

Indeed, if dist is the metric on the cube,

$$\text{dist}(x, y) = \sum_{i=1}^{n} |x_i - y_i| \quad \text{where} \quad x = (x_1, \dots, x_n) \quad \text{and} \quad y = (y_1, \dots, y_n)$$

then

$$|f(x) - f(y)| \leq L(f)\, \text{dist}(x, y).$$

We consider $\{-1, 1\}^n$ as a finite probability space endowed with the uniform measure.

For $\lambda \in \mathbb{C}$ and a polynomial $f : \{-1, 1\}^n \longrightarrow \mathbb{C}$, we introduce the *partition function*

$$\frac{1}{2^n} \sum_{x \in \{-1,1\}^n} e^{\lambda f(x)} = \mathbf{E} e^{\lambda f}.$$

Our first main result bounds from below the distance from the zeros of the partition function to the origin.

Theorem 1.1 *Let $f : \{-1, 1\}^n \longrightarrow \mathbb{C}$ be a polynomial and let $\lambda \in \mathbb{C}$ be such that*

$$|\lambda| \leq \frac{0.55}{L(f)\sqrt{\deg f}}.$$

Then

$$\mathbf{E}\, e^{\lambda f} \neq 0.$$

If, additionally, the constant term of f is 0 *then*

$$\left|\mathbf{E}\, e^{\lambda f}\right| \geq (0.41)^n.$$

We prove Theorem 1.1 in Sect. 4. As a simple example, let $f(x_1, \dots, x_n) = x_1 + \dots + x_n$. Then

$$\mathbf{E}\, e^{\lambda f} = \left(\mathbf{E}\, e^{\lambda x_1}\right) \cdots \left(\mathbf{E}\, e^{\lambda x_n}\right) = \left(\frac{e^{\lambda} + e^{-\lambda}}{2}\right)^n.$$

We have $L(f) = \deg f = 1$ and Theorem 1.1 predicts that $\mathbf{E}\, e^{\lambda f} \neq 0$ provided $|\lambda| \leq 0.55$. Indeed, the smallest in the absolute value root of $\mathbf{E}\, e^{\lambda f}$ is $\lambda = \pi i/2$ with $|\lambda| = \pi/2 \approx 1.57$. If we pick $f(x_1, \dots, x_n) = ax_1 + \dots + ax_n$ for some real constant $a > 0$ then the smallest in the absolute value root of $\mathbf{E}\, e^{\lambda f}$ is $\pi i/2a$ with $|\lambda|$ inversely proportional to $L(f)$, just as Theorem 1.1 predicts. It is not clear at the moment whether the dependence of the bound in Theorem 1.1 on $\deg f$ is optimal.

As we will see shortly, Theorem 1.1 implies that $\mathbf{E}\, e^{\lambda f}$ can be efficiently computed if $|\lambda|$ is strictly smaller than the bound in Theorem 1.1. When computing $\mathbf{E}\, e^{\lambda f}$, we may assume that the constant term of f is 0, since

$$\mathbf{E}\, e^{\lambda(f+\alpha)} = e^{\lambda\alpha}\mathbf{E}\, e^{\lambda f}$$

and hence adding a constant to f results in multiplying the partition function by a constant.

For a given f, we consider a univariate function

$$\lambda \longmapsto \mathbf{E}\, e^{\lambda f}.$$

As follows from Theorem 1.1, we can choose a branch of

$$g(\lambda) = \ln\left(\mathbf{E}\, e^{\lambda f}\right) \quad \text{for} \quad |\lambda| \leq \frac{0.55}{L(f)\sqrt{\deg f}}$$

such that $g(0) = 0$. It follows that $g(\lambda)$ is well-approximated by a low degree Taylor polynomial at 0.

Theorem 1.2 *Let $f : \{-1,1\}^n \longrightarrow \mathbb{C}$ be a polynomial with zero constant term and let*

$$g(\lambda) = \ln\left(\mathbf{E}\, e^{\lambda f}\right) \quad \textit{for} \quad |\lambda| \leq \frac{0.55}{L(f)\sqrt{\deg f}}.$$

For a positive integer $m \leq 5n$, let

$$T_m(f;\lambda) = \sum_{k=1}^{m} \frac{\lambda^k}{k!} \frac{d^k}{d\lambda^k} g(\lambda)\Big|_{\lambda=0}$$

be the degree m Taylor polynomial of $g(\lambda)$ computed at $\lambda = 0$. Then for $n \geq 2$

$$|g(\lambda) - T_m(f;\lambda)| \leq \frac{50n}{(m+1)(1.1)^m} + e^{-n}$$

provided

$$|\lambda| \leq \frac{1}{2L(f)\sqrt{\deg f}}. \tag{2}$$

In Sect. 3, we deduce Theorem 1.2 from Theorem 1.1.

As we discuss in Sect. 3.1, for a polynomial f given by (1), the value of $T_m(f;\lambda)$ can be computed in $nN^{O(m)}$ time, where N is the number of monomials in the representation (1). Theorem 1.2 implies that as long as $\epsilon \gg e^{-n}$, by choosing $m = O\big(\ln n - \ln \epsilon\big)$, we can compute the value of $\mathbf{E}\, e^{\lambda f}$ within relative error ϵ in $N^{O(\ln n - \ln \epsilon)}$ time provided λ satisfies the inequality (2). For ϵ exponentially small in n, it is more efficient to evaluate $\mathbf{E}\, e^{\lambda f}$ directly from the definition.

1.2 Relation to Prior Work

This paper is a continuation of a series of papers by the author [3, 4] and by the author and P. Soberón [5, 6] on algorithms to compute partition functions in combinatorics, see also [16]. The main idea of the method is that the logarithm of the partition function is well-approximated by a low-degree Taylor polynomial at the temperatures above the phase transition (the role of the temperature is played by $1/\lambda$), while the phase transition is governed by the complex zeros of the partition function, cf. [15, 18].

The main work of the method consists of bounding the complex roots of the partition function, as in Theorem 1.1. While the general approach of this paper looks similar to the approach of [3–5] and [6] (a martingale type and a fixed point type arguments), in each case bounding complex roots requires some effort and new ideas. Once the roots are bounded, it is relatively straightforward to approximate the partition function as in Theorem 1.2.

Another approach to computing partition functions, also rooted in statistical physics, is the correlation decay approach, see [17] and [1]. While we did not pursue that approach, in our situation it could conceivably work as follows: given a polynomial $f : \{-1, 1\}^n \longrightarrow \mathbb{R}$ and a real $\lambda > 0$, we consider the Boolean cube as a finite probability space, where the probability of a point $x \in \{-1, 1\}^n$ is $e^{\lambda f(x)}/\mathbf{E}\, e^{\lambda f}$. This makes the coordinates $x_1, \ldots, x_n$ random variables. We consider a graph with vertices $x_1, \ldots, x_n$ and edges connecting two vertices x_i and x_j if there is a monomial of f containing both x_i and x_j. This introduces a graph metric on the variables $x_1, \ldots, x_n$ and one could hope that if λ is sufficiently small, we have correlation decay: the random variable x_i is almost independent on the random variables sufficiently distant from x_i in the graph metric. This would allow us to efficiently approximate the probabilities $\mathbf{P}\,(x_i = 1)$ and $\mathbf{P}\,(x_i = -1)$ and then recursively estimate $\mathbf{E}\, e^{\lambda f}$.

While both approaches treat the phase transition as a natural threshold for computability, the concepts of phase transition in our method (complex zeros of the partition function) and in the correlation decay approach (non-uniqueness of Gibbs measures) though definitely related and even equivalent for some spin systems [8], in general are different.

Theorem 1.2 together with the algorithm of Sect. 3.1 below implies that to approximate $\mathbf{E}\, e^{\lambda f}$ within a relative error of $\epsilon > 0$, it suffices to compute moments $\mathbf{E} f^k$ for $k = O\left(\ln \epsilon^{-1}\right)$. This suggests some similarity with one of the results of [13], where (among other results) it is shown that the number of satisfying assignments of a DNF on n Boolean variables is uniquely determined by the numbers of satisfying assignments for all possible conjunctions of $k \leq 1 + \log_2 n$ clauses of the DNF (though this is a purely existential result with no algorithm attached). Each conjunction of the DNF can be represented as a polynomial

$$\phi_j(x) = \frac{1}{2^{|S_j|}} \prod_{i \in S_j} (1 + \epsilon_i x_i) \quad \text{where}$$

$$S_j \subset \{1, \ldots, n\} \quad \text{and} \quad \epsilon_i \in \{-1, 1\},$$

and we let

$$f(x) = \sum_{j=1}^{m} \phi_j(x).$$

Then the number of points $x \in \{-1, 1\}^n$ such that $f(x) > 0$ is uniquely determined by various expectations $\mathbf{E}\, \phi_{j_1} \cdots \phi_{j_k}$ for $k \leq 1 + \log_2 n$. The probability that $f(x) = 0$ for a random point $x \in \{-1, 1\}^n$ sampled from the uniform distribution, can be approximated by $\mathbf{E}\, e^{-\lambda f}$ for a sufficiently large $\lambda > 0$. The expectations are precisely those that arise when we compute the moments $\mathbf{E} f^k$. It is not clear at the moment whether the results of this paper can produce an efficient way to compute the number of satisfying assignments.

2 Applications to Optimization

2.1 Maximizing a Polynomial on the Boolean Cube

Let $f : \{-1, 1\}^n \longrightarrow \mathbb{R}$ be a polynomial with real coefficients defined by its monomial expansion (1). As is known, various computationally hard problems of discrete optimization, such as finding the maximum cardinality of an independent set in a graph, finding the minimum cardinality of a vertex cover in a hypergraph and the maximum constraint satisfaction problem can be reduced to finding the maximum of f on the Boolean cube $\{-1, 1\}^n$, see, for example, [7].

The problem is straightforward if $\deg f \leq 1$. If $\deg f = 2$, it may already be quite hard even to solve approximately: Given an undirected simple graph $G = (V, E)$ with set $V = \{1, \ldots, n\}$ of vertices and set $E \subset \binom{V}{2}$ of edges, one can express the largest cardinality of an *independent set* (a set vertices no two of which are connected by an edge of the graph), as the maximum of

$$f(x) = \frac{1}{2} \sum_{i=1}^{n} (x_i + 1) - \frac{1}{4} \sum_{\{i,j\} \in E} (1 + x_i) \left(1 + x_j\right)$$

on the cube $\{-1, 1\}^n$. It is an NP-hard problem to approximate the size of the largest independent set in a given graph on n vertices within a factor of $n^{1-\epsilon}$ for any $0 < \epsilon \leq 1$, fixed in advance [10, 19]. If $\deg f = 2$ and f does not contain linear or constant terms, the problem reduces to the max cut problem in a weighted graph (with both positive and negative weights allowed on the edges), where there exists a polynomial time algorithm achieving an $O(\ln n)$ approximation factor, see [14] for a survey.

If $\deg f \geq 3$, no efficient algorithm appears to be known that would outperform choosing a random point $x \in \{-1, 1\}^n$. The maximum of a polynomial f with $\deg f = 3$ and no constant, linear or quadratic terms can be approximated within an $O\left(\sqrt{n/\ln n}\right)$ factor in polynomial time, see [14]. Finding the maximum of a general real polynomial (1) on the Boolean cube $\{-1, 1\}^n$ is equivalent to the problem of finding the maximum weight of a subset of a system of weighted linear equations over $\mathbb{Z}_2$ that can be simultaneously satisfied [12]. Assuming that $\deg f$ is fixed in advance, f contains N monomials and the constant term of f is 0, a polynomial time algorithm approximating the maximum of f within a factor of $O(\sqrt{N})$ is constructed in [12]. More precisely, the algorithm from [12] constructs a point x such that $f(x)$ is within a factor of $O(\sqrt{N})$ from $\sum_I |\alpha_I|$ for f defined by (1). If $\deg f \geq 3$, it is unlikely that a polynomial time algorithm exists approximating the maximum of f within a factor of $2^{(\ln N)^{1-\epsilon}}$ for any fixed $0 < \epsilon \leq 1$ [12], see also [10].

Let us choose

$$\lambda = \frac{1}{2L(f)\sqrt{\deg f}}$$

as in Theorem 1.2. As is discussed in Sect. 3.2, by successive conditioning, we can compute in $N^{O(\ln n - \ln \epsilon)}$ time a point $y \in \{-1, 1\}^n$ which satisfies

$$e^{\lambda f(y)} \ \geq \ (1-\epsilon)\mathbf{E}\, e^{\lambda f} \tag{3}$$

for any given $0 < \epsilon \leq 1$.

How well a point y satisfying (3) approximates the maximum value of f on the Boolean cube $\{-1, 1\}^n$? We consider polynomials with coefficients -1, 0 and 1, where the problem of finding an $x \in \{-1, 1\}^n$ maximizing $f(x)$ is equivalent to finding a vector in $\mathbb{Z}_2^n$ satisfying the largest number of linear equations from a given list of linear equations over $\mathbb{Z}_2$.

Theorem 2.1 *Let*

$$f(x) = \sum_{I \in \mathcal{F}} \alpha_I \mathbf{x}^I$$

be a polynomial with zero constant term, where $\mathcal{F}$ is a family of non-empty subsets of the set $\{1, \dots, n\}$ and $\alpha_I = \pm 1$ for all $I \in \mathcal{F}$. Let

$$\max_{x \in \{-1,1\}^n} f(x) = \delta|\mathcal{F}| \quad \textit{for some} \quad 0 \leq \delta \leq 1.$$

Suppose further that every variable x_i enters at most four monomials $\mathbf{x}^I$ for $I \in \mathcal{F}$. Then

$$\mathbf{E}\, e^{\lambda f} \ \geq \ \exp\left\{\frac{3\lambda^2\delta^2}{16}|\mathcal{F}|\right\} \quad \textit{for} \quad 0 \leq \lambda \leq 1.$$

Since $\mathbf{E} f = 0$, the maximum of f is positive unless $\mathcal{F} = \emptyset$ and $f \equiv 0$. It is not clear whether the restriction on the number of occurrences of variables in Theorem 2.1 is essential or an artifact of the proof. We can get a similar estimate for any number occurrences provided the maximum of f is sufficiently close to $|\mathcal{F}|$.

Theorem 2.2 *Let*

$$f(x) = \sum_{I \in \mathcal{F}} \alpha_I \mathbf{x}^I$$

be a polynomial with zero constant term, where $\mathcal{F}$ is a family of non-empty subsets of the set $\{1, \dots, n\}$ and $\alpha_I = \pm 1$ for all $I \in \mathcal{F}$. Let $k > 2$ be an integer and suppose *that every variable x_i enters at most k monomials $\mathbf{x}^I$ for $I \in \mathcal{F}$. If*

$$\max_{x \in \{-1,1\}^n} f(x) \ \geq \ \frac{k-1}{k}|\mathcal{F}|$$

then

$$\mathbf{E}\, e^{\lambda f} \ \geq \ \exp\left\{\frac{3\lambda^2}{16}|\mathcal{F}|\right\} \quad \textit{for all} \quad 0 \leq \lambda \leq 1.$$

We prove Theorems 2.1 and 2.2 in Sect. 5.

Let f be a polynomial of Theorem 2.1 and suppose that, additionally, $|I| \leq d$ for all $I \in \mathcal{F}$, so that $\deg f \leq d$. We have $L(f) \leq 4$ and we choose

$$\lambda = \frac{1}{8\sqrt{d}}.$$

Let $y \in \{-1, 1\}^n$ be a point satisfying (3). Then

$$f(y) \ \geq \ \frac{1}{\lambda}\ln \mathbf{E}\, e^{\lambda f} + \frac{\ln(1-\epsilon)}{\lambda} \ \geq \ \frac{3\lambda\delta^2}{16}|\mathcal{F}| + \frac{\ln(1-\epsilon)}{\lambda}.$$

That is, if the maximum of f is at least $\delta|\mathcal{F}|$ for some $0 < \delta \leq 1$, we can approximate the maximum in quasi-polynomial time within a factor of $O\left(\delta^{-1}\sqrt{d}\right)$. Equivalently, if for some $0 < \delta \leq 0.5$ there is a vector in $\mathbb{Z}_2^n$ satisfying at least $(0.5 + \delta)|\mathcal{F}|$ equations of a set $\mathcal{F}$ of linear equations over $\mathbb{Z}_2$, where each variable enters at most 4 equations, in quasi-polynomial time we can compute a vector $v \in \mathbb{Z}_2^n$ satisfying at least $(0.5 + \delta_1)|\mathcal{F}|$ linear equations from the system, where $\delta_1 = \Omega(\delta^2/\sqrt{d})$ and d is the largest number of variables per equation.

Similarly, we can approximate in quasi-polynomial time the maximum of f in Theorem 2.2 within a factor of $O(k\sqrt{d})$ provided the maximum is sufficiently close to $|\mathcal{F}|$, that is, is at least $\frac{k-1}{k}|\mathcal{F}|$.

In Theorems 2.1 and 2.2, one can check in polynomial time whether the maximum of f is equal to $|\mathcal{F}|$, as this reduces to testing the feasibility of a system of linear equations over $\mathbb{Z}_2$. However, for any fixed $0 < \delta < 1$, testing whether the maximum is at least $\delta|\mathcal{F}|$ is computationally hard, cf. [10].

Håstad [9] constructed a polynomial time algorithm that approximates the maximum of f within a factor of $O(kd)$. In [2], see also [11], a polynomial algorithm is constructed that finds the maximum of f within a factor of $e^{O(d)}\sqrt{k}$, provided f is an odd function. More precisely, the algorithm finds a point x such that $f(x)$ is within a factor of $e^{O(d)}\sqrt{k}$ from $|\mathcal{F}|$.

3 Computing the Partition Function

3.1 Computing the Taylor Polynomial of $g(\lambda) = \ln\left(\mathbf{E}\, e^{\lambda f}\right)$

First, we discuss how to compute the degree m Taylor polynomial $T_m(f;\lambda)$ at $\lambda = 0$ of the function

$$g(\lambda) = \ln\left(\mathbf{E}\, e^{\lambda f}\right),$$

see Theorem 1.2. Let us denote

$$h(\lambda) = \mathbf{E}\, e^{\lambda f} \quad \text{and} \quad g(\lambda) = \ln h(\lambda).$$

Then

$$g' = \frac{h'}{h} \quad \text{and hence} \quad h' = g'h.$$

Therefore,

$$h^{(k)}(0) = \sum_{j=1}^{k} \binom{k-1}{j-1} g^{(j)}(0) h^{(k-j)}(0) \quad \text{for} \quad k = 1, \ldots, m. \tag{4}$$

If we calculate the derivatives

$$h(0),\ h^{(1)}(0), \ldots, h^{(m)}(0), \tag{5}$$

then we can compute

$$g(0),\ g^{(1)}(0), \ldots, g^{(m)}(0)$$

by solving a non-singular triangular system of linear equations (4) which has $h(0) = 1$ on the diagonal. Hence our goal is to calculate the derivatives (5).

We observe that

$$h^{(k)}(0) = \frac{1}{2^n} \sum_{x \in \{-1,1\}^n} f^k(x) = \mathbf{E} f^k.$$

For a polynomial f defined by its monomial expansion (1) we have

$$\mathbf{E} f = \alpha_{\emptyset}.$$

We can consecutively compute the monomial expansion of $f, f^2, \ldots, f^m$ by using the following multiplication rule for monomials on the Boolean cube $\{-1, 1\}^n$:

$$\mathbf{x}^I \mathbf{x}^J = \mathbf{x}^{I \Delta J},$$

where $I \Delta J$ is the symmetric difference of subsets $I, J \subset \{1, \ldots, n\}$. It follows then that for a polynomial $f : \{-1, 1\}^n \longrightarrow \mathbb{C}$ given by its monomial expansion (1) and a positive integer m, the Taylor polynomial

$$T_m(f; \lambda) = \sum_{k=1}^{m} \frac{\lambda^k}{k!} \frac{d^k}{d\lambda^k} g(\lambda) \Big|_{\lambda=0}$$

can be computed in $nN^{O(m)}$ time, where N is the number of monomials in f.

Our next goal is deduce Theorem 1.2 from Theorem 1.1. The proof is based on the following lemma.

Lemma 3.1 *Let $p : \mathbb{C} \longrightarrow \mathbb{C}$ be a univariate polynomial and suppose that for some $\beta > 0$ we have*

$$p(z) \neq 0 \quad \textit{provided} \quad |z| \leq \beta.$$

Let $0 < \gamma < \beta$ and for $|z| \leq \gamma$, let us choose a continuous branch of

$$g(z) = \ln p(z).$$

Let

$$T_m(z) = g(0) + \sum_{k=1}^{m} \frac{z^k}{k!} \frac{d^k}{dz^k} g(z) \Big|_{z=0}$$

be the degree m Taylor polynomial of $g(z)$ computed at $z = 0$. Then for

$$\tau = \frac{\beta}{\gamma} > 1$$

we have

$$|g(z) - T_m(z)| \leq \frac{\deg p}{(m+1)\tau^m(\tau - 1)} \quad \textit{for all} \quad |z| \leq \gamma.$$

Proof Let $n = \deg p$ and let $\alpha_1, \ldots, \alpha_n$ be the roots of p, so we may write

$$p(z) = p(0) \prod_{i=1}^{n} \left(1 - \frac{z}{\alpha_i}\right) \quad \text{where} \quad |\alpha_i| \geq \beta \quad \text{for} \quad i = 1, \ldots, n.$$

Then

$$g(z) = g(0) + \sum_{i=1}^{n} \ln\left(1 - \frac{z}{\alpha_i}\right),$$

where we choose the branch of the logarithm which is 0 when $z = 0$. Using the Taylor series expansion of the logarithm, we obtain

$$\ln\left(1 - \frac{z}{\alpha_i}\right) = -\sum_{k=1}^{m} \frac{z^k}{k\alpha_i^k} + \zeta_m \quad \text{provided} \quad |z| \leq \gamma,$$

where

$$|\zeta_m| = \left| - \sum_{k=m+1}^{+\infty} \frac{z^k}{k\alpha_i^k} \right| \leq \sum_{k=m+1}^{+\infty} \frac{\gamma^k}{k\beta^k} \leq \frac{1}{(m+1)\tau^m(\tau-1)}.$$

Therefore,

$$g(z) = g(0) - \sum_{i=1}^{n} \sum_{k=1}^{m} \frac{z^k}{k\alpha_i^k} + \eta_m \quad \text{for} \quad |z| \leq \gamma,$$

where

$$|\eta_m| \leq \frac{n}{(m+1)\tau^m(\tau-1)}.$$

It remains to notice that

$$T_m(z) = g(0) - \sum_{i=1}^{n} \sum_{k=1}^{m} \frac{z^k}{k\alpha_i^k}.$$

□

Next, we need a technical bound on the approximation of e^z by its Taylor polynomial.

Lemma 3.2 *Let $\rho > 0$ be a real number and let $m \geq 5\rho$ be an integer. Then*

$$\left| e^z - \sum_{k=0}^{m} \frac{z^k}{k!} \right| \leq e^{-2\rho} \quad \textit{for all} \quad z \in \mathbb{C} \quad \textit{such that} \quad |z| \leq \rho.$$

Proof For all $z \in \mathbb{C}$ such that $|z| \leq \rho$, we have

$$\left| e^z - \sum_{k=0}^{m} \frac{z^k}{k!} \right| = \left| \sum_{k=m+1}^{+\infty} \frac{z^k}{k!} \right| \leq \sum_{k=m+1}^{+\infty} \frac{\rho^k}{k!} = \frac{\rho^{m+1}}{(m+1)!} \sum_{k=0}^{+\infty} \frac{\rho^k (m+1)!}{(k+m+1)!}$$
$$\leq \frac{\rho^{m+1}}{(m+1)!} \sum_{k=0}^{+\infty} \frac{\rho^k}{k!} = \frac{\rho^{m+1} e^{\rho}}{(m+1)!} \leq \frac{\rho^{m+1} e^{\rho+m+1}}{(m+1)^{m+1}}.$$

Since $m \geq 5\rho$, we obtain

$$\left| e^z - \sum_{k=0}^{+\infty} \frac{z^k}{k!} \right| \leq \frac{\rho^{m+1} e^{\rho+m+1}}{5^{m+1} \rho^{m+1}} = \frac{e^{\rho}}{(5/e)^{m+1}} \leq \frac{e^{\rho}}{(5/e)^{5\rho}} \leq e^{-2\rho}.$$

and the proof follows. □

Proof of Theorem 1.2 Without loss of generality, we assume that $L(f) = 1$. Since the constant term of f is 0, for any $x \in \{-1, 1\}^n$, we have

$$|f(x)| \leq \sum_{i=1}^{n} \sum_{I:\, i \in I} |\alpha_I| \leq n.$$

Applying Lemma 3.2, we conclude that

$$\left| e^{\lambda f(x)} - \sum_{k=0}^{5n} \frac{\left(\lambda f(x)\right)^k}{k!} \right| \leq e^{-2n} \quad \text{for all} \quad x \in \{-1, 1\}^n \tag{6}$$

provided $|\lambda| \leq 1$. Let

$$p(\lambda) = 1 + \sum_{k=1}^{5n} \frac{\lambda^k}{k!} \frac{d^k}{d\lambda^k} \left(\mathbf{E}\, e^{\lambda f} \right) \Big|_{\lambda=0}$$

be the degree $5n$ Taylor polynomial of the function $\lambda \longmapsto \mathbf{E}\, e^{\lambda f}$ at $\lambda = 0$. From (6) it follows that

$$\left| \mathbf{E}\, e^{\lambda f} - p(\lambda) \right| \leq e^{-2n} \quad \text{provided} \quad |\lambda| \leq 1.$$

From Theorem 1.1, we conclude that

$$p(\lambda) \neq 0 \quad \text{for all} \quad \lambda \in \mathbb{C} \quad \text{such that} \quad |\lambda| \leq \frac{0.55}{\sqrt{\deg f}}$$

and, moreover,

$$\left|\ln p(\lambda) - \ln\left(\mathbf{E}\, e^{\lambda f}\right)\right| \leq e^{-n} \quad \text{provided} \quad |\lambda| \leq \frac{0.55}{\sqrt{\deg f}} \quad \text{and} \quad n \geq 2. \tag{7}$$

Applying Lemma 3.1 with

$$\beta = \frac{0.55}{\sqrt{\deg f}}, \quad \gamma = \frac{0.5}{\sqrt{\deg f}} \quad \text{and} \quad \tau = \frac{\beta}{\gamma} = 1.1,$$

we conclude that for the Taylor polynomial of $\ln p(\lambda)$ at $\lambda = 0$,

$$T_m(\lambda) = \ln p(0) + \sum_{k=1}^{m} \frac{\lambda^k}{k!} \frac{d^k}{d\lambda^k} \ln p(\lambda)\Big|_{\lambda=0}$$

we have

$$|T_m(\lambda) - \ln p(\lambda)| \leq \frac{50n}{(m+1)(1.1)^m} \quad \text{provided} \quad |\lambda| \leq \frac{1}{2\sqrt{\deg f}}. \tag{8}$$

It remains to notice that the Taylor polynomials of degree $m \leq 5n$ of the functions

$$\lambda \longmapsto \ln\left(\mathbf{E}\, e^{\lambda f}\right) \quad \text{and} \quad \lambda \longmapsto \ln p(\lambda)$$

at $\lambda = 0$ coincide, since both are determined by the first m derivatives of respectively $\mathbf{E}\, e^{\lambda f}$ and $p(\lambda)$ at $\lambda = 0$, cf. Sect. 3.1, and those derivatives coincide. The proof now follows by (7) and (8). □

3.2 Computing a Point y in the Cube with a Large Value of $f(y)$

We discuss how to compute a point $y \in \{-1, 1\}^n$ satisfying (3). We do it by successive conditioning and determine one coordinate of $y = (y_1, \ldots, y_n)$ at a time. Let F^+ and F^- be the facets of the cube $\{-1, 1\}^n$ defined by the equations $x_n = 1$ and $x_n = -1$ respectively for $x = (x_1, \ldots, x_n)$, $x \in \{-1, 1\}^n$. Then F^+ and F^- can be identified with the $(n-1)$-dimensional cube $\{-1, 1\}^{n-1}$ and we have

$$\mathbf{E}\, e^{\lambda f} = \frac{1}{2}\mathbf{E}\left(e^{\lambda f}|F^+\right) + \frac{1}{2}\mathbf{E}\left(e^{\lambda f}|F^-\right).$$

Moreover, for the restrictions f^+ and f^- of f onto F^+ and F^- respectively, considered as polynomials on $\{-1, 1\}^{n-1}$, we have

$$\deg f^+, \deg f^- \leq \deg f \quad \text{and} \quad L(f^+), L(f^-) \leq L(f).$$

Using the algorithm of Sect. 3.1 and Theorem 1.2, we compute $\mathbf{E}\left(e^{\lambda f}|F^+\right)$ and $\mathbf{E}\left(e^{\lambda f}|F^-\right)$ within a relative error $\epsilon/2n$, choose the facet with the larger computed value, let $y_n = 1$ if the value of $\mathbf{E}\left(e^{\lambda f}|F^+\right)$ appears to be larger and let $y_n = -1$ if the value of $\mathbf{E}\left(e^{\lambda f}|F^-\right)$ appears to be larger and proceed further by conditioning on the value of y_{n-1}. For polynomials with N monomials, the complexity of the algorithm is $N^{O(\ln n)}$.

4 Proof of Theorem 1.1

To prove Theorem 1.1, we consider restrictions of the partition function onto faces of the cube.

4.1 Faces

A *face* $F \subset \{-1, 1\}^n$ consists of the points x where some of the coordinates of x are fixed at 1, some are fixed at -1 and others are allowed to vary (a face is always non-empty). With a face F, we associate three subsets $I_+(F), I_-(F), I(F) \subset \{1, \ldots, n\}$ as follows:

$$\begin{aligned} I_+(F) =& \big\{i : \ x_i = 1 \quad \text{for all} \quad x \in F, \ x = (x_1, \ldots, x_n)\big\}, \\ I_-(F) =& \big\{i : \ x_i = -1 \quad \text{for all} \quad x \in F, \ x = (x_1, \ldots, x_n)\big\} \quad \text{and} \\ I(F) =& \{1, \ldots, n\} \setminus (I_+(F) \cup I_-(F)) \,. \end{aligned}$$

Consequently,

$$\begin{aligned} F = \Big\{ (x_1, \ldots, x_n) \quad \text{where} \quad & x_i = 1 \quad \text{for} \quad i \in I_+(F) \quad \text{and} \\ & x_i = -1 \quad \text{for} \quad i \in I_-(F) \Big\}. \end{aligned}$$

In particular, if $I_+(F) = I_-(F) = \emptyset$ and hence $I(F) = \{1, \ldots, n\}$, we have $F = \{-1, 1\}^n$. We call the number

$$\dim F = |I(F)|$$

the *dimension* of F.

For a subset $J \in \{1, \ldots, n\}$, we denote by $\{-1, 1\}^J$ the set of all points

$$x = \big(x_j : j \in J\big) \quad \text{where} \quad x_j = \pm 1.$$

Let $F \subset \{-1, 1\}^n$ be a face. For a subset $J \subset I(F)$ and a point $\epsilon \in \{-1, 1\}^J$, $\epsilon = \big(\epsilon_j : j \in J\big)$, we define

$$F^\epsilon = \big\{x \in F,\ x = (x_1, \ldots, x_n) : x_j = \epsilon_j \quad \text{for} \quad j \in J\big\}.$$

In words: F^ϵ is obtained from F by fixing the coordinates from some set $J \subset I(F)$ of free coordinates to 1 or to -1. Hence F^ϵ is also a face of $\{-1, 1\}^n$ and we think of $F^\epsilon \subset F$ as a face of F. We can represent F as a disjoint union

$$F = \bigcup_{\epsilon \in \{-1,1\}^J} F^\epsilon \quad \text{for any} \quad J \subset I(F). \tag{9}$$

4.2 The Space of Polynomials

Let us fix a positive integer d. We identify the set of all polynomials f as in (1) such that $\deg f \leq d$ and the constant term of f is 0 with $\mathbb{C}^N$, where

$$N = N(n, d) = \sum_{k=1}^{d} \binom{n}{k}.$$

For $\delta > 0$, we consider a closed convex set $\mathcal{U}(\delta) \subset \mathbb{C}^N$ consisting of the polynomials $f : \{-1, 1\}^n \longrightarrow \mathbb{C}$ such that $\deg f \leq d$ and $L(f) \leq \delta$. In other words, $\mathcal{U}(\delta)$ consists of the polynomials

$$f(x) = \sum_{\substack{I \subset \{1,\ldots,n\} \\ 1 \leq |I| \leq d}} \alpha_I \mathbf{x}^I \quad \text{where} \quad \sum_{I:\ i \in I} |\alpha_I| \ \leq \ \delta \quad \text{for} \quad i = 1, \ldots, n.$$

4.3 Restriction of the Partition Function onto a Face

Let $f : \{-1, 1\}^n \longrightarrow \mathbb{C}$ be a polynomial and let $F \subset \{-1, 1\}^n$ be a face. We define

$$\mathbf{E}\ \big(e^f | F\big) = \frac{1}{2^{\dim F}} \sum_{x \in F} e^{f(x)}.$$

We suppose that f is defined by its monomial expansion as in (1) and consider $\mathbf{E}\left(e^f|F\right)$ as a function of the coefficients α_I. Using (9) we deduce

$$\begin{aligned}
\frac{\partial}{\partial \alpha_J}\mathbf{E}\left(e^f|F\right) &= \frac{1}{2^{\dim F}}\sum_{x\in F}\mathbf{x}^J e^{f(x)} \\
&= \frac{(-1)^{|I_-(F)\cap J|}}{2^{|I(F)|}} \\
&\times \sum_{\substack{\epsilon\in\{-1,1\}^{I(F)\cap J} \\ \epsilon=(\epsilon_j:\ j\in I(F)\cap J)}}\left(\prod_{j\in I(F)\cap J}\epsilon_j\right)\sum_{x\in F^\epsilon} e^{f(x)} \\
&= \frac{(-1)^{|I_-(F)\cap J|}}{2^{|I(F)\cap J|}} \\
&\times \sum_{\substack{\epsilon\in\{-1,1\}^{I(F)\cap J} \\ \epsilon=(\epsilon_j:\ j\in I(F)\cap J)}}\left(\prod_{j\in I(F)\cap J}\epsilon_j\right)\mathbf{E}\left(e^f|F^\epsilon\right).
\end{aligned} \tag{10}$$

In what follows, we identify complex numbers with vectors in $\mathbb{R}^2 = \mathbb{C}$ and measure angles between non-zero complex numbers.

Lemma 4.1 *Let $0 < \tau \leq 1$ and $\delta > 0$ be real numbers and let $F \subset \{-1,1\}^n$ be a face. Suppose that for every $f \in \mathcal{U}(\delta)$ we have $\mathbf{E}\left(e^f|F\right) \neq 0$ and, moreover, for any $K \subset I(F)$ we have*

$$\left|\mathbf{E}\left(e^f|F\right)\right| \geq \left(\frac{\tau}{2}\right)^{|K|}\sum_{\epsilon\in\{-1,1\}^K}\left|\mathbf{E}\left(e^f, F^\epsilon\right)\right|.$$

Given $f \in \mathcal{U}(\delta)$ and a subset $J \subset \{1,\ldots,n\}$ such that $|J| \leq d$, let $\widehat{f} \in \mathcal{U}(\delta)$ be the polynomial obtained from f by changing the coefficient α_J of the monomial $\mathbf{x}^J$ in f to $-\alpha_J$ and leaving all other coefficients intact. Then the angle between the two non-zero complex numbers $\mathbf{E}\left(e^f|F\right)$ and $\mathbf{E}\left(e^{\widehat{f}}|F\right)$ does not exceed

$$\frac{2|\alpha_J|}{\tau^d}.$$

Proof Without loss of generality, we assume that $\alpha_J \neq 0$.

We note that for any $f \in \mathcal{U}(\delta)$, we have $\widehat{f} \in \mathcal{U}(\delta)$. Since $\mathbf{E}\left(e^f|F\right) \neq 0$ for all $f \in \mathcal{U}(\delta)$, we may consider a branch of $\ln\mathbf{E}\left(e^f|F\right)$ for $f \in \mathcal{U}(\delta)$.

Let us fix coefficients α_I for $I \neq J$ in

$$f(x) = \sum_{\substack{I \subset \{1,\dots,n\} \\ 1 \leq |I| \leq d}} \alpha_I \mathbf{x}^I \tag{11}$$

and define a univariate function

$$g(\alpha) = \ln \mathbf{E}\,\left(e^f | F\right) \quad \text{where} \quad |\alpha| \leq |\alpha_J|$$

obtained by replacing α_J with α in (11).

We obtain

$$g'(\alpha) = \frac{\partial}{\partial \alpha_J} \ln \mathbf{E}\,\left(e^f | F\right) = \left(\frac{\partial}{\partial \alpha_J} \mathbf{E}\,\left(e^f | F\right)\right) \Big/ \mathbf{E}\,\left(e^f | F\right). \tag{12}$$

Let

$$k = |I(F) \cap J| \;\leq\; |J| \;\leq\; d.$$

Using (10) we conclude that

$$\left|\frac{\partial}{\partial \alpha_J} \mathbf{E}\,\left(e^f | F\right)\right| \;\leq\; \frac{1}{2^k} \sum_{\epsilon \in \{-1,1\}^{I(F) \cap J}} \left|\mathbf{E}\,\left(e^f | F^{\epsilon}\right)\right|. \tag{13}$$

On the other hand,

$$\left|\mathbf{E}\,\left(e^f | F\right)\right| \;\geq\; \left(\frac{\tau}{2}\right)^k \sum_{\epsilon \in \{-1,1\}^{I(F) \cap J}} \left|\mathbf{E}\,\left(e^f | F^{\epsilon}\right)\right|. \tag{14}$$

Comparing (12), (13), and (14), we conclude that

$$|g'(\alpha)| = \left|\frac{\partial}{\partial \alpha_J} \ln \mathbf{E}\,\left(e^f | F\right)\right| \;\leq\; \frac{1}{\tau^k} \;\leq\; \frac{1}{\tau^d}.$$

Then

$$\left|\ln \mathbf{E}\,\left(e^f | F\right) - \ln \mathbf{E}\,\left(\widehat{e^f} | F\right)\right| = |g(\alpha_J) - g(-\alpha_J)| \;\leq\; 2|\alpha_J| \max_{|\alpha| \leq |\alpha_J|} \left|g'(\alpha)\right| \;\leq\; \frac{2|\alpha_J|}{\tau^d}$$

and the proof follows. □

Lemma 4.2 *Let $\theta \geq 0$ and $\delta > 0$ be real numbers such that $\theta\delta < \pi$, let $F \subseteq \{-1,1\}^n$ be a face such that $\dim F < n$ and suppose that $\mathbf{E}\,\left(e^f | F\right) \neq 0$ for all $f \in \mathcal{U}(\delta)$. Assume that for any $f \in \mathcal{U}(\delta)$, for any $J \subset \{1,\dots,n\}$ such that $|J| \leq d$,* *and for the polynomial $\widehat{f}$ obtained from f by changing the coefficient α_J to $-\alpha_J$ and leaving all other coefficients intact, the angle between non-zero complex numbers* $\mathbf{E}\left(e^f|F\right)$ *and* $\mathbf{E}\left(e^{\widehat{f}}|F\right)$ *does not exceed* $\theta|\alpha_J|$.

Suppose that $\widehat{F} \subset \{-1,1\}^n$ *is a face obtained from F by changing the sign of one of the coordinates in* $I_+(F) \cup I_-(F)$. *Then* $G = F \cup \widehat{F}$ *is a face of* $\{-1,1\}^n$ *and for*

$$\tau = \cos\frac{\theta\delta}{2}$$

we have

$$\left|\mathbf{E}\left(e^f|G\right)\right| \geq \frac{\tau}{2}\left(\left|\mathbf{E}\left(e^f|F\right)\right| + \left|\mathbf{E}\left(e^f|\widehat{F}\right)\right|\right)$$

for any $f \in \mathcal{U}(\delta)$.

Proof Suppose that $\widehat{F}$ is obtained from F by changing the sign of the i-th coordinate. Let $\tilde{f}$ be a polynomial obtained from f by replacing the coefficients α_I by $-\alpha_I$ whenever $i \in I$ and leaving all other coefficients intact. Then $\tilde{f} \in \mathcal{U}(\delta)$ and the angle between $\mathbf{E}\left(e^f|F\right)$ and $\mathbf{E}\left(e^{\tilde{f}}|F\right)$ does not exceed

$$\theta \sum_{I:\ i\in I} |\alpha_I| \leq \theta\delta.$$

On the other hand, $\mathbf{E}\left(e^{\tilde{f}}|F\right) = \mathbf{E}\left(e^f|\widehat{F}\right)$ and

$$\mathbf{E}\left(e^f|G\right) = \frac{1}{2}\mathbf{E}\left(e^f|F\right) + \frac{1}{2}\mathbf{E}\left(e^f|\widehat{F}\right) = \frac{1}{2}\mathbf{E}\left(e^f|F\right) + \frac{1}{2}\mathbf{E}\left(e^{\tilde{f}}|F\right).$$

Thus $\mathbf{E}\left(e^f|G\right)$ is the sum of two non-zero complex numbers, the angle between which does not exceed $\theta\delta < \pi$. Interpreting the complex numbers as vectors in $\mathbb{R}^2 = \mathbb{C}$, we conclude that the length of the sum is at least as large as the length of the sum of the orthogonal projections of the vectors onto the bisector of the angle between them, and the proof follows. □

Proof of Theorem 1.1 Let us denote $d = \deg f$.

One can observe that the equation

$$\frac{2}{\cos\left(\dfrac{\theta\beta}{2}\right)} = \theta$$

has a solution $\theta \geq 0$ for all sufficiently small $\beta > 0$. Numerical computations show that one can choose

$$\beta = 0.55,$$

in which case

$$\theta \approx 2.748136091.$$

Let

$$\delta = \frac{\beta}{\sqrt{d}} = \frac{0.55}{\sqrt{d}}.$$

We observe that

$$0 \ < \ \theta\delta \ \leq \ \theta\beta \approx 1.511474850 \ < \ \pi.$$

Let

$$\tau = \cos\frac{\theta\delta}{2} = \cos\frac{\theta\beta}{2\sqrt{d}}.$$

In particular,

$$\tau \ \geq \ \cos\frac{\theta\beta}{2} \approx 0.7277659962.$$

Next, we will use the inequality

$$\left(\cos\frac{\alpha}{\sqrt{d}}\right)^d \ \geq \ \cos\alpha \quad \text{for} \quad 0 \leq \alpha \leq \frac{\pi}{2} \quad \text{and} \quad d \ \geq \ 1. \tag{15}$$

One can obtain (15) as follows. Since $\tan(0) = 0$ and the function $\tan\alpha$ is convex for $0 \leq \alpha < \pi/2$, we have

$$\sqrt{d}\tan\frac{\alpha}{\sqrt{d}} \ \leq \ \tan\alpha \quad \text{for} \quad 0 \leq \alpha < \frac{\pi}{2}.$$

Integrating, we obtain

$$d\ln\cos\frac{\alpha}{\sqrt{d}} \ \geq \ \ln\cos\alpha \quad \text{for} \quad 0 \leq \alpha < \frac{\pi}{2}$$

and (15) follows.

Using (15), we obtain

$$\frac{2}{\left(\cos\frac{\theta\delta}{2}\right)^d} = \frac{2}{\left(\cos\frac{\theta\beta}{2\sqrt{d}}\right)^d} \leq \frac{2}{\cos\left(\frac{\theta\beta}{2}\right)} = \theta. \tag{16}$$

We prove by induction on $m = 0, 1, \ldots, n$ the following three statements.

1. Let $F \subset \{-1, 1\}^n$ be a face of dimension m. Then, for any $f \in \mathcal{U}(\delta)$, we have $\mathbf{E}\left(e^f|F\right) \neq 0$.
2. Let $F \subset \{-1, 1\}^n$ be a face of dimension m, let $f \in \mathcal{U}(\delta)$ and let $\widehat{f}$ be a polynomial obtained from f by changing one of the coefficients α_J to $-\alpha_J$ and leaving all other coefficients intact. Then the angle between two non-zero complex numbers $\mathbf{E}\left(e^f|F\right)$ and $\mathbf{E}\left(e^{\widehat{f}}|F\right)$ does not exceed $\theta|\alpha_J|$.
3. Let $F \subset \{-1, 1\}^n$ be a face of dimension m and let $f \in \mathcal{U}(\delta)$. Assuming that $m > 0$ and hence $I(F) \neq \emptyset$, let us choose any $i \in I(F)$ and let F^+ and F^- be the corresponding faces of F obtained by fixing $x_i = 1$ and $x_i = -1$ respectively. Then

$$\left|\mathbf{E}\left(e^f|F\right)\right| \geq \frac{\tau}{2}\left(\left|\mathbf{E}\left(e^f|F^+\right)\right| + \left|\mathbf{E}\left(e^f|F^-\right)\right|\right).$$

If $m = 0$ then F consists of a single point $x \in \{-1, 1\}^n$, so

$$\mathbf{E}\left(e^f|F\right) = e^{f(x)} \neq 0$$

and statement 1 holds. Assuming that $\widehat{f}$ is obtained from f by replacing the coefficient α_J with $-\alpha_J$ and leaving all other coefficients intact, we get

$$\frac{\mathbf{E}\left(e^f|F\right)}{\mathbf{E}\left(e^{\widehat{f}}|F\right)} = \exp\left\{2\alpha_J \mathbf{x}^J\right\}.$$

Since

$$|2\alpha_J \mathbf{x}^J| = 2|\alpha_J| \leq \theta|\alpha_J|,$$

the angle between $\mathbf{E}\left(e^f|F\right)$ and $\mathbf{E}\left(e^{\widehat{f}}|F\right)$ does not exceed $\theta|\alpha_J|$ and statement 2 follows. The statement 3 is vacuous for $m = 0$.

Suppose that statements 1 and 2 hold for faces of dimension $m < n$. Lemma 4.2 implies that if F is a face of dimension $m + 1$ and F^+ and F^- are m-dimensional faces obtained by fixing x_i for some $i \in I(F)$ to $x_i = 1$ and $x_i = -1$ respectively,

then

$$\begin{aligned}\left|\mathbf{E}\left(e^f|F\right)\right| \geq & \left(\cos\frac{\theta\delta}{2}\right)\frac{\left|\mathbf{E}\left(e^f|F^+\right)\right| + \left|\mathbf{E}\left(e^f|F^-\right)\right|}{2}\\ =&\frac{\tau}{2}\left(\left|\mathbf{E}\left(e^f|F^+\right)\right| + \left|\mathbf{E}\left(e^f|F^-\right)\right|\right)\end{aligned}$$

and the statement 3 holds for $(m+1)$-dimensional faces.

The statement 3 for $(m+1)$-dimensional faces and the statement 1 for m-dimensional faces imply the statement 1 for $(m+1)$-dimensional faces.

Finally, suppose that the statements 1 and 3 hold for all faces of dimension at most $m+1$. Let us pick a face $F \subset \{-1,1\}^n$ of dimension $m+1$, where $0 \leq m < n$. Applying the condition of statement 3 recursively to the faces of F, we get that for any $K \subset I(F)$,

$$\left|\mathbf{E}\left(e^f|F\right)\right| \geq \left(\frac{\tau}{2}\right)^{|K|} \sum_{\epsilon\in\{-1,1\}^K} \left|\mathbf{E}\left(e^f|F^\epsilon\right)\right|.$$

Then, by Lemma 4.1, the angle between two non-zero complex numbers $\mathbf{E}\left(e^f|F\right)$ and $\mathbf{E}\left(\widehat{e^f}|F\right)$ does not exceed

$$\frac{2|\alpha_J|}{\tau^d} = \frac{2|\alpha_J|}{\left(\cos\frac{\theta\delta}{2}\right)^d} \leq \theta|\alpha_J|$$

by (16), and the statement 2 follows for faces of dimension $m+1$.

This proves that statements 1–3 hold for faces F of all dimensions. Iterating statement 3, we obtain that for any $f \in \mathcal{U}(\delta)$, we have

$$\left|\mathbf{E}\, e^f\right| \geq \left(\frac{\tau}{2}\right)^n \sum_{x\in\{-1,1\}^n} |e^{f(x)}|.$$

Since for any $x \in \{-1,1\}^n$ and for any $f \in \mathcal{U}(\delta)$, we have

$$|f(x)| \leq \sum_{i=1}^{n} \sum_{\substack{I\subset\{1,\ldots,n\}\\ i\in I}} |\alpha_I| \leq n\delta \leq \beta n,$$

we conclude that

$$\left|\mathbf{E}\, e^f\right| \geq \tau^n e^{-\beta n} \geq (0.41)^n.$$

The proof follows since if $f : \{-1,1\}^n \longrightarrow \mathbb{C}$ is a polynomial with zero constant term and

$$|\lambda| \ \leq \ \frac{0.55}{L(f)\sqrt{\deg f}},$$

then $\lambda f \in \mathcal{U}(\delta)$. □

5 Proofs of Theorems 2.1 and 2.2

The proofs of Theorems 2.1 and 2.2 are based on the following lemma.

Lemma 5.1 *Let*

$$f(x) = \sum_{I \in \mathcal{F}} \alpha_I \mathbf{x}^I$$

be a polynomial such that $\alpha_I \geq 0$ *for all* $I \in \mathcal{F}$*. Then*

$$\mathbf{E}\, e^f \ \geq \ \prod_{I \in \mathcal{F}} \left(\frac{e^{\alpha_I} + e^{-\alpha_I}}{2} \right).$$

Proof Since

$$e^{\alpha x} = \left(\frac{e^{\alpha} + e^{-\alpha}}{2} \right) + x \left(\frac{e^{\alpha} - e^{-\alpha}}{2} \right) \quad \text{for} \quad x = \pm 1,$$

we have

$$\mathbf{E}\, e^f = \mathbf{E} \prod_{I \in \mathcal{F}} e^{\alpha_I \mathbf{x}^I} = \mathbf{E} \prod_{I \in \mathcal{F}} \left(\left(\frac{e^{\alpha_I} + e^{-\alpha_I}}{2} \right) + \mathbf{x}^I \left(\frac{e^{\alpha_I} - e^{-\alpha_I}}{2} \right) \right). \tag{17}$$

Since

$$\frac{e^{\alpha_I} - e^{-\alpha_I}}{2} \geq 0 \quad \text{provided} \quad \alpha_I \geq 0$$

and

$$\mathbf{E} \left(\mathbf{x}^{I_1} \cdots \mathbf{x}^{I_k} \right) \ \geq \ 0 \quad \text{for all} \quad I_1, \ldots, I_k,$$

expanding the product in (17) and taking the expectation, we get the desired inequality. □

Next, we prove a similar estimate for functions f that allow some monomials with negative coefficients.

Lemma 5.2 *Let* $f(x) = g(x) - h(x)$ *where*

$$g(x) = \sum_{I \in \mathcal{G}} \mathbf{x}^I, \quad h(x) = \sum_{I \in \mathcal{H}} \mathbf{x}^I, \quad \mathcal{G} \cap \mathcal{H} = \emptyset.$$

Suppose that the constant terms of g *and* h *are* 0 *and that every variable* x_i *enters not more than* k *monomials of* f *for some integer* $k > 0$. *Then*

$$\mathbf{E}\, e^{\lambda f} \;\geq\; \exp\left\{ \frac{3\lambda^2}{8} \left(|\mathcal{G}| - (k-1)|\mathcal{H}|\right) \right\} \quad \textit{for} \quad 0 \leq \lambda \leq 1.$$

Proof Since $\mathbf{E} f = 0$, by Jensen's inequality we have

$$\mathbf{E}\, e^{\lambda f} \;\geq\; 1$$

and the estimate follows if $|\mathcal{G}| \leq (k-1)|\mathcal{H}|$. Hence we may assume that $|\mathcal{G}| > (k-1)|\mathcal{H}|$.

Given a function $f : \{-1, 1\}^n \longrightarrow \mathbb{R}$ and a set $J \subset \{1, \ldots, n\}$ of indices, we define a function (conditional expectation) $f_J : \{-1, 1\}^{n-|J|} \longrightarrow \mathbb{R}$ obtained by averaging over variables x_j with $j \in J$:

$$f_J\left(x_i :\; i \notin J\right) = \frac{1}{2^{|J|}} \sum_{\substack{x_j = \pm 1 \\ j \in J}} f\left(x_1, \ldots, x_n\right).$$

In particular, $f_J = f$ if $J = \emptyset$ and $f_J = \mathbf{E} f$ if $J = \{1, \ldots, n\}$. We obtain the monomial expansion of f_J by erasing all monomials of f that contain x_j with $j \in J$. By Jensen's inequality we have

$$\mathbf{E}\, e^{\lambda f} \;\geq\; \mathbf{E}\, e^{\lambda f_J} \quad \text{for all real} \quad \lambda. \tag{18}$$

Let us choose a set J of indices with $|J| \leq |\mathcal{H}|$ such that every monomial in $h(x)$ contains at least one variable x_j with $j \in J$. Then every variable x_j with $j \in J$ is contained in at most $(k-1)$ monomials of $g(x)$ and hence f_J is a sum of at least $|\mathcal{G}| - (k-1)|\mathcal{H}|$ monomials.

From (18) and Lemma 5.1, we obtain

$$\mathbf{E}\, e^{\lambda f} \;\geq\; \mathbf{E}\, e^{\lambda f_J} \;\geq\; \left(\frac{e^{\lambda} + e^{-\lambda}}{2} \right)^{|\mathcal{G}| - (k-1)|\mathcal{H}|} \;\geq\; \left(1 + \frac{\lambda^2}{2} \right)^{|\mathcal{G}| - (k-1)|\mathcal{H}|}.$$

Using that

$$\ln(1+x) \;\geq\; x - \frac{x^2}{2} = x\left(1 - \frac{x}{2}\right) \quad \text{for} \quad x \geq 0, \tag{19}$$

we conclude that

$$\mathbf{E}\, e^{\lambda f} \;\geq\; \exp\left\{\frac{\lambda^2}{2}\left(1 - \frac{\lambda^2}{4}\right)(|\mathcal{G}| - (k-1)|\mathcal{H}|)\right\} \;\geq\; \exp\left\{\frac{3\lambda^2}{8}(|\mathcal{G}| - (k-1)|\mathcal{H}|)\right\}$$

as desired. □

Now we are ready to prove Theorem 2.2.

Proof of Theorem 2.2 Let $x_0 \in \{-1,1\}^n$, $x_0 = (\xi_1, \ldots, \xi_n)$ be a maximum point of f, so that

$$\max_{x \in \{-1,1\}^n} f(x) = f(x_0).$$

Let us define $\tilde{f} : \{-1,1\}^n \longrightarrow \mathbb{R}$ by

$$\tilde{f}(x_1, \ldots, x_n) = f(\xi_1 x_1, \ldots, \xi_n x_n).$$

Then

$$\max_{x \in \{-1,1\}^n} f(x) = \max_{x \in \{-1,1\}^n} \tilde{f}(x), \quad \mathbf{E}\, e^{\lambda f} = \mathbf{E}\, e^{\lambda \tilde{f}}$$

and the maximum value of $\tilde{f}$ on the cube $\{-1,1\}^n$ is attained at $u = (1, \ldots, 1)$. Hence without loss of generality, we may assume that the maximum value of f on the cube $\{-1,1\}^n$ is attained at $u = (1, \ldots, 1)$.

We write

$$f(x) = g(x) - h(x) \quad \text{where} \quad g(x) = \sum_{I \in \mathcal{G}} \mathbf{x}^I \quad \text{and} \quad h(x) = \sum_{I \in \mathcal{H}} \mathbf{x}^I$$

for some disjoint sets $\mathcal{G}$ and $\mathcal{H}$ of indices. Moreover,

$$\max_{x \in \{-1,1\}^n} f(x) = f(u) = |\mathcal{G}| - |\mathcal{H}| \;\geq\; \frac{k-1}{k}|\mathcal{F}|.$$

Since

$$|\mathcal{G}| + |\mathcal{H}| = |\mathcal{F}|,$$

we conclude that

$$|\mathcal{G}| \ \geq \ \frac{2k-1}{2k}|\mathcal{F}| \quad \text{and} \quad |\mathcal{H}| \ \leq \ \frac{1}{2k}|\mathcal{F}|.$$

By Lemma 5.2,

$$\mathbf{E}\, e^{\lambda f} \ \geq \ \exp\left\{\frac{3\lambda^2}{8}\left(|\mathcal{G}|-(k-1)|\mathcal{H}|\right)\right\} \ \geq \ \exp\left\{\frac{3\lambda^2}{16}|\mathcal{F}|\right\}$$

as desired. □

To prove Theorem 2.1, we need to handle negative terms with more care.

Lemma 5.3 *Let* $f(x) = g(x) - h(x)$ *where*

$$g(x) = \sum_{I \in \mathcal{G}} \mathbf{x}^I, \quad h(x) = \sum_{I \in \mathcal{H}} \mathbf{x}^I, \quad \mathcal{G} \cap \mathcal{H} = \emptyset$$

and

$$|\mathcal{G}| \ \geq \ |\mathcal{H}|.$$

Suppose that the constant terms of g *and* h *are* 0 *and that the supports* $I \in \mathcal{H}$ *of monomials in* $h(x)$ *are pairwise disjoint. Then*

$$\mathbf{E}\, e^{\lambda f} \ \geq \ \exp\left\{\frac{3\lambda^2}{8}\left(\sqrt{|\mathcal{G}|}-\sqrt{|\mathcal{H}|}\right)^2\right\} \quad \text{for} \quad 0 \leq \lambda \leq 1.$$

Proof By Jensen's inequality we have

$$\mathbf{E}\, e^{\lambda f} \ \geq \ \exp\{\lambda \mathbf{E} f\} = 1,$$

which proves the lemma in the case when $|\mathcal{G}| = |\mathcal{H}|$. Hence we may assume that $|\mathcal{G}| > |\mathcal{H}|$.

If $|\mathcal{H}| = 0$ then, applying Lemma 5.1, we obtain

$$\mathbf{E}\, e^{\lambda f} = \mathbf{E}\, e^{\lambda g} \ \geq \ \left(\frac{e^{\lambda}+e^{-\lambda}}{2}\right)^{|\mathcal{G}|} \ \geq \ \left(1+\frac{\lambda^2}{2}\right)^{|\mathcal{G}|}.$$

Using (19), we conclude that

$$\mathbf{E}\, e^{\lambda f} \ \geq \ \exp\left\{\frac{\lambda^2}{2}\left(1-\frac{\lambda^2}{4}\right)|\mathcal{G}|\right\} \ \geq \ \exp\left\{\frac{3\lambda^2}{8}|\mathcal{G}|\right\},$$

which proves the lemma in the case when $|\mathcal{H}| = 0$. Hence we may assume that $|\mathcal{G}| > |\mathcal{H}| > 0$.

Since the supports $I \in \mathcal{H}$ of monomials in h are pairwise disjoint, we have

$$\mathbf{E}\, e^{\lambda h} = \prod_{I \in \mathcal{H}} \mathbf{E}\, e^{\lambda \mathbf{x}^I} = \left(\frac{e^{\lambda} + e^{-\lambda}}{2} \right)^{|\mathcal{H}|}. \tag{20}$$

Let us choose real $p, q \geq 1$, to be specified later, such that

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Applying the Hölder inequality, we get

$$\mathbf{E}\, e^{\lambda g/p} = \mathbf{E}\, \left(e^{\lambda f/p} e^{\lambda h/p} \right) \ \leq\ \left(\mathbf{E}\, e^{\lambda f} \right)^{1/p} \left(\mathbf{E}\, e^{\lambda q h/p} \right)^{1/q}$$

and hence

$$\mathbf{E}\, e^{\lambda f} \ \geq\ \frac{\left(\mathbf{E}\, e^{\lambda g/p} \right)^p}{\left(\mathbf{E}\, e^{\lambda q h/p} \right)^{p/q}}.$$

Applying Lemma 5.1 and formula (20), we obtain

$$\mathbf{E}\, e^{\lambda f} \ \geq\ \left(\frac{e^{\lambda/p} + e^{-\lambda/p}}{2} \right)^{|\mathcal{G}|p} \left(\frac{e^{\lambda q/p} + e^{-\lambda q/p}}{2} \right)^{-|\mathcal{H}|p/q}.$$

Since

$$e^{x^2/2} \ \geq\ \frac{e^x + e^{-x}}{2} \ \geq\ 1 + \frac{x^2}{2} \quad \text{for} \quad x \geq 0,$$

we obtain

$$\mathbf{E}\, e^{\lambda f} \ \geq\ \left(1 + \frac{\lambda^2}{2p^2} \right)^{|\mathcal{G}|p} \exp\left\{ -\frac{\lambda^2 q |\mathcal{H}|}{2p} \right\}.$$

Applying (19), we obtain

$$\mathbf{E}\, e^{\lambda f} \ \geq\ \exp\left\{ \frac{\lambda^2 |\mathcal{G}|}{2p} - \frac{\lambda^2 q |\mathcal{H}|}{2p} - \frac{\lambda^4 |\mathcal{G}|}{8p^3} \right\}.$$

Let us choose

$$p = \frac{\sqrt{|\mathcal{G}|}}{\sqrt{|\mathcal{G}|} - \sqrt{|\mathcal{H}|}} \quad \text{and} \quad q = \frac{\sqrt{|\mathcal{G}|}}{\sqrt{|\mathcal{H}|}}.$$

Then

$$\begin{aligned}\mathbf{E}\, e^{\lambda f} \ &\geq \ \exp\left\{\frac{\lambda^2}{2}\left(\sqrt{|\mathcal{G}|}-\sqrt{|\mathcal{H}|}\right)^2 - \frac{\lambda^4\left(\sqrt{|\mathcal{G}|}-\sqrt{|\mathcal{H}|}\right)^3}{8\sqrt{|\mathcal{G}|}}\right\}\\ &= \ \exp\left\{\frac{\lambda^2}{2}\left(\sqrt{|\mathcal{G}|}-\sqrt{|\mathcal{H}|}\right)^2\left(1-\frac{\lambda^2\left(\sqrt{|\mathcal{G}|}-\sqrt{|\mathcal{H}|}\right)}{4\sqrt{|\mathcal{G}|}}\right)\right\}\\ &\geq \ \exp\left\{\frac{3\lambda^2}{8}\left(\sqrt{|\mathcal{G}|}-\sqrt{|\mathcal{H}|}\right)^2\right\}\end{aligned}$$

and the proof follows. □

Lemma 5.4 *Let* $f(x) = g(x) - h(x)$ *where*

$$g(x) = \sum_{I\in\mathcal{G}} \mathbf{x}^I, \quad h(x) = \sum_{I\in\mathcal{H}} \mathbf{x}^I, \quad \mathcal{G}\cap\mathcal{H} = \emptyset$$

and

$$|\mathcal{G}| \ \geq \ |\mathcal{H}|.$$

Suppose that the constant terms of g *and* h *are* 0, *that every variable* x_i *enters at most two monomials in* $h(x)$ *and that if* x_i *enters exactly two monomials in* $h(x)$ *then* x_i *enters at most two monomials in* $g(x)$. *Then for* $0 \leq \lambda \leq 1$, *we have*

$$\mathbf{E}\, e^{\lambda f} \ \geq \ \exp\left\{\frac{3\lambda^2}{8}\left(\sqrt{|\mathcal{G}|}-\sqrt{|\mathcal{H}|}\right)^2\right\}.$$

Proof We proceed by induction on the number k of variables x_i that enter exactly two monomials in $h(x)$. If $k = 0$ then the result follows by Lemma 5.3.

Suppose that $k > 0$ and that x_i is a variable that enters exactly two monomials in $h(x)$ and hence at most two monomials in $g(x)$. As in the proof of Lemma 5.2, let $f_i : \{0,1\}^{n-1} \longrightarrow \mathbb{R}$ be the polynomial obtained from f by averaging with respect to x_i. As in the proof of Lemma 5.2, we have

$$\mathbf{E}\, e^{\lambda f} \ \geq \ \mathbf{E}\, e^{\lambda f_i} \quad \text{where} \quad f_i(x) = \sum_{I\in\mathcal{G}_i} \mathbf{x}^I - \sum_{I\in\mathcal{H}_i} \mathbf{x}^I$$

and $\mathcal{G}_i$, respectively $\mathcal{H}_i$, is obtained from $\mathcal{G}$, respectively $\mathcal{H}$, by removing supports of monomials containing x_i. In particular,

$$|\mathcal{H}_i| = |\mathcal{H}| - 2 \quad \text{and} \quad |\mathcal{G}_i| \geq |\mathcal{G}| - 2.$$

Applying the induction hypothesis to f_i, we obtain

$$\begin{aligned}\mathbf{E}\, e^{\lambda f} &\geq \mathbf{E}\, e^{\lambda f_i} \geq \exp\left\{\frac{3\lambda^2}{8}\left(\sqrt{|\mathcal{G}_i|}-\sqrt{|\mathcal{H}_i|}\right)^2\right\} \\ &\geq \exp\left\{\frac{3\lambda^2}{8}\left(\sqrt{|\mathcal{G}|-2}-\sqrt{|\mathcal{H}|-2}\right)^2\right\} \geq \exp\left\{\frac{3\lambda^2}{8}\left(\sqrt{|\mathcal{G}|}-\sqrt{|\mathcal{H}|}\right)^2\right\}\end{aligned}$$

and the proof follows. □

Finally, we are ready to prove Theorem 2.1.

Proof of Theorem 2.1 As in the proof of Theorem 2.2, without loss of generality we may assume that the maximum of f is attained at $u = (1, \ldots, 1)$.

We write

$$f(x) = g(x) - h(x) \quad \text{where} \quad g(x) = \sum_{I \in \mathcal{G}} \mathbf{x}^I \quad \text{and} \quad h(x) = \sum_{I \in \mathcal{H}} \mathbf{x}^I$$

for some disjoint sets $\mathcal{G}$ and $\mathcal{H}$ of indices. Moreover,

$$\max_{x \in \{-1,1\}^n} f(x) = f(u) = |\mathcal{G}| - |\mathcal{H}| = \delta |\mathcal{F}|.$$

Since

$$|\mathcal{G}| + |\mathcal{H}| = |\mathcal{F}|,$$

we conclude that

$$|\mathcal{G}| = \frac{1+\delta}{2}|\mathcal{F}| \quad \text{and} \quad |\mathcal{H}| = \frac{1-\delta}{2}|\mathcal{F}|. \tag{21}$$

For $i = 1, \ldots, n$ let μ_i^+ be the number of monomials in g that contain variable i and let μ_i^- be the number of monomials in h that contain x_i. Then

$$\mu_i^+ + \mu_i^- \leq 4 \quad \text{for} \quad i = 1, \ldots, n. \tag{22}$$

If for some i we have $\mu_i^+ < \mu_i^-$ then for the point u_i obtained from u by switching the sign of the i-th coordinate, we have

$$f(u_i) = \left(|\mathcal{G}| - 2\mu_i^+\right) - \left(|\mathcal{H}| - 2\mu_i^-\right) = |\mathcal{G}| - |\mathcal{H}| + 2\left(\mu_i^- - \mu_i^+\right) > f(u),$$

contradicting that u is a maximum point of f. Therefore,

$$\mu_i^+ \geq \mu_i^- \quad \text{for} \quad i = 1, \ldots, n$$

and, in view of (22), we conclude that

$$\mu_i^- \ \leq\ 2 \quad \text{for} \quad i = 1, \ldots, n \quad \text{and if} \quad \mu_i^- = 2 \quad \text{then} \quad \mu_i^+ = 2.$$

By Lemma 5.4,

$$\mathbf{E}\, e^{\lambda f} \ \geq\ \exp\left\{ \frac{3\lambda^2}{8} \left(\sqrt{|\mathcal{G}|} - \sqrt{|\mathcal{H}|} \right)^2 \right\}.$$

Using (21), we deduce that

$$\begin{aligned} \mathbf{E}\, e^{\lambda f} \ &\geq\ \exp\left\{ \frac{3\lambda^2}{8} \left(\sqrt{\frac{1+\delta}{2}} - \sqrt{\frac{1-\delta}{2}} \right)^2 |\mathcal{F}| \right\} \\ &= \exp\left\{ \frac{3\lambda^2}{8} \left(1 - \sqrt{1-\delta^2} \right) |\mathcal{F}| \right\} \ \geq\ \exp\left\{ \frac{3\lambda^2\delta^2}{16} |\mathcal{F}| \right\}, \end{aligned}$$

which completes the proof. □

Acknowledgements I am grateful to Johan Håstad for advice and references on optimizing a polynomial on the Boolean cube and to the anonymous referees for careful reading of the paper and useful suggestions.

References

1. A. Bandyopadhyay, D. Gamarnik, Counting without sampling: asymptotics of the log-partition function for certain statistical physics models. Random Struct. Algorithm **33**(4), 452–479 (2008)
2. B. Barak, A. Moitra, R. O'Donnell, P. Raghavendra, O. Regev, D. Steurer, L. Trevisan, A. Vijayaraghavan, D. Witmer, J. Wright, Beating the random assignment on constraint satisfaction problems of bounded degree, in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques.* LIPIcs. Leibniz International Proceedings in Informatics, vol. 40 (Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2015), pp. 110–123
3. A. Barvinok, Computing the partition function for cliques in a graph. Theor. Comput. **11**, Article 13, 339–355 (2015)
4. A. Barvinok, Computing the permanent of (some) complex matrices. Found. Comput. Math. **16**(2), 329–342 (2016)
5. A. Barvinok, P. Soberón, Computing the partition function for graph homomorphisms. Combinatorica (2016). doi:10.1007/s00493-016-3357-2
6. A. Barvinok, P. Soberón, Computing the partition function for graph homomorphisms with multiplicities. J. Comb. Theory Ser. A **137**, 1–26 (2016)
7. E. Boros, P.L. Hammer, Pseudo-boolean optimization. Discret. Appl. Math. **123**(1–3), 155–225 (2002). *Workshop on Discrete Optimization* (DO'99), Piscataway
8. R.L. Dobrushin, S.B. Shlosman, Completely analytical interactions: constructive description. J. Stat. Phys. **46**(5–6), 983–1014 (1987)

9. J. Håstad, On bounded occurrence constraint satisfaction. Inf. Process. Lett. **74**(1–2), 1–6 (2000)
10. J. Håstad, Some optimal inapproximability results. J. ACM **48**(4), 798–859 (2001)
11. J. Håstad, Improved bounds for bounded occurrence constraint satisfaction, manuscript (2005). Available at https://www.nada.kth.se/~johanh/bounded2.pdf
12. J. Håstad, S. Venkatesh, On the advantage over a random assignment. Random Struct. Algorithm **25**(2), 117–149 (2004)
13. J. Kahn, N. Linial, A. Samorodnitsky, Inclusion-exclusion: exact and approximate. Combinatorica **16**(4), 465–477 (1996)
14. S. Khot, A. Naor, Grothendieck-type inequalities in combinatorial optimization. Commun. Pure Appl. Math. **65**(7), 992–1035 (2012)
15. T.D. Lee, C.N. Yang, Statistical theory of equations of state and phase transitions. II. Lattice gas and Ising model. Phys. Rev. (2) **87**, 410–419 (1952)
16. G. Regts, Zero-free regions of partition functions with applications to algorithms and graph limits. Combinatorica (2017). doi:10.1007/s00493-016-3506-7
17. D. Weitz, Counting independent sets up to the tree threshold, in *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, STOC'06 (ACM, New York, 2006), pp. 140–149
18. C.N. Yang, T.D. Lee, Statistical theory of equations of state and phase transitions. I. Theory of condensation. Phys. Rev. (2) **87**, 404–409 (1952)
19. D. Zuckerman, Linear degree extractors and the inapproximability of max clique and chromatic number. Theor. Comput. **3**, 103–1283 (2007)

Siegel's Lemma Is Sharp

József Beck

Abstract Siegel's Lemma is concerned with finding a "small" nontrivial integer solution of a large system of homogeneous linear equations with integer coefficients, where the number of variables substantially exceeds the number of equations (for example, n equations and N variables with $N \geq 2n$), and "small" means small in the maximum norm. Siegel's Lemma is a clever application of the Pigeonhole Principle, and it is a pure existence argument. The basically combinatorial Siegel's Lemma is a key tool in transcendental number theory and diophantine approximation. David Masser (a leading expert in transcendental number theory) asked the question whether or not the Siegel's Lemma is best possible. Here we prove that the so-called "Third Version of Siegel's Lemma" is best possible apart from an absolute constant factor. In other words, we show that no other argument can beat the Pigeonhole Principle proof of Siegel's Lemma (apart from an absolute constant factor). To prove this, we combine a concentration inequality (i.e., Fourier analysis) with combinatorics.

1 Introduction

What we study is a discrepancy problem at the crossroads of number theory and combinatorics. It is about the sharpness of the well-known Siegel's Lemma, which was formally introduced in 1929 (but it was already used by others before, e.g., Thue already used "Siegel's Lemma" in his famous paper from 1909). Siegel's Lemma is a key tool in transcendental number theory and diophantine approximation; see e.g. Lemma 1 of Chapter 2 in Baker's well known book [3]. (In particular, Siegel's Lemma is a key step in the applications of the method of "constructing auxiliary polynomials in several variables", that includes the so-called Thue method. The two most famous applications of the Thue method are Roth's $2 + \epsilon$ theorem and Schmidt's Subspace Theorem, representing the main results in the theory of rational approximations of algebraic numbers. Another famous application of Siegel's Lemma is Baker's method about linear form in logarithms.)

J. Beck (✉)
Mathematics Department, Rutgers University, New Brunswick, NJ, USA
e-mail: jbeck@math.rutgers.edu

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_8

To formulate the original form of Siegel's Lemma, we consider a linear system

$$\sum_{1\le j\le N} d_{i,j}x_j = 0, \quad 1 \le i \le n$$

with n equations and N variables, where $N > n$, and for all coefficients we have $d_{i,j} \in \mathbb{Z}$ with $|d_{i,j}| \le A$. (Note that Siegel's Lemma was enormously generalized in the last 30 years in number theory to such an extent that the original form is almost unrecognizable. Here we do not discuss these far-reaching generalizations that include different "heights" and "subspace" versions; see e.g. the book [4].) Of course we can rewrite the linear system in the short matrix form $\mathbf{Dx} = \mathbf{0}$, where the matrix $\mathbf{D} = (d_{i,j})$, $1 \le i \le n$, $1 \le j \le N$ has n rows and N columns. The problem is to give an upper bound to the maximum norm of the smallest nontrivial solution $\mathbf{x} = (x_1, \ldots, x_N) \in \mathbb{Z}^N \setminus \mathbf{0}$. That is, we are looking for the minimum of

$$\max_{1\le j\le N} |x_j|.$$

Note that the maximum norm is the hard one; it is substantially harder than (say) the euclidean norm (see Vaaler [11]).

The simplest statement of the original Siegel's Lemma goes as follows. Consider all integer vectors $\mathbf{v} = (v_1, \ldots, v_N) \in \mathbb{Z}^N \setminus \mathbf{0}$ with $0 \le v_j \le B$, $1 \le j \le N$, where B is a positive integer to be specified later. Note that every row-sum $\sum_{1\le j\le N} d_{i,j}v_j$ $(1 \le i \le n)$ is in an interval of at most NBA integers, so, if

$$(B+1)^N > (NBA)^n, \tag{1}$$

then the pigeonhole principle implies that there exist two different vectors $\mathbf{v}_h = (v_{h,1}, \ldots, v_{h,N}) \in \mathbb{Z}^N \setminus \mathbf{0}$, $h = 1, 2$ with $0 \le v_{h,j} \le B$, $1 \le j \le N$ such that $\mathbf{Dv}_1 = \mathbf{Dv}_2$, and so $\mathbf{Dx} = \mathbf{0}$ with $\mathbf{x} = \mathbf{v}_1 - \mathbf{v}_2$. Thus we obtain a nontrivial solution $\mathbf{x} = (x_1, \ldots, x_N) \in \mathbb{Z}^N \setminus \mathbf{0}$ such that

$$\max_{1\le j\le N} |x_j| \le B.$$

Inequality (1) holds if

$$B = \left\lfloor (NA)^{n/(N-n)} \right\rfloor$$

(lower integral part), implying

$$\max_{1\le j\le N} |x_j| \le \left\lfloor (NA)^{n/(N-n)} \right\rfloor. \tag{2}$$

Note that (2) suffices for many number-theoretic applications (in fact, (2) is Lemma 1 of Chapter 2 in [3]), but, by using probability theory, we can substantially improve on (2). To apply probability theory, it is convenient to have zero expectation: we replace the integer vectors $\mathbf{v} = (v_1, \ldots, v_N) \in \mathbb{Z}^N \setminus \mathbf{0}$, $0 \le v_j \le B$ with $|v_j| \le B$, $1 \le j \le N$ (where B is a positive integer to be specified later). First, for illustration, we simply repeat the argument of (1)–(2) working with these new vectors that may have negative coordinates. If

$$(2B+1)^N > (2NBA+1)^n, \tag{3}$$

then the pigeonhole principle implies that there exist two different vectors $\mathbf{v}_h = (v_{h,1}, \ldots, v_{h,N}) \in \mathbb{Z}^N \setminus \mathbf{0}$, $h = 1,2$ with $|v_{h,j}| \le B$, $1 \le j \le N$ such that $\mathbf{D}\mathbf{v}_1 = \mathbf{D}\mathbf{v}_2$, and so $\mathbf{D}\mathbf{x} = \mathbf{0}$ with $\mathbf{x} = \mathbf{v}_1 - \mathbf{v}_2$. Thus we obtain a nontrivial solution $\mathbf{x} = (x_1, \ldots, x_N) \in \mathbb{Z}^N \setminus \mathbf{0}$ such that

$$\max_{1 \le j \le N} |x_j| \le 2B. \tag{4}$$

Inequality (3) holds if

$$2B = (2NA)^{n/(N-n)},$$

and so by (4),

$$\max_{1 \le j \le N} |x_j| \le (2NA)^{n/(N-n)},$$

which is basically the same as (2).

So far we did not take advantage of allowing negative coordinates (in fact, we got a slightly weaker result). The term $(2NBA+1)^n$ on the right-hand side of (3) is a trivial upper bound on the number of "pigeonholes", and it is based on the fact that NBA is a trivial upper bound for the absolute value of a row-sum. However, by allowing negative coordinates, we can apply the central limit theorem with expected value zero, which implies that a *typical* row-sum has absolute value $\le const\sqrt{N}BA$, and using the large deviation theorem, we can guarantee that a typical *maximum* row-sum has absolute value $\le const\sqrt{N}\sqrt{\log n} \cdot BA$. (Here the factor $\sqrt{\log n}$ comes from the superexponentially small tail of the normal distribution.) Thus, the application of the pigeonhole principle (1) or (3) is replaced by

$$\frac{1}{2}(2B+1)^N > (2const\sqrt{N\log n} \cdot BA + 1)^n. \tag{5}$$

That is, by using the large deviation theorem, we can substantially reduce the number of "pigeonholes". Inequality (5) holds if

$$2B = \left(c\sqrt{N\log n}A\right)^{n/(N-n)}, \tag{6}$$

where $c > 0$ is some absolute constant. So, by (4),

$$\max_{1\le j\le N} |x_j| \le \left(c\sqrt{N\log nA}\right)^{n/(N-n)}. \tag{7}$$

Since $\sqrt{N\log n} < \sqrt{N\log N} < N$ if N is large, (7) represents a substantial improvement on (2).

By using an extra twist in the argument above, we can even get rid of the relatively small factor $\sqrt{\log n}$ in (7). Indeed, a more sophisticated application of the large deviation theorem gives that, for a typical set of n row-sums we can guarantee the following: at least $n/2$ row-sums have absolute value $\le c\sqrt{N}BA$, at most $n/4$ row-sums have absolute value between $c\sqrt{N}BA$ and $2c\sqrt{N}BA$, at most $n/8$ row-sums have absolute value between $2c\sqrt{N}BA$ and $3c\sqrt{N}BA$, at most $n/16$ row-sums have absolute value between $3c\sqrt{N}BA$ and $4c\sqrt{N}BA$, and so on. This implies, via routine calculations, that the number of "pigeonholes" reduces to

$$\le (const\sqrt{N}BA)^n.$$

(Note that a similar argument shows up in Spencer [10].)

The choice

$$2B = \left(c\sqrt{N}A\right)^{n/(N-n)}$$

guarantees that

$$\frac{1}{2}(2B+1)^m > (const\sqrt{N}\cdot BA)^n \ge \text{number of pigenholes},$$

so, applying the pigeonhole principle the usual way, we have

$$\max_{1\le j\le N} |x_j| \le \left(c\sqrt{N}A\right)^{n/(N-n)}. \tag{8}$$

Here is a precise form of (8).

Third Version of Siegel's Lemma *Let*

$$\sum_{1\le j\le N} d_{i,j}x_j = 0, \quad 1 \le i \le n,$$

be a linear system such that $N > n$, and for all coefficients we have $d_{i,j} \in \mathbb{Z}$ with $|d_{i,j}| \le A$. Then there exists a nontrivial solution $\mathbf{x} = (x_1,\dots,x_N) \in \mathbb{Z}^N \setminus \mathbf{0}$ such that

$$\max_{1\le j\le N} |x_j| \le \left(70\sqrt{N}A\right)^{n/(N-n)}. \tag{9}$$

For the sake of completeness, and the convenience of the reader, we include a detailed proof of (9), which converts the intuition outlined above into a precise argument; see the "Appendix: Proof of the Third Version of Siegel's Lemma" at the end of the paper.

Note that in this paper we do not make any serious effort to find the best constant factors. For example, the constant factor 70 in (9) is far from optimal—in fact, it can be eliminated completely. Indeed, see e.g. the book Bombieri–Gubler [4], Ch.2, Section 2.9, Corollary 2.9.9, or the papers Bombieri–Vaaler [5] and Vaaler–van der Poorten [12] in the appropriate special case. The novelty of these proofs is to use arguments from the geometry of numbers, which is just another—more sophisticated—way of applying the pigeonhole principle.

The reason why we nevertheless include our "naive" pigeonhole principle proof in the Appendix (with the weaker constant factor 70) is that our proof is much more accessible for a typical combinatorist—the likely reader of this paper.

The Third Version raises the question: can the factor $\sqrt{N}A$ in (9) be further improved to $o(\sqrt{N}A)$?

It is easy to see that the factor A in (9) cannot be replaced by $o(A)$. Indeed, in the simplest case $N = 2$ and $n = 1$, we choose two different primes p and q in the interval $A/2 < p < q < A$ (A is large enough); then the equation $px_1 = qx_2$ has the property that for every nontrivial solution $\mathbf{x} = (x_1, x_2) \in \mathbb{Z}^2 \setminus \mathbf{0}$ there is a nonzero coordinate of $\mathbf{x}$ that is divisible by p or q, implying

$$\max_{1\le j\le 2} |x_j| \ge p > A/2.$$

In the general case $N > n \ge 1$, we refine the construction above as follows. Write $m = N - n$, and choose $n(m+1)$ primes $p_{i,j}$, $1 \le i \le n$, $1 \le j \le m+1$ between $(1-\varepsilon)A^{1/m}$ and $A^{1/m}$, where $0 < \varepsilon < \frac{1}{2n}$ and A is large. Write

$$P_i = \prod_{1\le j\le m+1} p_{i,j}, \quad 1 \le i \le n,$$

and consider the linear system

$$\sum_{1\le j\le m} \frac{P_i}{p_{i,j}} x_j = \frac{P_i}{p_{i,m+1}} x_{m+i}, \quad 1 \le i \le n \tag{10}$$

of n equations and $m+n = N$ variables. Every coefficient in (10) has absolute value $\le \left(A^{1/m}\right)^m = A$. Notice that every nontrivial solution $\mathbf{x} = (x_1, \dots, x_N) \in \mathbb{Z}^N \setminus \mathbf{0}$ of (10) has a nonzero coordinate x_j with some $1 \le j \le m$. Then by (10), x_j is divisible by the product

$$\prod_{1\le i\le n} p_{i,j},$$

implying that

$$|x_j| \geq \prod_{1\leq i\leq n} p_{i,j} > \left((1-\varepsilon)A^{1/m}\right)^n > \frac{1}{2}A^{n/(N-n)} \text{ since } \varepsilon < \frac{1}{2n}. \tag{11}$$

(11) proves that the factor A in (9) cannot be replaced by $o(A)$. (For a similar example, see also page 2 in Schmidt [9].)

This raises the question: Can one reduce the other factor $\sqrt{N}$ in (9)? For simplicity we just study the case

$$N \geq 3n/2. \tag{12}$$

Theorem 1 below shows that in the special case $A = 1$ and (12) the factor $\sqrt{N}$ in (9) cannot be replaced by $o\left(\sqrt{N}\right)$. Note that $A = 1$ is the most interesting special case, because we cannot take advantage of "large" coefficients.

Theorem 1 *There is a (small) positive absolute constant $c_0 > 0$ with the following property: for every pair $N > n \geq 1$ of positive integers satisfying (12), there exists a matrix $\mathbf{D} = (d_{i,j})$, $1 \leq i \leq n$, $1 \leq j \leq N$ with n rows, N columns, entries $d_{i,j} \in \{1, -1\}$ such that for every nontrivial integer solution*

$$\mathbf{x} = (x_1, x_2, \ldots, x_N) \in \mathbb{Z}^N \setminus \mathbf{0}$$

of the homogeneous linear system $\mathbf{Dx} = \mathbf{0}$, meaning the long form

$$\sum_{1\leq j\leq N} d_{i,j}x_j = 0, \quad 1 \leq i \leq n,$$

the maximum norm of $\mathbf{x}$ has the lower bound

$$\max_{1\leq j\leq N} |x_j| > c_0 \left(\sqrt{N}\right)^{n/(N-n)}.$$

Actually we have much more than pure existence: for large n the overwhelming majority of the n-by-N ± 1 matrices $\mathbf{D}$ satisfy the theorem. In fact, the violators $\mathbf{D} = (d_{i,j})$, $1 \leq i \leq n$, $1 \leq j \leq N$ of Theorem 1 represent an exponentially small $O(2^{-n/2})$ part of the total 2^{nN}.

For an explicit value of c_0 that works for all sufficiently large n; see (87).

Note that Theorem 1 is a "large discrepancy" type result in the following sense. A homogeneous linear system always has the trivial solution (with maximum norm zero). The message of Theorem 1 is that, for the overwhelming majority of large homogeneous linear systems, every nontrivial integer solution is "large" (in terms of the maximum norm). That is, there is a large discrepancy between the trivial and nontrivial integer solutions.

Next we explain how to extend Theorem 1 beyond the special case $A = 1$. For illustration, we start with the case where

$$A \approx N^{n/(2N-4n)} \text{ and } N \geq 3n. \tag{13}$$

Let p be a prime between $A/2$ and A, write $m = N - n$, and consider the linear system

$$\sum_{j=1}^{m} p d_{i,j} x_j = x_{m+i}, \quad 1 \leq i \leq n, \tag{14}$$

where $\mathbf{D} = (d_{i,j})$, $1 \leq i \leq n$, $1 \leq j \leq m$ is a matrix with n rows, m columns and entries $d_{i,j} \in \{1, -1\}$, satisfying Theorem 1 for the pair m, n. Note that $m = N-n \geq 2n$ follows from (13), so (12) clearly holds (for notational simplicity we just apply Theorem 1 in the range $N \geq 2n$). Let

$$\mathbf{x} = (x_1, x_2, \ldots, x_N) \in \mathbb{Z}^N \setminus \mathbf{0}$$

be a nontrivial integer solution of the homogeneous linear system (14).

If $x_{m+i} = 0$ for all $1 \leq i \leq n$, then (dividing by p) Theorem 1 yields,

$$\max_{1 \leq j \leq m} |x_j| > c_0 \left(\sqrt{m}\right)^{n/(m-n)}. \tag{15}$$

If $x_{m+i} \neq 0$ for some $1 \leq i \leq n$, then of course x_{m+i} is divisible by p, so

$$|x_{m+i}| \geq p > A/2. \tag{16}$$

Combining (15) and (16), we have

$$\max_{1 \leq j \leq N} |x_j| > \min\left\{c_0 \left(\sqrt{m}\right)^{n/(m-n)}, A/2\right\}. \tag{17}$$

On the other hand, assuming

$$c_0 \left(\sqrt{m}\right)^{n/(m-n)} = A/2, \tag{18}$$

we have

$$\min\left\{c_0 \left(\sqrt{m}\right)^{n/(m-n)}, A/2\right\} = \left(c_0 \left(\sqrt{m}\right)^{n/(m-n)}\right)^{(m-n)/(N-n)} (A/2)^{(N-m)/N-n)} =$$

$$= c_1 \left(\sqrt{m} A/2\right)^{n/(N-n)}. \tag{19}$$

Combining (17), (18), and (19), we have

$$\max_{1\leq j\leq N}|x_j| > c_1\left(\sqrt{N}A\right)^{n/(N-n)}, \tag{20}$$

if (13) holds. (20) proves the sharpness of (9) under the condition (13).

Next, let n be even, let p_1, p_2 be two different primes between $A/2$ and A, write $m = N-(n/2)$, and assume $m \geq 2n$. Consider the linear system

$$\sum_{j=1}^{m} p_1 d_{i,j}x_j = x_{m+i},\ \ 1\leq i\leq n/2,$$

$$\sum_{j=1}^{m} p_2 d_{i,j}x_j = x_{m+i-(n/2)},\ \ n/2 < i\leq n, \tag{21}$$

where $\mathbf{D} = (d_{i,j})$, $1\leq i\leq n$, $1\leq j\leq m$ is a matrix with n rows, m columns and entries $d_{i,j}\in\{1,-1\}$, satisfying Theorem 1 for the pair m,n (note that $m = N-(n/2)\geq 2n$ follows from the hypothesis). Let

$$\mathbf{x} = (x_1, x_2, \ldots, x_N)\in\mathbb{Z}^N\setminus\mathbf{0}$$

be a nontrivial integer solution of the homogeneous linear system (21).

If $x_{m+i} = 0$ for all $1\leq i\leq n/2$, then (dividing by p_1 and p_2, respectively) Theorem 1 yields,

$$\max_{1\leq j\leq m}|x_j| > c_0\left(\sqrt{m}\right)^{n/(m-n))}. \tag{22}$$

If $x_{m+i}\neq 0$ for some $1\leq i\leq n/2$, then of course x_{m+i} is divisible by both p_1 and p_2, so

$$|x_{m+i}|\geq p_1p_2 > A^2/4. \tag{23}$$

Combining (22) and (23), we have

$$\max_{1\leq j\leq N}|x_j| > \min\left\{c_0\left(\sqrt{m}\right)^{n/(m-n)}, A^2/4\right\}. \tag{24}$$

On the other hand, assuming

$$A\approx N^{n/(4N-6n)}\ \text{ and }\ N\geq 5n/2, \tag{25}$$

we have

$$\min\left\{c_0\left(\sqrt{m}\right)^{n/(m-n)}, A^2/4\right\} = \left(c_0\left(\sqrt{m}\right)^{n/(m-n)}\right)^{(m-n)/(N-n)}(A^2/4)^{(N-m)/N-n)} =$$

$$= c_1\left(\sqrt{m}A\right)^{n/(N-n)}. \tag{26}$$

(26) proves the sharpness of (9) under the condition (25). Notice that (25) is substantially different from (13).

In general, let $r \geq 3$ be an integer, let n be divisible by r, let $p_1, p_2, \ldots, p_r$ be r different primes between $(1-\frac{1}{r})A$ and A (n and A are sufficiently large), write $m = N - (n/r)$, and assume $m \geq 2n$. Consider the linear system

$$\sum_{j=1}^{m} p_1 d_{i,j}x_j = x_{m+i}, \quad 1 \leq i \leq n/r,$$

$$\sum_{j=1}^{m} p_2 d_{i,j}x_j = x_{m+i-(n/r)}, \quad n/r < i \leq 2n/r,$$

and so on, where the last block of n/r equations goes as follows:

$$\sum_{j=1}^{m} p_r d_{i,j}x_j = x_{m+i-((r-1)n/r)}, \quad (r-1)n/r < i \leq n, \tag{27}$$

where $\mathbf{D} = (d_{i,j})$, $1 \leq i \leq n$, $1 \leq j \leq m$ is a matrix with n rows, m columns and entries $d_{i,j} \in \{1,-1\}$, satisfying Theorem 1 for the pair m, n (note that $m = N - (n/r) \geq 2n$ follows from the hypothesis). Let

$$\mathbf{x} = (x_1, x_2, \ldots, x_N) \in \mathbb{Z}^N \setminus \mathbf{0}$$

be a nontrivial integer solution of the homogeneous linear system (27).

If $x_{m+i} = 0$ for all $1 \leq i \leq n/r$, then (dividing by $p_1, p_2, \ldots, p_r$, respectively) Theorem 1 yields,

$$\max_{1\leq j\leq m}|x_j| > c_0\left(\sqrt{m}\right)^{n/(m-n))}. \tag{28}$$

If $x_{m+i} \neq 0$ for some $1 \leq i \leq n/r$, then of course x_{m+i} is divisible by the r primes $p_1, p_2, \ldots, p_r$, so

$$|x_{m+i}| \geq p_1 p_2 \cdots p_r > \left(1-\frac{1}{r}\right)^r A^r > A^r/4. \tag{29}$$

Combining (28) and (29), we have

$$\max_{1\le j\le N} |x_j| > \min\left\{c_0\left(\sqrt{m}\right)^{n/(m-n)}, A^r/4\right\}. \tag{30}$$

On the other hand, assuming

$$A \approx N^{n/(2rN-(2r+2)n)} \text{ and } N \ge (2r+1)n/r, \tag{31}$$

we have

$$\min\left\{c_0\left(\sqrt{m}\right)^{n/(m-n)}, A^r/4\right\} = \left(c_0\left(\sqrt{m}\right)^{n/(m-n)}\right)^{(m-n)/(N-n)} (A^r/4)^{(N-m)/(N-n)} =$$

$$= c_1\left(\sqrt{m}A\right)^{n/(N-n)}. \tag{32}$$

(32) proves the sharpness of (9) under the condition (31). Notice that (31) with $r \ge 3$ is substantially different from (13) and (25).

Summarizing, (31)–(32) prove the sharpness of (9) for infinitely many very different types of triples N, n, A with $N \ge 2n$. These corollaries of Theorem 1 show that the Third Version (9) of Siegel's Lemma is best possible (apart from an absolute constant factor). Or, we may say that no other argument can beat the Pigeonhole Principle proof of Siegel's Lemma (apart from an absolute constant factor).

In the Remark at the end of Sect. 3 we explain why the proof technique of this paper (based on the concentration inequality Lemma 4) breaks down in the range

$$\left(\frac{3}{2}-\varepsilon\right)n > N > n.$$

We also discuss the "between range"

$$3n/2 > N \ge \left(\frac{3}{2}-\varepsilon\right)n,$$

see Theorem 2 there.

Theorem 1 is about the majority of the 2^{nN} homogeneous linear systems $\mathbf{Dx} = \mathbf{0}$, where $\mathbf{D} = (d_{i,j})$ with entries $d_{i,j} \in \{1, -1\}$, $1 \le i \le n$, $1 \le j \le N$, and it states that every nontrivial integer solution is "large".

In the other direction, it is easy to construct a huge family of homogeneous linear systems $\mathbf{Dx} = \mathbf{0}$ with entries $d_{i,j} \in \{1, -1\}$ such that there exists a small nontrivial integer solution. Indeed, assume that N is even, and for every i in $1 \le i \le n$, exactly half of the entries $d_{i,j} \in \{1, -1\}$, $1 \le j \le N$ are equal to 1. There are

$$\binom{N}{N/2}^n = \left((1+o(1))\frac{2^N}{\sqrt{\pi N/2}}\right)^n$$

such homogeneous linear systems $\mathbf{Dx} = \mathbf{0}$, and each has the small solution $(1, 1, 1, \ldots, 1)$ (for which the maximum norm is as small as possible).

Of course,

$$\binom{N}{N/2}^n = \left((1+o(1))\frac{2^N}{\sqrt{\pi N/2}}\right)^n \leq \frac{2^{nN}}{N^{n/2}} = o\left(2^{nN}\right)$$

represents a minority; nevertheless,

$$\binom{N}{N/2}^n = \left((1+o(1))\frac{2^N}{\sqrt{\pi N/2}}\right)^n \geq \frac{2^{nN}}{(2N)^{n/2}}$$

is certainly very large—it is in the "rough order of magnitude" of 2^{nN}.

For an analog of Theorem 1 for inhomogeneous linear systems; see Theorem 3 at the end of Sect. 3.

2 Proof of Theorem 1

We can clearly assume that n is "large". We can also assume that $N \leq n \log N$, since otherwise

$$\left(\sqrt{N}\right)^{n/(N-n)} \leq \exp\left(\frac{\frac{1}{2}\log N}{\log N - 1}\right) < 2,$$

and then already the trivial lower bound

$$\max_{1\leq j\leq N} |x_j| \geq 1$$

implies Theorem 1.

We cannot explicitly construct the desired matrix $\mathbf{D}$; we just prove the existence of such a matrix by applying the so-called "probabilistic method" (also called "Erdős's method"). It is very interesting that the proof of Siegel's Lemma and the proof of Theorem 1 are both non-constructive, pure existence arguments.

To apply the "probabilistic method", we need a "concentration inequality" in combinatorial number theory (see Lemma 4 below). Up to Lemma 4 we closely follow a paper of Halász [7]. Lemma 4 is not covered by any of the theorems in [7], but we use Halász's method to prove it.

We start with a simple lemma in additive measure theory.

Lemma 1 *Let U and V be two periodic sets on the real line, both having period one, and assume that both sets are Jordan measurable (i.e., the characteristic function of each set restricted to the unit interval* $[0, 1)$ *is Riemann integrable). Define the* *set-sum* $U + V = \{u + v : u \in U, v \in V\}$. *Then either* $U + V$ *is the whole line, or* $\mathrm{meas}(U + V) \geq \mathrm{meas}(U) + \mathrm{meas}(V)$. *Here* $\mathrm{meas}(W)$ *is the Lebesgue measure of* $W \cap [0, 1)$.

Remarks Lemma 1 is in Macbeath [8] who proved it in a few lines (in fact for Lebesgue measurable sets). Lemma 1 is also a simple corollary of the well known Cauchy–Davenport theorem in additive group theory. Indeed, the Cauchy–Davenport theorem states that, for any prime p and for any nonempty subsets S_1 and S_2 of the additive cyclic group $\mathbb{Z}/p\mathbb{Z}$ of order p, we have the inequality

$$|S_1 + S_2| \geq \min\{p, |S_1| + |S_2| - 1\}$$

(where we use the standard notation $|S|$ for the number of elements of a finite set S). We apply the unit circle representation of the group: the elements are the vertices of the regular p-gon on the unit circle with one vertex at $(1, 0)$, and the group operation means to add the angles.

Since the Riemann sum approaches the Riemann integral, the continuous Lemma 1 follows from the discrete Cauchy–Davenport theorem (with the unit circle representation) by taking limit with $p \to \infty$.

We actually need the following corollary of Lemma 1.

Lemma 2 *Let* U *be a periodic set on the real line with period one, and assume that* U *is Jordan measurable. For every integer* $\ell \geq 2$ *write* $\ell U = \{u + v : u \in U, v \in (\ell - 1)U\}$. *Then for every integer* $\ell \geq 2$, *either* ℓU *is the whole line, or* $\mathrm{meas}(\ell U) \geq \ell\mathrm{meas}(U)$.

Proof of Lemma 2 It follows from Lemma 1 by induction on ℓ. □

Let $\mathcal{B} = \{b_1, b_2, \ldots, b_r\}$ be an arbitrary finite multiset of integers (i.e., some integers may have multiplicity greater than one). We introduce the unconventional concept of *strength* of the multiset $\mathcal{B}$. First, for every positive nonzero integer $s \geq 1$ write

$$\mathrm{mult}(\mathcal{B}; s) = \sum_{\substack{1 \leq h \leq r: \\ |b_h| = s}} 1,$$

that is, $\mathrm{mult}(\mathcal{B}; s)$ represents the multiplicity of $\pm s$ in $\mathcal{B}$. Now let

$$\mathrm{Str}(\mathcal{B}) = \mathrm{Strength}(\mathcal{B}) = \sum_{s=1}^{\infty} (\mathrm{mult}(\mathcal{B}; s))^2 . \tag{33}$$

We emphasize that $\mathcal{B}$ may contain several zeros, but the multiplicity of zero does not contribute to (33).

We apply Lemma 2 in the proof of the following technical result about certain trigonometric sums.

Lemma 3 *Let $\mathcal{B} = \{b_1, \dots, b_r\}$ be an arbitrary multiset of nonzero integers, and let y be a positive real number with $r/2 \geq y > 0$. Then*

$$\text{meas}\left\{ t \in [0,1) : \sum_{j=1}^{r} (1 - cos(2\pi b_j t)) \leq y \right\} \leq \min_{\substack{\mathcal{C} \subseteq \mathcal{B}: \\ |\mathcal{C}| \geq 2y}} \frac{2\text{Str}(\mathcal{C})}{\left\lfloor \sqrt{\frac{|\mathcal{C}|}{2y}} \right\rfloor |\mathcal{C}|^2}.$$

Remark We basically repeat some of Halász's arguments in [7].

Proof of Lemma 3 Note that for every integer $k \geq 2$,

$$\left| \sin\left(\sum_{h=1}^{k} x_h \right) \right| \leq \sum_{h=1}^{k} |\sin x_h|, \tag{34}$$

which follows from the elementary fact

$$|\sin(x+y)| = |\sin x \cos y + \sin y \cos x| \leq |\sin x| + |\sin y|$$

by induction.

Combining (34) with the inequality between the arithmetic and quadratic means, we obtain the inequality

$$1 - \cos\left(2 \sum_{h=1}^{k} \alpha_h \right) = 2 \left(\sin\left(\sum_{j=1}^{k} \alpha_h \right) \right)^2 \leq 2 \left(\sum_{h=1}^{k} |\sin \alpha_h| \right)^2 \leq$$

$$\leq 2k \sum_{h=1}^{k} (\sin \alpha_h)^2 = k \sum_{h=1}^{k} (1 - \cos(2\alpha_h)). \tag{35}$$

For an arbitrary nonempty subset $\mathcal{C} \subseteq \mathcal{B}$ write

$$T(\mathcal{C}; y) = \left\{ t \in \mathbb{R} : \sum_{b_j \in \mathcal{C}} (1 - cos(2\pi b_j t)) \leq y \right\}, \tag{36}$$

which clearly implies

$$T(\mathcal{B}; y) \subseteq T(\mathcal{C}; y). \tag{37}$$

If $t_h \in T(\mathcal{C}; y)$, $1 \leq h \leq k$, then for every fixed $b_j \in \mathcal{C}$ we use (35) with $\alpha_h = \pi b_j t_h$, and obtain

$$\sum_{b_j \in \mathcal{C}} \left(1 - \cos\left(2\pi b_j \sum_{h=1}^{k} t_h \right) \right) \leq k \sum_{h=1}^{k} \sum_{b_j \in \mathcal{C}} \left(1 - \cos(2\pi b_j t_h) \right) \leq k^2 y.$$

Combining this with (36), we obtain the following information about the set-sum $kT(\mathcal{C};y)$:

$$kT(\mathcal{C};y) = T(\mathcal{C};y) + \cdots + T(\mathcal{C};y) \subset T(\mathcal{C};k^2y). \tag{38}$$

Applying Lemma 2 in (38), we have

$$\text{meas}\left(T(\mathcal{C};k^2y)\right) \geq \text{meas}\left(kT(\mathcal{C};y)\right) \geq \min\left\{1, k\cdot \text{meas}(T(\mathcal{C};y))\right\}, \tag{39}$$

which holds for every integer $k \geq 2$.

Let

$$c = |\mathcal{C}| = \sum_{b_j\in\mathcal{C}} 1 \geq 1$$

(of course counted with multiplicity). Choosing $y = c/2$ in (36), we have

$$T(\mathcal{C};c/2) = \left\{t \in \mathbb{R} : \sum_{b_j\in\mathcal{C}} (1 - cos(2\pi b_j t)) \leq c/2\right\} =$$

$$= \left\{t \in \mathbb{R} : \sum_{b_j\in\mathcal{C}} cos(2\pi b_j t) \geq c/2\right\}. \tag{40}$$

Clearly $\text{meas}(T(\mathcal{C};c/2)) < 1$. Combining (33), (40) and Parseval's formula, we obtain the following non-trivial upper bound

$$\text{meas}(T(\mathcal{C};c/2))\left(\frac{c}{2}\right)^2 \leq \int_0^1 \left(\sum_{b_j\in\mathcal{C}} cos(2\pi b_j t)\right)^2 dt = \frac{1}{2}\text{Str}(\mathcal{C}),$$

which implies

$$\text{meas}(T(\mathcal{C};c/2)) \leq \frac{2}{c^2}\text{Str}(\mathcal{C}). \tag{41}$$

Let k be the largest integer satisfying $k^2y \leq c/2$, that is, let

$$k = \left\lfloor \sqrt{\frac{c}{2y}} \right\rfloor \text{ (lower integral part).} \tag{42}$$

By (39), (41) and (42),

$$\frac{2}{c^2}\mathrm{Str}(\mathcal{C}) \geq \mathrm{meas}(T(\mathcal{C}; c/2)) \geq \mathrm{meas}\left(T(\mathcal{C}; k^2y)\right) \geq \min\left\{1, k \cdot \mathrm{meas}(T(\mathcal{C}; y))\right\}. \tag{43}$$

Since trivially $1 > \mathrm{meas}(T(\mathcal{C}; c/2))$, (43) implies

$$\frac{2}{c^2}\mathrm{Str}(\mathcal{C}) \geq k \cdot \mathrm{meas}(T(\mathcal{C}; y)),$$

and so

$$\mathrm{meas}(T(\mathcal{C}; y)) \leq \frac{1}{k}\frac{2}{c^2}\mathrm{Str}(\mathcal{C}) = \frac{2\mathrm{Str}(\mathcal{C})}{\left\lfloor \sqrt{\frac{|\mathcal{C}|}{2y}} \right\rfloor |\mathcal{C}|^2},$$

assuming $|\mathcal{C}| \geq 2y$. Combining this with (37), Lemma 3 follows. □

Next we use Lemma 3 to prove an upper bound on a "concentration problem" in combinatorial number theory. Again we closely follow the paper of Halász [7].

Let $\mathcal{A} = \{a_1, a_2, \ldots, a_k\}$ be a multiset of positive nonzero integers. For an arbitrary integer $d \in \mathbb{Z}$ let $\mathrm{Z}(\mathcal{A}; d)$ denote the number of solutions $(z_1, z_2, \ldots, z_k) \in \{-1, 1\}^k$ of the equation

$$\sum_{h=1}^{k} a_h z_h = d. \tag{44}$$

Clearly (where of course $\mathrm{i} = \sqrt{-1}$)

$$\mathrm{Z}(\mathcal{A}; d) = \int_0^1 e^{-2\pi \mathrm{i} d} \prod_{h=1}^{k} \left(e^{2\pi \mathrm{i} a_h t} + e^{-2\pi \mathrm{i} a_h t}\right) dt =$$

$$= 2^k \int_0^1 e^{-2\pi \mathrm{i} d} \prod_{h=1}^{k} \cos(2\pi a_h t)\ dt \leq 2^k \int_0^1 \prod_{h=1}^{k} |\cos(2\pi a_h t)|\ dt,$$

which imples

$$\mathrm{Z}(\mathcal{A}) = \max_{d \in \mathbb{Z}} \mathrm{Z}(\mathcal{A}; d) \leq 2^k \int_0^1 \prod_{h=1}^{k} |\cos(2\pi a_h t)|\ dt. \tag{45}$$

Applying the inequality $y \le e^{-(1-y)}$ for $y \ge 0$ in (45), and also using the elementary fact $2\cos^2 x - 2 = \cos(2x) - 1$, we have

$$|\cos(2\pi a_h t)| = \left(\cos^2(2\pi a_h t)\right)^{1/2} \le$$

$$\le \exp\left\{-\frac{1}{2}\left(1-\cos^2(2\pi a_h t)\right)\right\} = \exp\left\{-\frac{1}{4}\left(1-\cos(4\pi a_h t)\right)\right\},$$

and so

$$\prod_{h=1}^{k} |\cos(2\pi a_h t)| \le \exp\left\{-\frac{1}{4}\sum_{h=1}^{k}(1-\cos(4\pi a_h t))\right\}. \tag{46}$$

Write

$$f(t) = f(\mathcal{A};t) = \sum_{h=1}^{k}(1-\cos(4\pi a_h t)). \tag{47}$$

Combining (45), (46) and (47), we have

$$\frac{\mathrm{Z}(\mathcal{A})}{2^k} \le \int_0^1 \prod_{h=1}^{k}|\cos(2\pi a_h t)|\,dt \le \int_0^1 \exp\left\{-\frac{1}{4}f(\mathcal{A};t)\right\}\,dt. \tag{48}$$

Using $f(t) = f(\mathcal{A};t)$ we obtain the trivial upper bound

$$\int_0^1 \exp\left\{-\frac{1}{4}f(t)\right\}\,dt \le \sum_{j=0}^{\infty} \mathrm{meas}(\{t \in [0,1) : j \le f(t) < j+1\})e^{-j/4}. \tag{49}$$

To estimate the right-hand side of (49), we apply Lemma 3 for the multiset $2\mathcal{A} = \{2a_1, 2a_2, \ldots, 2a_k\}$, and obtain the concentration upper bound (see (48) and (49))

$$\frac{\mathrm{Z}(\mathcal{A})}{2^k} \le \sum_{j=0}^{\infty} \mathrm{meas}(\{t \in [0,1) : j \le f(t) < j+1\})e^{-(j-1)/4} \le$$

$$\le \left(\sum_{1\le\ell\le k/2} e^{-(\ell-1)/4} \min_{\substack{\mathcal{C}\subseteq 2\mathcal{A}:\\ |\mathcal{C}|\ge 2\ell}} \frac{2\mathrm{Str}(\mathcal{C})}{\left\lfloor\sqrt{\frac{|\mathcal{C}|}{2\ell}}\right\rfloor |\mathcal{C}|^2}\right) + e^{-(k-2)/8} =$$

$$= \left(\sum_{1\le\ell\le k/2} e^{-(\ell-1)/4} \min_{\substack{\mathcal{B}\subseteq \mathcal{A}:\\ |\mathcal{B}|\ge 2\ell}} \frac{2\mathrm{Str}(\mathcal{B})}{\left\lfloor\sqrt{\frac{|\mathcal{B}|}{2\ell}}\right\rfloor |\mathcal{B}|^2}\right) + e^{-(k-2)/8}, \tag{50}$$

with $\mathcal{B} = \mathcal{C}/2$ (since clearly $\mathrm{Str}(\mathcal{B}) = \mathrm{Str}(\mathcal{C})$).

Using the trivial fact

$$\lfloor x \rfloor \geq x/2 \text{ for every real } x \geq 1$$

in (50), we have

$$\frac{\mathrm{Z}(\mathcal{A})}{2^k} \leq \left(\sum_{1 \leq \ell \leq k/2} \min_{\substack{\mathcal{B} \subseteq \mathcal{A}: \\ |\mathcal{B}| \geq 2\ell}} \frac{8\sqrt{\ell} e^{-(\ell-1)/4} \mathrm{Str}(\mathcal{B})}{|\mathcal{B}|^{5/2}} \right) + e^{-(k-2)/8}.$$

Moreover, we have the alternative upper bound

$$\frac{\mathrm{Z}(\mathcal{A})}{2^k} \leq \min \left\{ \frac{1}{\sqrt{k}}, \frac{1}{2} \right\}.$$

Here the first half

$$\frac{\mathrm{Z}(\mathcal{A})}{2^k} \leq \frac{1}{\sqrt{k}}$$

immediately follows from a classical result of Erdős [6]

$$\mathrm{Z}(\mathcal{A}) \leq \binom{k}{\lfloor k/2 \rfloor}.$$

For the sake of completeness we outline the elegant idea: it is an application of Sperner's theorem in combinatorics, and goes as follows. For a fixed integer d, we associate with every integer solution of the equation $\sum_{h=1}^{k} a_h z_h = d$, $\mathbf{z} = (z_1, z_2, \ldots, z_k) \in \{-1, 1\}^k$ the subset

$$S(\mathbf{z}) = \{j \in \{1, 2, 3, \ldots, k\} : z_j = 1\}$$

of $\{1, 2, 3, \ldots, k\}$. The family $\mathcal{F}_d$ of sets $S(\mathbf{z})$ is an *antichain* of $\{1, 2, 3, \ldots, k\}$, i.e., there are no two sets $S(\mathbf{z}_1), S(\mathbf{z}_2) \in \mathcal{F}_d$ such that one is a subset of the other. By a well-known theorem of Sperner every antichain on a k-element underlying set has at most

$$\binom{k}{\lfloor k/2 \rfloor}$$

sets, proving Erdős's upper bound

$$\mathrm{Z}(\mathcal{A}) \leq \binom{k}{\lfloor k/2 \rfloor}.$$

We finish with the simple upper bound

$$\frac{\mathrm{Z}(\mathcal{A})}{2^k} \leq \binom{k}{\lfloor k/2 \rfloor} 2^{-k} \leq \frac{1}{\sqrt{k}},$$

Note that the last inequality is clear for "small" k by brute force checking, and for "large" k it easily follows from applying Stirling's formula

$$n! = (1 + o(1)) \left(\frac{n}{e}\right) \sqrt{2\pi n}$$

for the three factorials in

$$\binom{k}{\lfloor k/2 \rfloor} = \frac{k!}{\lfloor k/2 \rfloor ! \lceil k/2 \rceil !}.$$

This completes the proof of the first half.

Finally, note that the upper bound

$$\frac{\mathrm{Z}(\mathcal{A})}{2^k} \leq \frac{1}{2}$$

is completely trivial. Indeed, if the first $k - 1$ variables z_h, $1 \leq h \leq k - 1$ are already fixed, then from the two values $z_k = \pm 1$ of the last variable at least one does not work (i.e., $\sum_{h=1}^{k} a_h z_h \neq d$), proving the upper bound $\frac{1}{2}$.

Combining the three upper bounds, we obtain the following lemma.

Lemma 4 *Let* $\mathcal{A} = \{a_1, a_2, \ldots, a_k\}$ *be a multiset of positive nonzero integers, let* $\mathrm{Z}(\mathcal{A}; d)$ *denote the number of solutions* $(z_1, z_2, \ldots, z_k) \in \{-1, 1\}^k$ *of the equation*

$$\sum_{h=1}^{k} a_h z_h = d,$$

and write

$$\mathrm{Z}(\mathcal{A}) = \max_{d \in \mathbb{Z}} \mathrm{Z}(\mathcal{A}; d).$$

Then

$$\frac{\mathrm{Z}(\mathcal{A})}{2^k} \le \min\left\{\left(\sum_{1\le\ell\le k/2} \min_{\substack{\mathcal{B}\subseteq\mathcal{A}:\\ |\mathcal{B}|\ge 2\ell}} \frac{12\sqrt{\ell}e^{-\ell/4}\mathrm{Str}(\mathcal{B})}{|\mathcal{B}|^{5/2}}\right) + e^{-(k-2)/8}, \frac{1}{\sqrt{k}}, \frac{1}{2}\right\}$$

where $\mathrm{Str}(\mathcal{A})$ *is defined in* (33). □

After these preparations we are now ready to apply the "probabilistic method". Since we work with discrete probability, the "probabilistic method" is just a "counting method"; in fact, "counting the average value".

Let $\mathbf{x} = (x_1, x_2, \ldots, x_N) \in \mathbb{Z}^N \setminus \mathbf{0}$ be an arbitrary but fixed nontrivial integer vector such that

$$|x_i| \le L = \left\lfloor c_0 \left(\sqrt{N}\right)^{n/(N-n)} \right\rfloor, \ \ 1 \le i \le N$$

(the value of the absolute constant $c_0 > 0$ will be specified later; see (87)). Let $k = k(\mathbf{x}) \ge 1$ denote the number of nonzero coordinates of vector $\mathbf{x}$, and let

$$x_{i_1}, x_{i_2}, \ldots, x_{i_k} \text{ be the nonzero coordinates of vector } \mathbf{x} \tag{51}$$

with $1 \le i_1 < i_2 < \ldots < i_k \le N$.

Let $Q(\mathbf{x})$ denote the cardinality of the set of solutions $(y_1, y_2, \ldots, y_k) \in \{1, -1\}^k$ of the equation

$$\sum_{j=1}^{k} x_{i_j} y_j = 0. \tag{52}$$

We emphasize that here $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ are the fixed coefficients, and $y_1, y_2, \ldots, y_k \in \{1, -1\}$ are the variables.

Write (see (51))

$$\mathcal{X} = \mathcal{X}(\mathbf{x}) = \{x_{i_1}, x_{i_2}, \ldots, x_{i_k}\}. \tag{53}$$

We can give an upper bound for $Q(\mathbf{x})$ by applying Lemma 4 with $a_h = |x_{i_h}|$ and $z_h = y_h$ $(1 \le h \le k)$:

$$\frac{Q(\mathbf{x})}{2^k} \le \frac{\mathrm{Z}(\mathcal{X})}{2^k} \le \min\left\{\left(\sum_{1\le\ell\le k/2} \min_{\substack{\mathcal{Z}\subseteq\mathcal{X}:\\ |\mathcal{Z}|\ge 2\ell}} \frac{12\sqrt{\ell}e^{-\ell/4}\mathrm{Str}(\mathcal{Z})}{|\mathcal{Z}|^{5/2}}\right) + e^{-(k-2)/8}, \frac{1}{\sqrt{k}}, \frac{1}{2}\right\}, \tag{54}$$

where $k = k(\mathbf{x}) \ge 1$ denotes the number of nonzero coordinates of vector $\mathbf{x}$.

Next we introduce a discrete probability space. Let $\Omega = \Omega_{n,N}$ denote the sets of all matrices $\mathbf{Y} = (y_{i,j})$, $1 \le i \le n$, $1 \le j \le N$ with n rows and N columns such that every entry $y_{i,j}$ is 1 or -1. So $|\Omega| = 2^{nN}$, and assume that the matrices $\mathbf{Y}$ are equally likely. In other words, the entries $y_{i,j}$, $1 \le i \le n$, $1 \le j \le N$ of the matrix $\mathbf{Y}$ represent nN independent random variables, each having values ± 1 with probability $1/2$.

Let $\Phi = \Phi_{n,N}$ denote the set of all nontrivial integer vectors $\mathbf{x} = (x_1, x_2, \dots, x_N) \in \mathbb{Z}^N \setminus \mathbf{0}$ such that

$$|x_i| \le L = \left\lfloor c_0 \left(\sqrt{N}\right)^{n/(N-n)} \right\rfloor, \;\; 1 \le i \le N. \tag{55}$$

For every $\mathbf{x} \in \Phi$ define the set

$$\text{Zero}(\mathbf{x}) = \{\mathbf{Y} \in \Omega : \; \mathbf{Y}\mathbf{x} = \mathbf{0}\},$$

where of course $\mathbf{Y}\mathbf{x} = \mathbf{0}$ has the long form

$$\sum_{j=1}^{N} y_{i,j} x_j = 0, \;\; 1 \le i \le n.$$

By using (51), (52), (53), and (54), we have the upper bound for the corresponding (discrete) probability

$$\Pr[\text{Zero}(\mathbf{x})] = |\text{Zero}(\mathbf{x})| 2^{-nN} \le$$

$$\le \left(\min \left\{ \left(\sum_{1 \le \ell \le k/2} \min_{\substack{\mathcal{Z} \subseteq \mathcal{X}: \\ |\mathcal{Z}| \ge 2\ell}} \frac{12\sqrt{\ell} e^{-\ell/4} \text{Str}(\mathcal{Z})}{|\mathcal{Z}|^{5/2}} \right) + e^{-(k-2)/8}, \frac{1}{\sqrt{k}}, \frac{1}{2} \right\} \right)^n, \tag{56}$$

where the n-th power comes from the independence of the n rows of $\mathbf{Y}$. (More precisely, different rows of $\mathbf{Y}$ contain disjoint sets of independent random variables $y_{i,j}$, so the product rule applies.)

By using upper bound (56), we are able to prove the following lemma.

Lemma 5 *There is a positive absolute constant $c_0 > 0$ (see the definition of L in (55)) such that for all $N \ge 3n/2$,*

$$\sum_{\mathbf{x} \in \Phi_{n,N}} |\text{Zero}(\mathbf{x})| 2^{-nN} = O(2^{-n/2}).$$

First we derive Theorem 1 from Lemma 5. Simple double-counting gives the equality

$$\sum_{\mathbf{x}\in\Phi_{n,N}} |\mathrm{Zero}(\mathbf{x})| = \sum_{\mathbf{Y}\in\Omega_{n,N}} \sum_{\substack{\mathbf{x}\in\Phi_{n,N}:\\ \mathbf{Y}\mathbf{x}=\mathbf{0}}} 1.$$

Combining (2) and the fact $|\Omega| = 2^{nN}$ with Lemma 5, we have

$$\frac{1}{|\Omega|} \sum_{\mathbf{Y}\in\Omega_{n,N}} \sum_{\substack{\mathbf{x}\in\Phi_{n,N}:\\ \mathbf{Y}\mathbf{x}=\mathbf{0}}} 1 = O(2^{-n/2}). \tag{57}$$

(57) means that the average number of "small" (=max norm is $\leq L$) nontrivial integral solutions $\mathbf{x} \in \Phi_{n,N}$ is $O(2^{-n/2})$. This implies that for the overwhelming majority of the 2^{nN} homogeneous linear systems $\mathbf{Y}\mathbf{x} = \mathbf{0}$, $\mathbf{Y} \in \Omega_{n,N}$ there is *no* "small" nontrivial integral solution (if n is large).

In fact, (57) implies that the violators of Theorem 1 represent exponentially small $O(2^{-n/2})$ part of the total 2^{nN}.

This completes the deduction of Theorem 1 from Lemma 5.

3 Proof of Lemma 5

We recall the notation introduced in (33): for every $1 \leq s \leq L$ we use $m_s = \mathrm{mult}(\mathcal{X}; s)$ to denote the multiplicity of $\pm s$ in $\mathcal{X} = \mathcal{X}(\mathbf{x})$, where $\mathcal{X}(\mathbf{x})$ is defined in (53).

Let $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$ denote the set of natural numbers *including zero*. Given a multiplicity vector $\mathbf{m} = (m_1, m_2, m_3, \ldots, m_L) \in \mathbb{N}^L$, we can define its *Strength* in the analogous way (see (33))

$$\mathrm{Str}(\mathbf{m}) = \sum_{s=1}^{L} m_s^2.$$

Given two vectors

$$\mathbf{r} = (r_1, r_2, r_3, \ldots) \in \mathbb{N}^L \text{ and } \mathbf{m} = (m_1, m_2, m_3, \ldots) \in \mathbb{N}^L,$$

we write $\mathbf{r} \leq \mathbf{m}$ if and only if $r_s \leq m_s$ for all $1 \leq s \leq L$.

For a fixed index-sequence $1 \leq i_1 < i_2 < \ldots < i_k \leq N$ in (51), let $\Phi(i_1, i_2, \ldots, i_k)$ denote the set of vectors $\mathbf{x} = (x_1, \ldots, x_N) \in \Phi$ such that

$$x_{i_1}, x_{i_2}, \ldots, x_{i_k} \text{ are the nonzero coordinates of } \mathbf{x}.$$

Then by (56)

$$\sum_{\mathbf{x}\in\Phi(i_1,i_2,\ldots,i_k)} |\mathrm{Zero}(\mathbf{x})|2^{-nN} \le$$

$$\le 2^k \sum_{\substack{\mathbf{m}=(m_1,m_2,m_3,\ldots)\in\mathbb{N}^L:\\ m_1+m_2+m_3+\ldots=k}} \binom{k}{m_1}\binom{k-m_1}{m_2}\binom{k-m_1-m_2}{m_3}\cdots$$

$$\cdot\left(\min\left\{\left(\sum_{1\le\ell\le k/2}\ \min_{\substack{\mathbf{r}\le\mathbf{m}:\\ r_1+r_2+r_3+\ldots\ge 2\ell}} \frac{12\sqrt{\ell}e^{-\ell/4}\mathrm{Str}(\mathbf{r})}{(r_1+r_2+r_3+\ldots)^{5/2}}\right)+e^{-(k-2)/8},\frac{1}{\sqrt{k}},\frac{1}{2}\right\}\right)^n, \tag{58}$$

where the factor 2^k at the beginning of the middle line comes from the k choices of the signs ± in $\pm s$.

Therefore, by (58) we have

$$\sum_{\mathbf{x}\in\Phi} |\mathrm{Zero}(\mathbf{x})|2^{-nN} =$$

$$= \sum_{1\le k\le N}\ \sum_{\substack{(i_1,i_2,\ldots,i_k):\\ 1\le i_1<i_2<\ldots<i_k\le N}}\ \sum_{\mathbf{x}\in\Phi(i_1,i_2,\ldots,i_k)} |\mathrm{Zero}(\mathbf{x})|2^{-nN} \le$$

$$\le \sum_{k=1}^{N}\binom{N}{k}2^k \sum_{\substack{\mathbf{m}=(m_1,m_2,m_3,\ldots)\in\mathbb{N}^L:\\ m_1+m_2+m_3+\ldots=k}} \binom{k}{m_1}\binom{k-m_1}{m_2}\binom{k-m_1-m_2}{m_3}\cdots$$

$$\cdot\,(\Lambda(\mathbf{m}))^n\,, \tag{59}$$

where we used the new notation

$$\Lambda(\mathbf{m}) = \min\left\{\left(\sum_{1\le\ell\le k/2}\ \min_{\substack{\mathbf{r}\le\mathbf{m}:\\ \sum_i r_i\ge 2\ell}} \frac{12\sqrt{\ell}e^{-\ell/4}\left(\sum_i r_i^2\right)}{\left(\sum_i r_i\right)^{5/2}}\right)+e^{-(k-2)/8},\frac{1}{\sqrt{k}},\frac{1}{2}\right\}. \tag{60}$$

We recall (55):

$$L = \left\lfloor c_0\left(\sqrt{N}\right)^{n/(N-n)}\right\rfloor,\ \text{ and of course } L\ge 1,$$

since otherwise Theorem 1 is trivial. Theorem 1 and Lemma 5 are about the case $N \geq 3n/2$, but for technical reasons we study the slightly larger range

$$N \geq 11n/8.$$

Then we have

$$L \leq N^{4/3}. \tag{61}$$

Also, we have $N \leq n \log N$.

We split the right-hand side of (59) into two parts according as $1 \leq k \leq c_1 N$ and $c_1 N < k \leq N$, where $c_1 > 0$ is a sufficiently small absolute constant to be specified later (note in advance that $c_1 = 1/4$ is a good choice; see (70)). Our goal is to show that the contribution of the first part $1 \leq k \leq c_1 N$ is negligible if n is large.

By using the trivial estimation (see (60))

$$\Lambda(\mathbf{m}) \leq \min\left\{\frac{1}{\sqrt{k}}, \frac{1}{2}\right\},$$

we have

$$\text{first part of the right hand side of (59)} =$$

$$= \sum_{1 \leq k \leq c_1 N} \binom{N}{k} 2^k \sum_{\substack{\mathbf{m}=(m_1,m_2,m_3,\ldots)\in\mathbb{N}^L: \\ m_1+m_2+m_3+\ldots=k}} \binom{k}{m_1}\binom{k-m_1}{m_2}\binom{k-m_1-m_2}{m_3}\cdots$$

$$\cdot (\Lambda(\mathbf{m}))^n \leq$$

$$\leq \sum_{1 \leq k \leq c_1 N} \binom{N}{k} 2^k \sum_{\substack{(m_1,m_2,m_3,\ldots)\in\mathbb{N}^L: \\ m_1+m_2+m_3+\ldots=k}} \binom{k}{m_1}\binom{k-m_1}{m_2}\binom{k-m_1-m_2}{m_3}\cdots$$

$$\cdot \left(\min\left\{\frac{1}{\sqrt{k}}, \frac{1}{2}\right\}\right)^n \leq$$

$$\leq \sum_{1 \leq k \leq \frac{1}{10} n/\log n} \binom{N}{k} 2^k \cdot$$

$$\cdot \left(\sum_{\substack{(m_1,m_2,m_3,\ldots)\in\mathbb{N}^L: \\ m_1+m_2+m_3+\ldots=k}} \binom{k}{m_1}\binom{k-m_1}{m_2}\binom{k-m_1-m_2}{m_3}\cdots \right) \left(\frac{1}{2}\right)^n +$$

$$+\sum_{\frac{1}{10}n/\log n<k\le c_1N}\binom{N}{k}2^k\cdot$$

$$\cdot\left(\sum_{\substack{(m_1,m_2,m_3,\ldots)\in\mathbb{N}^L:\\ m_1+m_2+m_3+\ldots=k}}\binom{k}{m_1}\binom{k-m_1}{m_2}\binom{k-m_1-m_2}{m_3}\cdots\right)\left(\frac{1}{\sqrt{k}}\right)^n\le$$

$$\le\sum_{1\le k\le\frac{1}{10}n/\log n}\binom{N}{k}2^kL^k2^{-n}+\sum_{\frac{1}{10}n/\log n<k\le c_1N}\binom{N}{k}2^kL^kk^{-n/2},\tag{62}$$

where the factor L^k comes from the multinomial theorem.

By using the inequality $k!\ge(k/e)^k$—we call it the "weak form of Stirling's formula"—we have

$$\sum_{1\le k\le\frac{1}{10}n/\log n}\binom{N}{k}2^kL^k2^{-n}\le\sum_{1\le k\le\frac{1}{10}n/\log n}\left(\frac{2eNL}{k}\right)^k2^{-n}\le$$

$$\le\sum_{1\le k\le\frac{1}{10}n/\log n}\left(\frac{2eNN^{4/3}}{k}\right)^k2^{-n}<2^{-n/2},\tag{63}$$

if n is large (here we used $N\le n\log N$ and (61)).

Next we split the last sum in (61) into two parts

$$\sum_{\frac{1}{10}n/\log n<k\le c_1N}\binom{N}{k}2^kL^kk^{-n/2}=$$

$$=\sum_{\frac{1}{10}n/\log n<k\le n/3}\binom{N}{k}2^kL^kk^{-n/2}+\sum_{n/3<k\le c_1N}\binom{N}{k}2^kL^kk^{-n/2},\tag{64}$$

and estimate the first part by using the weak form of Stirling's formula as above:

$$\sum_{\frac{1}{10}n/\log n<k\le n/3}\binom{N}{k}2^kL^kk^{-n/2}\le\sum_{\frac{1}{10}n/\log n<k\le n/3}\left(\frac{2eNL}{k}\right)^k k^{-n/2}\le$$

$$\le\sum_{\frac{1}{10}n/\log n<k\le n/3}\left(N^{(4/3)+o(1)}\right)^k\left(N^{-1+o(1)}\right)^{-n/2}<2^{-n},\tag{65}$$

since $4/9<1/2$ and $N\le n\log N$.

Next we estimate the second part in (65); again by using the weak form of Stirling's formula,

$$\sum_{n/3<k\le c_1N}\binom{N}{k}2^kL^kk^{-n/2}\le\sum_{n/3<k\le c_1N}f(k),\tag{66}$$

where

$$f(x)=\left(\frac{2eNL}{x}\right)^x x^{-n/2}=\exp\left(x\log A-(x+B)\log x\right)\ \text{ with } A=2eNL \text{ and } B=n/2.\tag{67}$$

The derivative of the function

$$g(x)=g_{A,B}(x)=x\log A-(x+B)\log x$$

is very simple:

$$g'(x)=\log A-\log x-1-\frac{B}{x}=\log\frac{A}{ex}-\frac{B}{x}.$$

Since

$$\min_{n/3\le x\le c_1N}\left(\log\frac{A}{ex}-\frac{B}{x}\right)\ge\log\frac{2L}{c_1}-\frac{n/2}{n/3}>0\ \text{ if }\ c_1\le 1/3,\tag{68}$$

we obtain that $f(x)$ is monotone increasing in the interval $n/3\le x\le c_1N$, and using this fact in (66), (67), and (68), we have the upper bound

$$\sum_{n/3<k\le c_1N}\binom{N}{k}2^kL^kk^{-n/2}\le N\cdot\left(\frac{2eNL}{c_1N}\right)^{c_1N}(c_1N)^{-n/2}\le$$

$$\le N\cdot\left(\frac{2e}{c_1}\right)^{c_1N}\left(c_0N^{\frac{n}{2(N-n)}}\right)^{c_1N}(c_1N)^{-n/2}\le$$

$$\le N\cdot\left(\frac{2ec_0}{c_1}\right)^{c_1N}\left(\frac{N^{\frac{c_1N}{N-n}-1}}{c_1}\right)^{n/2}.\tag{69}$$

If

$$2ec_0 \le c_1, \quad N \ge \frac{7}{5}n \text{ and } c_1 = \frac{1}{4}, \tag{70}$$

then

$$\frac{c_1 N}{N-n} \le c_1 7/2 = 7/8, \tag{71}$$

and using (70)–(71) in (69), we have

$$\sum_{n/3<k\le c_1 N} \binom{N}{k} 2^k L^k k^{-n/2} \le N \cdot \left(\frac{N^{\frac{c_1 N}{N-n}-1}}{c_1} \right)^{n/2} \le N \cdot \left(\frac{N^{(7/8)-1}}{c_1} \right)^{n/2} \le 2^{-n}. \tag{72}$$

Thus in the rest it suffices to focus on the range $c_1 N = N/4 < k \le N$, which means that we study the second part at the end of (59):

second part of the right hand side of (59) =

$$= \sum_{c_1 N \le k \le N} \binom{N}{k} 2^k k! \sum_{\substack{\mathbf{m}=(m_1,m_2,m_3,\ldots)\in\mathbb{N}^L: \\ m_1+m_2+m_3+\ldots=k}} \frac{1}{m_1!m_2!m_3!\cdots} \cdot (\Lambda(\mathbf{m}))^n \,. \tag{73}$$

Given a multiplicity vector $\mathbf{m} = (m_1, m_2, m_3, \ldots) \in \mathbb{N}^L$ and an integer $j \ge 0$, consider the following power-of-two decomposition

$$M_j = M_j(\mathbf{m}) = \sum_{\substack{1\le i\le L: \\ 2^j \le m_i < 2^{j+1}}} m_i.$$

So

$$\sum_{j\ge 0} M_j = \sum_{1\le i\le L} m_i = k.$$

Let

$$\max_{j\ge 0} M_j^{3/2} 2^{-j} = M_{j_0}^{3/2} 2^{-j_0},$$

that is, the maximum is attained at $j = j_0$.

Clearly

$$M_{j_0}^{3/2}2^{-j_0} = \sqrt{M_{j_0}}\frac{M_{j_0}}{2^{j_0}} < \sqrt{k}\frac{2^{j_0+1}L}{2^{j_0}} = 2\sqrt{k}L.$$

For notational convenience to be used later, we write

$$M_{j_0}^{3/2}2^{-j_0} = 2^{-p}\sqrt{N}L, \tag{74}$$

where $p = p(\mathbf{m}) > -1$ is an appropriate real number.

Let $\mathbf{r} = \mathbf{r}(\mathbf{m}; j_0) = (r_1, r_2, r_3, \ldots)$ be defined as follows: $r_i = m_i$ if $2^{j_0} \le m_i < 2^{j_0+1}$, and $r_i = 0$ otherwise. By using the definition of $\Lambda(\mathbf{m})$ (see (60)) with this $\mathbf{r} = \mathbf{r}(\mathbf{m}; j_0)$, we have

$$\Lambda(\mathbf{m}) \le \sum_{1\le\ell\le M_{j_0}/2} 12\sqrt{\ell}e^{-\ell/4}\left(\sum_{\substack{1\le i\le L:\\ 2^{j_0}\le m_i<2^{j_0+1}}} m_i^2\right)M_{j_0}^{-5/2}+$$

$$+\sum_{M_{j_0}/2<\ell\le k/2} 12\sqrt{\ell}e^{-\ell/4} + e^{-(k-2)/8} \le$$

$$\le \sum_{1\le\ell\le M_{j_0}/2} 12\sqrt{\ell}e^{-\ell/4}\left(2^{j_0+1}M_{j_0}\right)M_{j_0}^{-5/2} + \sum_{M_{j_0}/2<\ell\le k/2} 12\sqrt{\ell}e^{-\ell/4} + e^{-(k-2)/8} =$$

$$= 2^{j_0}M_{j_0}^{-3/2}\sum_{1\le\ell\le M_{j_0}/2} 24\sqrt{\ell}e^{-\ell/4} + \sum_{M_{j_0}/2<\ell\le k/2} 12\sqrt{\ell}e^{-\ell/4} + e^{-(k-2)/8}. \tag{75}$$

We need the following almost trivial lower bound

$$M_{j_0} \ge k^{1/4}. \tag{76}$$

To prove (76), consider the maximum

$$M_{j_1} = \max_{j\ge0} M_j \ge \text{average} \ge \frac{\sum_{j\ge0} M_j}{1+\log_2 k} = \frac{k}{1+\log_2 k},$$

which implies

$$M_{j_0}^{3/2} \ge M_{j_0}^{3/2}2^{-j_0} = \max_{j\ge0} M_j^{3/2}2^{-j} \ge M_{j_1}^{3/2}2^{-j_1} \ge$$

$$\ge M_{j_1}^{3/2}k^{-1} \ge \left(\frac{k}{1+\log_2 k}\right)^{3/2}k^{-1} > k^{3/8}, \tag{77}$$

and (76) follows.

Applying (76) in (75), we have

$$\Lambda(\mathbf{m}) \leq 10^3 \frac{2^{j_0}}{M_{j_0}^{3/2}} = \frac{10^3 2^p}{\sqrt{N}L} < \frac{2^{p+10}}{\sqrt{N}L}, \tag{78}$$

where in the last step we used (74).

For an arbitrary integer $j \geq 0$ write

$$q(j) = \log_2 N - \log_2 L - j, \ \text{ or equivalently, } \ 2^j = \frac{N}{2^{q(j)}L}.$$

Using this, (74), (78) and Stirling's formula in (73), we have

$$\text{second part of the right hand side of (59)} =$$

$$= \sum_{c_1 N \leq k \leq N} (2N)^k \sum_{\substack{\mathbf{m}=(m_1,m_2,m_3,\ldots)\in\mathbb{N}^L:\\ m_1+m_2+m_3+\ldots=k}} \prod_{j\geq 0} \prod_{\substack{1\leq i\leq L:\\ 2^j\leq m_i<2^{j+1}}} (m_i!)^{-1} \cdot (\Lambda(\mathbf{m}))^n \leq$$

$$\leq \sum_{c_1 N \leq k \leq N} (2N)^k \sum_{\substack{\mathbf{m}=(m_1,m_2,m_3,\ldots)\in\mathbb{N}^L:\\ m_1+m_2+m_3+\ldots=k}} \prod_{j\geq 0} \prod_{\substack{1\leq i\leq L:\\ 2^j\leq m_i<2^{j+1}}} \left(\frac{m_i}{e}\right)^{-m_i} \cdot (\Lambda(\mathbf{m}))^n \leq$$

$$\leq \sum_{c_1 N \leq k \leq N} (2N)^k \sum_{\substack{\mathbf{m}=(m_1,m_2,m_3,\ldots)\in\mathbb{N}^L:\\ m_1+m_2+m_3+\ldots=k}} \prod_{j\geq 0} \left(\frac{2^j}{e}\right)^{-M_j} \cdot (\Lambda(\mathbf{m}))^n =$$

$$= \sum_{c_1 N \leq k \leq N} \sum_{\substack{\mathbf{m}=(m_1,m_2,m_3,\ldots)\in\mathbb{N}^L:\\ m_1+m_2+m_3+\ldots=k}} \prod_{j\geq 0} \left(\frac{2eN}{2^j}\right)^{M_j} \cdot (\Lambda(\mathbf{m}))^n \leq$$

$$\leq \sum_{c_1 N \leq k \leq N} \sum_{\substack{\mathbf{m}=(m_1,m_2,m_3,\ldots)\in\mathbb{N}^L:\\ m_1+m_2+m_3+\ldots=k}} \prod_{j\geq 0} \left(\frac{2eN}{2^j}\right)^{M_j} \cdot \left(\frac{2^{p+10}}{\sqrt{N}L}\right)^n =$$

$$= \sum_{c_1 N \leq k \leq N} \sum_{\substack{\mathbf{m}=(m_1,m_2,m_3,\ldots)\in\mathbb{N}^L:\\ m_1+m_2+m_3+\ldots=k}} \prod_{j\geq 0} \left(2e2^{q(j)}L\right)^{M_j} \cdot \left(\frac{2^{p+10}}{\sqrt{N}L}\right)^n =$$

$$= \sum_{c_1 N \le k \le N} \sum_{\substack{\mathbf{m}=(m_1,m_2,m_3,\ldots)\in\mathbb{N}^L:\\ m_1+m_2+m_3+\ldots=k}} \prod_{j\ge 0} \left(2e2^{q(j)} L \left(\frac{2^{p+10}}{\sqrt{N}L} \right)^{n/k} \right)^{M_j} =$$

$$= \sum_{c_1 N \le k \le N} \sum_{\substack{\mathbf{m}=(m_1,m_2,m_3,\ldots)\in\mathbb{N}^L:\\ m_1+m_2+m_3+\ldots=k}} \prod_{j\ge 0} \left(2e2^{q(j)} L^{1-(n/k)} \left(\frac{2^{p+10}}{\sqrt{N}} \right)^{n/k} \right)^{M_j}. \tag{79}$$

We claim that the function

$$h(x) = L^{-(n/x)} \left(\frac{2^{p+10}}{\sqrt{N}} \right)^{n/x} = \left(\frac{4^{p+10}}{NL^2} \right)^{(n/2)/x} \tag{80}$$

is monotone increasing in the interval $c_1 N = N/4 \le x \le N$. Indeed, it clearly suffices to verify the inequality

$$NL^2 \ge 4^{p+10}. \tag{81}$$

We recall (74):

$$2^{-p}\sqrt{N}L = M_{j_0}^{3/2} 2^{-j_0} > k^{3/8}, \tag{82}$$

where in the last step we used (77). By (82),

$$4^{-p-10} NL^2 > k^{3/4} 4^{-10} > 1$$

(if n is large), which gives (81).

Applying the monotonicity of the function $h(x)$ defined in (80), we have the upper bound

$$\max_{c_1 N \le k \le N} L^{1-(n/k)} \left(\frac{2^{p+10}}{\sqrt{N}} \right)^{n/k} = L^{1-(n/N)} \left(\frac{2^{p+10}}{\sqrt{N}} \right)^{n/N} \le$$

$$\le c_0^{(N-n)/N} \cdot 2^{(p+10)n/N}, \tag{83}$$

where in the last step we used the definition of L (see (55)), which implies

$$L^{1-(n/N)} \le \left(c_0 N^{\frac{n}{2(N-n)}} \right)^{1-(n/N)} = c_0^{(N-n)/N} (\sqrt{N})^{n/N}.$$

Using (83) in (79), we obtain

$$\text{second part of the right hand side of (59)} \le$$

$$\le \sum_{c_1 N \le k \le N} \sum_{\substack{\mathbf{m}=(m_1,m_2,m_3,\dots)\in\mathbb{N}^L:\\ m_1+m_2+m_3+\dots=k}} \prod_{j\ge 0} \left(c_0^{(N-n)/N} \cdot 2e2^{q(j)+((p+10)n/N)} \right)^{M_j}, \tag{84}$$

where

$$q(j) = \log_2 N - \log_2 L - j. \tag{85}$$

If $\mathbf{m}$ is fixed then of course $p = p(\mathbf{m})$ is also fixed. As $j \ge 0$ runs through the integers, $q(j) = \log_2 N - \log_2 L - j$ (see (85)) runs through an arithmetic progression with gap one.

Furthermore, by using the maximum property of the index j_0, we have (see also (74))

$$M_j^{3/2} 2^{-j} \le M_{j_0}^{3/2} 2^{-j_0} = \sqrt{N} L 2^{-p},$$

and so

$$M_j^{3/2} \le 2^j \sqrt{N} L 2^{-p} = \frac{N}{2^{q(j)} L} \sqrt{N} L 2^{-p} =$$
$$= N^{3/2} 2^{-p-q(j)} \le (k/c_1)^{3/2} 2^{-p-q(j)},$$

which implies

$$M_j \le k 2^{-2(p+q(j))/3} c_1^{-1}.$$

Combining this with the trivial upper bound $M_j \le k$, we have

$$M_j \le \min\left\{ k 2^{-2(p+q(j))/3} c_1^{-1}, k \right\} = k \min\left\{ 2^{-2(p+q(j))/3}/c_1, 1 \right\}. \tag{86}$$

We are ready now to define c_0 (expressed in terms of $c_1 = 1/4$): let

$$c_0 = \left(\frac{c_1^6}{32e} \right)^4 2^{-30} = 2^{-98} e^{-4}. \tag{87}$$

This choice of c_0 clearly satisfies the requirement $2ec_0 \le c_1 = 1/4$ in (70).

Note that in the range $N \ge 4n/3$ definition (87) implies

$$c_0 \le \left(\frac{c_1^6}{32e} \right)^{\frac{N}{N-n}} 2^{\frac{-10n}{N-n}},$$

which is equivalent to

$$c_0^{(N-n)/N} 2e2^{1+(10n/N)} c_1^{-6} \le \frac{1}{8}. \tag{88}$$

We distinguish two cases:
Case 1: Index $j \ge 0$ has the property that

$$2^{q(j)+((p+10)n/N)} \ge 2^{1+(10n)/N} c_1^{-6},$$

and
Case 2: Index $j \ge 0$ has the property that

$$2^{q(j)+((p+10)n/N)} < 2^{(1+(10n/N)} c_1^{-6}.$$

Let $j_2 \ge 0$ be the smallest index satisfying Case 2 (i.e., $q(j_2) = \log_2 N - \log_2 L - j_2$ is the largest member of the arithmetic progression of gap one in Case 2). Then by (88),

$$c_0^{(N-n)/N} \cdot 2e2^{q(j_2)+((p+10)n/N)} < c_0^{(N-n)/N} 2e2^{1+(10n/N)} c_1^{-6} \le \frac{1}{8}. \tag{89}$$

Combining (84), (86) and Cases 1–2, we have

$$\text{second part of the right hand side of (59)} \le$$

$$\le \sum_{c_1 N \le k \le N} \sum_{\substack{\mathbf{m}=(m_1,m_2,m_3,\ldots)\in\mathbb{N}^L:\\ m_1+m_2+m_3+\ldots=k}} \prod_{j\ge 0:\ \text{Case 1}} \left(c_0^{(N-n)/N} \cdot 2e2^{q(j)+((p+10)n/N)}\right)^{M_j} \cdot$$

$$\cdot \prod_{j\ge 0:\ \text{Case 2}} \left(c_0^{(N-n)/N} \cdot 2e2^{q(j)+((p+10)n/N)}\right)^{M_j} \le \sum_{c_1 N \le k \le N} \sum_{\substack{\mathbf{m}=(m_1,m_2,m_3,\ldots)\in\mathbb{N}^L:\\ m_1+m_2+m_3+\ldots=k}} \cdot$$

$$\cdot \prod_{j\ge 0:\ \text{Case 1}} \left(\max\left\{c_0^{(N-n)/N} \cdot 2e2^{q(j)+((p+10)n/N)}, 1\right\}\right)^{k\min\{1,2^{-2(p+q(j))/3}/c_1\}} \cdot$$

$$\cdot \prod_{j\ge 0:\ \text{Case 2}} \left(c_0^{(N-n)/N} \cdot 2e2^{q(j)+((p+10)n/N)}\right)^{M_j}. \tag{90}$$

Combining (89) and the definition of Case 2, we have

$$\prod_{j\geq 0:\text{ Case }2}\left(c_0^{(N-n)/N}\cdot 2e2^{q(j)+((p+10)n/N)}\right)^{M_j}\leq$$

$$\leq\prod_{j\geq 0:\text{ Case }2}8^{-M_j}=\prod_{j\geq j_2}8^{-M_j}=8^{-k}\prod_{j\geq 0:\text{ Case }1}8^{M_j}=$$

$$=8^{-k}\prod_{0\leq j<j_2}8^{M_j},\tag{91}$$

where at the end we used the fact

$$\sum_{j\geq 0}M_j=k.$$

Since $p=p(\mathbf{m})>-1$ (see (74)), in Case 1 we have

$$2^{q(j)+(p+1)}2^{10n/N}\geq 2^{q(j)+(pn/N)}2^{10n/N}=2^{q(j)+((p+10)n/N)}\geq 2^{1+(10n/N)}c_1^{-6},$$

which implies

$$2^{q(j)+p}\geq c_1^{-6},$$

and so

$$2^{-2(p+q(j))/3}/c_1\leq\left(c_1^6\right)^{2/3}/c_1=c_1^3.\tag{92}$$

Combining (89), (92), and the definition of Case 1, we have

$$\prod_{j\geq 0:\text{ Case }1}\left(\max\left\{c_0^{(N-n)/N}\cdot 2e2^{q(j)+((p+10)n/N)},1\right\}\right)^{k\min\{1,2^{-2(p+q(j))/3}/c_1\}}\leq$$

$$\leq\prod_{0\leq j<j_2}\left(2^{|j-j_2|-1}\right)^{k\min\{1,2^{-2(p+q(j))/3}/c_1\}}\leq\prod_{0\leq j<j_2}\left(2^{|j-j_2|-1}\right)^{c_1^3k2^{2(1-|j-j_2|)/3}}\leq$$

$$\leq\prod_{r\geq 0}(2^r)^{c_1^3k2^{-2r/3}}=\exp\left\{\log 2\cdot c_1^3k\sum_{r\geq 0}r2^{-2r/3}\right\},\tag{93}$$

where $r \geq 0$ runs over the non-negative integers. Using the well known fact

$$\sum_{r=0}^{\infty} rx^r = \frac{x}{(1-x)^2} \;\text{ for }\; |x| < 1$$

in (93), we obtain

$$\prod_{j\geq 0:\ \text{Case } 1} \left(\max\left\{c_0^{(N-n)/N} \cdot 2e2^{q(j)+((p+10)n/N)}, 1\right\}\right)^{k\min\{1,2^{-2(p+q(j))/3}/c_1\}} \leq$$

$$\leq \prod_{r\geq 0} (2^r)^{c_1^3 k 2^{-2r/3}} = \exp\left\{\log 2 \cdot c_1^3 k 2^{-2/3}\left(1-2^{-2/3}\right)^{-2}\right\} \leq 2^{10c_1^3 k}. \tag{94}$$

Repeating the same argument, we have

$$\prod_{j\geq 0:\ \text{Case } 1} 8^{M_j} \leq 8^{10c_1^3 k}. \tag{95}$$

By (90), (91), (94) and (95),

$$\text{second part of the right hand side of (59)} \leq \sum_{c_1 N\leq k\leq N} \sum_{\substack{\mathbf{m}=(m_1,m_2,m_3,\ldots)\in\mathbb{N}^L:\\ m_1+m_2+m_3+\ldots=k}} \cdot$$

$$\cdot \prod_{j\geq 0:\ \text{Case } 1} \left(\max\left\{c_0^{(N-n)/N} \cdot 2e2^{q(j)+((p+10)n/N)}, 1\right\}\right)^{k\min\{1,2^{-2(p+q(j))/3}/c_1\}} \cdot$$

$$\cdot \prod_{j\geq 0:\ \text{Case } 2} \left(c_0^{(N-n)/N} \cdot 2e2^{q(j)+((p+10)n/N)}\right)^{M_j} \leq$$

$$\leq \sum_{c_1 N\leq k\leq N} \sum_{\substack{\mathbf{m}=(m_1,m_2,m_3,\ldots)\in\mathbb{N}^L:\\ m_1+m_2+m_3+\ldots=k}} 2^{10c_1^3 k} 8^{-(1-10c_1^3)k} =$$

$$= \sum_{c_1 N\leq k\leq N} \sum_{\substack{\mathbf{m}=(m_1,m_2,m_3,\ldots)\in\mathbb{N}^L:\\ m_1+m_2+m_3+\ldots=k}} 2^{-(3-40c_1^3)k} =$$

$$= \sum_{c_1 N\leq k\leq N} \binom{k+L-1}{L-1} 2^{-(3-40c_1^3)k}, \tag{96}$$

where in the last step we used the well known combinatorial fact that the number of vectors $\mathbf{m} = (m_1, m_2, m_3, \ldots) \in \mathbb{N}^L$ satisfying $m_1 + m_2 + m_3 + \ldots = k$ is $\binom{k+L-1}{L-1}$ (called the number of k-combinations of a multiset with L types of elements, each type with unlimited repetition).

We are now ready to complete the proof of Lemma 5. By (59), (62), (63), (65), (72) and (96),

$$\sum_{\mathbf{x}\in\Phi} |\text{Zero}(\mathbf{x})| 2^{-nN} \leq \text{first part of the right hand side of (59)}+$$

$$+\text{second part of the right hand side of (59)} \leq$$

$$\leq \sum_{1\leq k\leq \frac{1}{10}n/\log n} \binom{N}{k} 2^k L^k 2^{-n} + \sum_{\frac{1}{10}n/\log n<k\leq n/3} \binom{N}{k} 2^k L^k k^{-n/2} +$$

$$+ \sum_{n/3<k\leq c_1 N} \binom{N}{k} 2^k L^k k^{-n/2} \leq$$

$$\leq 2^{-n/2} + 2^{-n} + 2^{-n} + \sum_{c_1 N\leq k\leq N} \binom{k+L-1}{L-1} 2^{-(3-40c_1^3)k}, \tag{97}$$

assuming we are in the range $N \geq 7n/5$. Since $c_1 = 1/4$, we have

$$(3 - 40c_1^3)k \geq (3 - 40c_1^3)N/4 = \left(3 - \frac{2}{3}\right) N/4 = 7N/12,$$

and using this in (97), we obtain

$$\sum_{\mathbf{x}\in\Phi} |\text{Zero}(\mathbf{x})| 2^{-nN} \leq 2 \cdot 2^{-n/2} + \binom{N+L}{L} 2^{-7N/12}. \tag{98}$$

If

$$L \leq N/120, \tag{99}$$

then by Stirling's formula,

$$\binom{N+L}{L} \leq \left(\frac{e(N+L)}{L}\right)^L \leq (121e)^{N/120} < \left(2^{10}\right)^{N/120} = 2^{N/12},$$

and using it in (98), we have

$$\sum_{\mathbf{x}\in\Phi} |\mathrm{Zero}(\mathbf{x})| 2^{-nN} \le 2\cdot 2^{-n/2} + 2^{N/12}2^{-7N/12} < 3\cdot 2^{-n/2}. \tag{100}$$

By (55),

$$L \le c_0 \left(\sqrt{N}\right)^{n/(N-n)}, \tag{101}$$

and the inequality

$$c_0 \left(\sqrt{N}\right)^{n/(N-n)} \le N/120 \tag{102}$$

clearly holds if $N \ge 3n/2$ (since the constant c_0 is much smaller than $1/120$; see (87)). Combining (99), (100), (101) and (102), Lemma 5 follows. □

This completes the proof of Theorem 1. □

Remark The proof technique of Theorem 1 is based on Lemma 4. Lemma 4 is a concentration inequality that makes use of "large multiplicities". If

$$\left(\frac{3}{2} - \varepsilon\right) n \ge N > n, \tag{103}$$

then

$$L = \left\lfloor c_0 \left(\sqrt{N}\right)^{n/(N-n)} \right\rfloor \ge N^{1+\varepsilon},$$

and a typical vector $(x_1, \ldots, x_N) \in \mathbb{Z}^N$ with $|x_i| \le L$ has very few coordinate repetitions, i.e., the contribution of "large multiplicities" is negligible. It means that in the range (103) Lemma 4 becomes useless.

Theorem 1 is about the range

$$N \ge 3n/2, \tag{104}$$

and we can save the proof of Theorem 1 in the slight extension of (104)

$$3n/2 > N \ge \left(\frac{3}{2} - \varepsilon\right) n \tag{105}$$

by replacing the definition of L in (55) with the slightly smaller threshold

$$L = \left\lfloor N^{\frac{(1-3\varepsilon)n}{2(N-n)}} \right\rfloor. \tag{106}$$

Indeed, all that we needed in the proof of Lemma 5 (the hard part of Theorem 1) was to be in the range $N \geq 7n/5$ and to satisfy the inequality $L \leq N/120$ (see (99)), and in the range (105) with $0 < \varepsilon \leq 1/10$ we have $N \geq 7n/5$ and

$$\frac{(1-3\varepsilon)n}{2(N-n)} \leq \frac{(1-3\varepsilon)n}{2\left(\frac{1}{2}-\varepsilon\right)n} = \frac{1-3\varepsilon}{1-2\varepsilon} < 1-\varepsilon,$$

and using it in (106), we obtain the desired inequality

$$L \leq N^{\frac{(1-3\varepsilon)n}{2(N-n)}} < N^{1-\varepsilon} < N/120.$$

Thus we obtain the following result.

Theorem 2 *For every $0 < \varepsilon \leq 10$ there is a finite threshold $n_0 = n_0(\varepsilon) < \infty$ with the following property: for every pair $N > n \geq n_0$ of positive integers satisfying (105), there exists a matrix $\mathbf{D} = (d_{i,j})$, $1 \leq i \leq n$, $1 \leq j \leq N$ with n rows, N columns, entries $d_{i,j} \in \{1, -1\}$ such that for every nontrivial integer solution*

$$\mathbf{x} = (x_1, x_2, \ldots, x_N) \in \mathbb{Z}^N \setminus \mathbf{0}$$

of the homogeneous linear system $\mathbf{Dx} = \mathbf{0}$, meaning the long form

$$\sum_{1 \leq j \leq N} d_{i,j}x_j = 0, \quad 1 \leq i \leq n,$$

the maximum norm of $\mathbf{x}$ has the lower bound

$$\max_{1 \leq j \leq N} |x_j| > N^{\frac{(1-3\varepsilon)n}{2(N-n)}}. \tag{107}$$

What is more, the overwhelming majority of the n-by-N ± 1 matrices $\mathbf{D}$ satisfy the theorem: the violators $\mathbf{D} = (d_{i,j})$, $1 \leq i \leq n$, $1 \leq j \leq N$ of Theorem 2 represent an exponentially small $O(2^{-n/2})$ part of the total 2^{nN}. □

To extend Theorem 1 to the range (103) one needs a more sophisticated version of Lemma 4. We are going to return to this problem in another paper. Theorem 2 will be the starting point of our study of the range (103).

Note that the same proof works for inhomogeneous linear systems, since every solution of an inhomogeneous linear system is automatically nontrivial. Thus we obtain the following theorem.

Theorem 3 *There is a positive absolute constant $c_0 > 0$ with the following property: for every pair $N > n \geq 1$ of positive integers satisfying $N \geq 3n/2$, and every nontrivial integer vector*

$$\mathbf{d} = (d_1, d_2, \ldots, d_n) \in \mathbb{Z}^n \setminus \mathbf{0},$$

there exists a matrix $\mathbf{D} = (d_{i,j})$, $1 \le i \le n$, $1 \le j \le N$ *with n rows, N columns, entries* $d_{i,j} \in \{1, -1\}$ *such that for every integer solution*

$$\mathbf{x} = (x_1, x_2, \ldots, x_N) \in \mathbb{Z}^N$$

of the inhomogeneous linear system $\mathbf{Dx} = \mathbf{d}$, *meaning the long form*

$$\sum_{1 \le j \le N} d_{i,j} x_j = d_i, \quad 1 \le i \le n,$$

the maximum norm of $\mathbf{x}$ *has the lower bound*

$$\max_{1 \le j \le N} |x_j| > c_0 \left(\sqrt{N}\right)^{n/(N-n)}. \tag{108}$$

Actually we have much more than pure existence: given an arbitrary but fixed vector

$$\mathbf{d} = (d_1, d_2, \ldots, d_n) \in \mathbb{Z}^n \setminus \mathbf{0},$$

for large n the overwhelming majority of the n-by-N ± 1 *matrices* $\mathbf{D}$ *satisfy* (108)*. In fact, the violators* $\mathbf{D} = (d_{i,j})$, $1 \le i \le n$, $1 \le j \le N$ *of Theorem* 3 *represent an exponentially small* $O(2^{-n/2})$ *part of the total* 2^{nN}*.* □

Note that Theorem 3 is trivial for "large" vectors $\mathbf{d} \in \mathbb{Z}^n \setminus \mathbf{0}$, where "large" means that the maximum norm of $\mathbf{d}$ is much larger than N. But for "small" vectors $\mathbf{d} \in \mathbb{Z}^n \setminus \mathbf{0}$ the statement of Theorem 3 is far from trivial.

Concluding Remarks After my talk about the sharpness of Siegel's Lemma in the Matousek Conference in Praga (Summer of 2016) Noga Alon mentioned two of his earlier results—with co-authors—that are somewhat related to the subject of this paper. The first one is Proposition 3.4.3 in Alon–Vu [2], which corresponds to the extreme case $N = n + 1$ *not* covered in Theorem 1. In the case $N = n + 1$ Proposition 3.4.3 gives a lower bound roughly $(n/2)^{n/4}$ for ±1-matrices. Moreover, Noga Alon pointed out that their method also gives a lower bound in the more general case of $N = k + n$ where $k \ge 2$ is fixed and n is large. The idea is to take tensor products with any non-singular k-by-k ±1-matrix, which gives a lower bound roughly $(n/2k)^{n/4k}$.

Noga Alon also mentioned the paper Alon–Kozlov [1] which is closer to the range of N and n that I am considering. They proved lower bounds—see Lemma 5.2—that are substantially weaker than mine in Theorem 1, but their paper represents a different approach that has its own advantages.

Note that some of these results are explicit constructions.

Acknowledgements I am grateful to David Masser (Basel) who called my attention to the problem.

Appendix: Proof of the Third Version of Siegel's Lemma

We combine the usual pigeonhole principle argument with probability theory; in particular, we borrow some ideas from the paper of Spencer [10]. We use the following variant of the large deviation theorem in probability theory.

Bernstein's inequality Let $Z_1, Z_2, \ldots, Z_N$ be real-valued independent random variables with zero expectation $\mathbf{E}Z_j = 0$ and $|Z_j| \le M$, $1 \le j \le N$. Then, for all positive $\tau > 0$,

$$\Pr\left[\left|\sum_{j=1}^{N} Z_j\right| \ge \tau\right] \le 2\exp\left(-\frac{\tau^2/2}{\left(\sum_{j=1}^{N} \mathbf{E}Z_j^2\right) + (\tau M/3)}\right).$$

Consider now the i-th row of the homogeneous linear system

$$\sum_{j=1}^{N} d_{i,j}x_j = 0, \text{ where } |d_{i,j}| \le A.$$

For a positive integer $B \ge 1$ and an integer j in $1 \le j \le N$, let X_j denote the random variable with $\Pr[X_j = b] = \frac{1}{2B+1}$, where b runs over the integers $-B$, $-B+1, -B+2, \ldots, B$. Moreover, assume that $X_1, X_2, \ldots, X_N$ are independent. We apply Bernstein's inequality with $Z_j = d_{i,j}X_j$, $\tau = \lambda\sqrt{N}AB$ and $M = AB$:

$$\Pr\left[\left|\sum_{j=1}^{N} Z_j\right| \ge \lambda\sqrt{N}AB\right] \le 2\exp\left(-\frac{\lambda^2NA^2B^2/2}{(NA^2B^2) + \left(\lambda\sqrt{N}A^2B^2/3\right)}\right) =$$

$$= 2\exp\left(-\frac{\lambda^2}{2 + 2N^{-1/2}\lambda/3}\right). \tag{109}$$

Let

$$\lambda_h = 6h, \quad 1 \le h \le \log n. \tag{110}$$

Then by (109), for every $1 \le h \le \log n$,

$$\Pr\left[\left|\sum_{j=1}^{N} d_{i,j}X_j\right| \ge 6h\sqrt{N}AB\right] \le 2\exp\left(-\frac{36h^2}{2 + (12N^{-1/2}h/3)}\right) \le$$

$$\le 2\exp\left(-\frac{36h^2}{2h + (12h/3)}\right) = 2e^{-6h}. \tag{111}$$

For every integer h in $1 \leq h \leq \log n$, define the random variable

$$Y_h = \left| \left\{ i \in \{1, 2, \ldots, n\} : \left| \sum_{j=1}^{N} d_{i,j} X_j \right| \geq 6h\sqrt{N}AB \right\} \right|. \tag{112}$$

By using (111), we obtain the following upper bound for the expected value of Y_h:

$$\mathbf{E}Y_h \leq 2e^{-6h}n, \ 1 \leq h \leq \log n. \tag{113}$$

Since the random variable Y_h has non-negative values, we can use the simple Markov inequality stating that for any random variable Y with non-negative values and finite expectation

$$\Pr[Y \geq a] \leq \frac{\mathbf{E}Y}{a} \text{ for any } a > 0.$$

Applying Markov inequality in (113) we have

$$\Pr\left[Y_h \geq 2h(h+1) \cdot 2e^{-6h}n\right] \leq \frac{1}{2h(h+1)}, \ 1 \leq h \leq \log n.$$

Using the telescoping sum

$$\sum_{h=1}^{\infty} \frac{1}{h(h+1)} = \sum_{h=1}^{\infty} \left(\frac{1}{h} - \frac{1}{h+1} \right) = 1,$$

we obtain that

$$\sum_{1 \leq h \leq \log n} \frac{1}{2h(h+1)} < \frac{1}{2},$$

so, with probability greater than $1/2$ we have

$$Y_h < 2h(h+1) \cdot 2e^{-6h}n \text{ for every } 1 \leq h \leq \log n. \tag{114}$$

(114) means that, with probability greater than $1/2$, the number of row-sums $\sum_{j=1}^{N} d_{i,j}X_j$ that have absolute value $\geq 6h\sqrt{N}AB$, is less than

$$2h(h+1) \cdot 2e^{-6h}n \text{ for every } 1 \leq h \leq \log n.$$

It follows that, with probability greater than $1/2$, the number of row-sums $\sum_{j=1}^{N} d_{i,j}X_j$ that have absolute value between $6h\sqrt{N}AB$ and $6(h+1)\sqrt{N}AB$, is less than

$$2h(h+1)\cdot 2e^{-6h}n \text{ for every } 1 \le h \le \log n, \tag{115}$$

and there is no row-sum with absolute value $\ge 6\log n\sqrt{N}AB$. In the last step we used the fact that

$$2h(h+1)\cdot 2e^{-6h}n < 1 \text{ with } h = \lfloor \log n \rfloor$$

(lower integral part).

Write (see (115))

$$k_h = \lfloor 2h(h+1)\cdot 2e^{-6h}n \rfloor, \ \ 1 \le h \le \log n, \tag{116}$$

and

$$k_0 = n - \sum_{1 \le h \le \log n} k_h. \tag{117}$$

The total number of row-sum vectors (with n coordinates) satisfying (115) can be estimated from above by using the parameters k_h in (116)–(117) as follows:

$$\le \binom{n}{k_0}\left(2\cdot 6\sqrt{N}AB+1\right)^{k_0} \prod_{1\le h\le \log n}\left(\binom{n}{k_h}\left(2\cdot 6(h+1)\sqrt{N}AB+1\right)^{k_h}\right). \tag{118}$$

On the other hand, “with probability greater than $1/2$” means more than

$$\frac{1}{2}(2B+1)^N \tag{119}$$

possible vectors $\mathbf{v} \in \{-B, -B+1, -B+2, \dots, B\}^N$.

So, if (119) is greater or equal to (118), than the Pigeonhole Principle applies, and there exist two different vectors

$$\mathbf{v}_1, \mathbf{v}_2 \in \{-B, -B+1, -B+2, \dots, B\}^N$$

such that they generate the same row-vector, i.e., $\mathbf{D}\mathbf{v}_1 = \mathbf{D}\mathbf{v}_2$. Then $\mathbf{x} = \mathbf{v}_1 - \mathbf{v}_2$ satisfies the homogeneous linear system $\mathbf{D}\mathbf{x} = \mathbf{0}$ with

$$\max_{1\le j\le N} |x_j| \le 2B. \tag{120}$$

The rest is routine estimation. Clearly

$$\binom{n}{k_0}\left(2\cdot 6\sqrt{N}AB+1\right)^{k_0}\prod_{1\le h\le \log n}\left(\binom{n}{k_h}\left(2\cdot 6(h+1)\sqrt{N}AB+1\right)^{k_h}\right)\le$$

$$\le\binom{n}{k_0}\left(13\sqrt{m}AB\right)^{k_0}\prod_{1\le h\le \log n}\left(\binom{n}{k_h}\left(13(h+1)\sqrt{N}AB\right)^{k_h}\right)=$$

$$=\left(13\sqrt{N}AB\right)^{n}\binom{n}{k_0}\prod_{1\le h\le \log n}\left(\binom{n}{k_h}(h+1)^{k_h}\right).\tag{121}$$

By using $s!\ge (s/e)^s$ and (116), we have

$$\binom{n}{k_0}\prod_{1\le h\le \log n}\left(\binom{n}{k_h}(h+1)^{k_h}\right)\le\binom{n}{k_0}\prod_{1\le h\le \log n}\left(e^{6h}(h+1)\right)^{k_h}\le$$

$$\le\binom{n}{k_0}\prod_{1\le h\le \log n}e^{7hk_h}=\binom{n}{k_0}\exp\left(\sum_{1\le h\le \log n}7hk_h\right)\le$$

$$\le\binom{n}{k_0}\exp\left(\sum_{1\le h\le \log n}7h\cdot 3h(h+1)e^{-6h}n\right)\le$$

$$\le\binom{n}{k_0}\exp\left(21n\sum_{h=1}^{\infty}h^2(h+1)e^{-6h}\right)\le 2^n e^n.\tag{122}$$

Using (122) in (121), we have

$$\binom{n}{k_0}\left(2\cdot 6\sqrt{N}AB+1\right)^{k_0}\prod_{1\le h\le \log n}\left(\binom{n}{k_h}\left(2\cdot 6(h+1)\sqrt{N}AB+1\right)^{k_h}\right)\le$$

$$\le\left(13\sqrt{N}AB\right)^{n}2^n e^n<\left(70\sqrt{N}AB\right)^{n}.\tag{123}$$

Combining (118), (119) and (123), it suffices to guarantee the inequality

$$\frac{1}{2}(2B+1)^N\ge\left(70\sqrt{N}AB\right)^{n}.\tag{124}$$

Inequality (124) clearly holds with

$$2B = \left\lfloor \left(70\sqrt{N}A\right)^{n/(N-n)} \right\rfloor, \tag{125}$$

and using (120) in (125), we conclude

$$\max_{1\le j\le N} |x_j| \le \left(70\sqrt{N}A\right)^{n/(N-n)},$$

completing the proof of the Third Version of Siegel's Lemma. □

References

1. N. Alon, D.N. Kozlov, Coins with arbitrary weights. J. Algorithms **25**, 162–176 (1997)
2. N. Alon, V.H. Vu, Anti-Hadamard matrices, coin weighing, threshold gates and indecomposable hypergraphs. J. Comb. Theory Ser. A **79**, 133–160 (1997)
3. A. Baker, *Transcendental Number Theory* (Cambridge University Press, Cambridge, 1975)
4. E. Bombieri, W. Gubler, *Heights in Diophantine Geometry*. New Mathematical Monographs, vol. 4 (Cambridge University Press, Cambridge, 2006)
5. E. Bombieri, J.D. Vaaler, On Siegel's lemma. Invent. Math. **73**(1), 11–32 (1983)
6. P. Erdős, On a lemma of Littlewood and Offord. Bull. Am. Math. Soc. **51**, 898–902 (1945)
7. G. Halász, Estimates for the concentration function of combinatorial number theory and probability. Period. Math. Hung. **8**, 197–211 (1977)
8. A.M. Macbeath, On measure of sum-sets, II. The sum-theorem for the torus. Proc. Camb. Philos. Soc. **49**, 40–43 (1953)
9. W.M. Schmidt, *Diophantine Approximations and Diophantine Equations*. Lecture Notes in Mathematics, vol. 1467 (Springer, Berlin, 1991)
10. J. Spencer, Six standard deviations suffice. Trans. Am. Math. Soc. **289**, 679–706 (1985)
11. J.D. Vaaler, The best constant in Siegel's lemma. Monatshaft. Math. **140**(1), 71–89 (2003)
12. J.D. Vaaler, A.J. van der Poorten, Bounds for solutions of systems of linear equations. Bull. Aust. Math. Soc. **25**, 125–132 (1982)

On Codimension One Embedding of Simplicial Complexes

Anders Björner and Afshin Goodarzi

Dedicated to the memory of Jiří Matoušek

Abstract We study d-dimensional simplicial complexes that are PL embeddable in $\mathbb{R}^{d+1}$. It is shown that such a complex must satisfy a certain homological condition. The existence of this obstruction allows us to provide a systematic approach to deriving upper bounds for the number of top-dimensional faces of such complexes, particularly in low dimensions.

1 Introduction

The question of embeddability of a d-dimensional simplicial complex into k-dimensional Euclidean space $\mathbb{R}^k$ has a long history. In the following section we sketch some of this background. Technical definitions and details appear in later sections. See J. Matoušek's book [13, chapter 5] and his paper with M. Tancer and U. Wagner [14] for nice introductions to the field.

In this note we provide a homological obstruction to codimension one ($k = d+1$) piecewise linear (PL) embeddability of simplicial complexes. For the case of graphs ($d = 1$) this kind of obstruction was used by S. Mac Lane [11] in his work on planarity.

A. Björner (✉)
Department of Mathematics, Royal Institute of Technology, S-100 44 Stockholm, Sweden
e-mail: bjorner@kth.se

A. Goodarzi
Department of Mathematics, Discrete Geometry Group, Free University of Berlin,
14195 Berlin, Germany
e-mail: goodarzi@math.fu-berlin.de

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_9

As corollaries we derive upper bounds for the number of top-dimensional faces in a complex with codimension one PL embedding, in terms of the lower dimensional face numbers and Betti numbers. For instance, we show that

$$f_d(\Sigma) \leq \frac{\mathrm{g}(\Sigma)}{\mathrm{g}(\Sigma)-2}\left(\left(\sum_{i=1}^{d}(-1)^{i-1}\left(f_{d-i}(\Sigma)-\beta_{d-i}(\Sigma)\right)\right)-1\right),$$

where $f_i(\Sigma)$ is the number of faces of dimension i, β_i is the Betti number in dimension i, and $\mathrm{g}(\Sigma)$ is the girth (smallest size of a d-cycle in non-zero homology). See Theorem 3 for details. For $d = 1$ and $\mathrm{g}(\Sigma) = 3$ this specializes to Euler's $3n-6$ upper bound for the maximal number of edges of a planar graph.

The method used enables us to provide a unified approach and to give more detailed versions of face number inequalities for such complexes in low dimensions. For instance, we obtain that

$$f_2(\Sigma) \leq 2\,(f_1(\Sigma) - f_0(\Sigma) - \beta_1(\Sigma))$$

for any connected 2-dimensional complex Σ that PL embeds into $\mathbb{R}^3$, see Proposition 8. Furthermore, we give a new upper bound for the number of facets of complexes with codimension one PL embedding, in terms only of the number of vertices. This slightly improves the upper bound given by Dey and Pach [6].

Finally, some of our face number inequalities are adapted to the case of balanced complexes, i.e., complexes whose 1-skeleton is $(d+1)$-colorable in the graph-theoretic sense.

2 Background

The concept of planarity has been of interest to mathematicians ever since the subject of graph theory was founded. For instance, the impossibility for a planar graph on n (≥ 3) vertices of having more than $3n-6$ edges was mentioned in a letter from L. Euler to C. Goldbach in 1750, see [1, p. 75].

A topological characterisation of planarity was given by K. Kuratowski in 1929 and independently (a few months later) by O. Frink and P.A. Smith. This result asserts that a finite graph is planar if and only if it does not contain a subgraph homeomorphic to K_5 or $K_{3,3}$. Since then other characterisations of planarity have been given. Among them one can mention the more combinatorial approaches by H. Whitney [23] and S. Mac Lane [11], and the more topological approach of H. Hanani and W.T. Tutte (see [20], for instance).

What can be said about the situation in higher dimensions? Let Σ be a finite d-dimensional simplicial complex. It was known since the early days of topology that Σ is linearly embeddable into $\mathbb{R}^{2d+1}$. In his 1933 article, E. R. van Kampen [21] showed that this result is best possible, by presenting d-dimensional complexes (now known as the van Kampen–Flores complexes) that do not embed into $\mathbb{R}^{2d}$. Thus, the natural question is, for $d \leq k \leq 2d$, does Σ admit an embedding into $\mathbb{R}^k$? The most intensively investigated cases are when $k = 2d$ or $k = d + 1$. Note that these are the two natural generalisations to higher dimensions of the concept of planarity.

There is no satisfactory analogue of Kuratowski's characterisation in higher dimensions. Indeed, for every $d > 1$ and $d + 1 \leq k \leq 2d$, J. Zaks [24] constructed infinitely many pairwise non-homeomorphic d-dimensional complexes that are minimal with respect to the property of being not embeddable in $\mathbb{R}^k$.

Based on the aforementioned work of van Kampen, in 1957 A. Shapiro [18] introduced the *van Kampen obstruction*; a cohomological obstruction to embeddability of d-dimensional complexes into $\mathbb{R}^{2d}$. See [14] for a geometric description. The van Kampen obstruction can be seen as a higher-dimensional analogue of the Hanani–Tutte theorem, though the strong version of Hanani–Tutte theorem appeared much later in [20].

3 Embedding

A simplicial complex Σ is said to admit a *linear embedding* into $\mathbb{R}^k$ if Σ has a geometric realisation $\|\Sigma\|$ in $\mathbb{R}^k$. More generally, Σ admits a *topological embedding* into $\mathbb{R}^k$ if there is a continuous injection $\|\Sigma\| \hookrightarrow \mathbb{R}^k$, from some geometric realisation of Σ to $\mathbb{R}^k$. An intermediate concept is that of *PL embedding*. We say that Σ is *piecewise linear* (*PL*) embeddable into $\mathbb{R}^k$ if there is a subdivision of $\|\Sigma\|$ that linearly embeds into $\mathbb{R}^k$. In this paper we focus on PL embeddings.

It is a consequence of Steinitz' Theorem [25, Lect. 4] that every planar graph can be drawn in the plane with straight edges. However, for higher dimensional objects the situation is more complicated.

Example 1 (Brehm's triangulated Möbius strip) In [3], Brehm presented a triangulation of the Möbius strip that can not be geometrically realised in $\mathbb{R}^3$. The idea is simple but elegant: A non null-homotopic curve, different from the center line, and the boundary curve of the Möbius strip are linked together, with absolute value of the linking number at least 2. This can easily be visualised by, for instance, considering the blue curve on the left hand side of Fig. 1 below. Now, triangulate the Möbius strip in such a way that the blue curve and the boundary curve are induced triangles; see the right hand side of Fig. 1. Two triangles with straight edges in $\mathbb{R}^3$ are either the unlink or the Hopf link. Hence, these two triangles cannot be realised by straight edges. Iterated simplicial suspensions produce examples of d-dimensional complexes that are PL embeddable into $\mathbb{R}^{d+1}$ but do not admit a linear embedding.

The difference between linear and PL embedding is even more dramatic. One can show that the problem of linear embeddability is algorithmically decidable. On the other hand, it is shown in [14, Theorem 1.1] that codimension one PL embeddability is algorithmically undecidable for $d \geq 4$. See [14] for a thorough discussion.

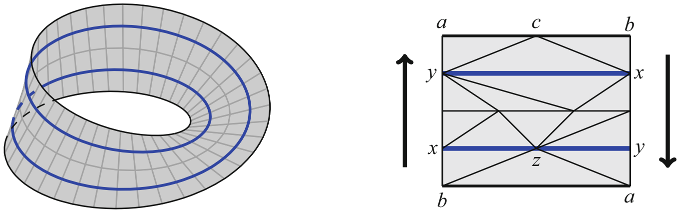


Fig. 1 Brehm's triangulated Möbius strip

Let us also remark that topological and PL embeddings do not coincide in codimension one. In fact, by the double suspension theorem [4], the suspension of the Poincaré homology 3-sphere topologically embeds into $\mathbb{R}^5$. However, it does not admit a PL embedding into $\mathbb{R}^5$ [22, p. 576].

4 Main Results

Let Σ be a d-dimensional simplicial complex. We consider simplicial homology of Σ with $\mathbb{Z}_2$ coefficients. Let $c = \sum \epsilon_\sigma \sigma$ be a d-chain, where the sum is over all d-dimensional faces of Σ and $\epsilon_\sigma \in \mathbb{Z}_2$. We let the *support* $\mathrm{supp}(c)$ of c be the set of all d-faces σ such that $\epsilon_\sigma = 1$.

Let us say that a basis $\mathcal{B}$ of $H_d(\Sigma; \mathbb{Z}_2)$ is *m-complete* if every d-dimensional face of Σ appears in the support of at most m elements in $\mathcal{B}$. When $d = 1$, this definition agrees with Mac Lane's concept of *m-fold complete set of cycles* for graphs. He showed that having a 2-fold complete set of cycles is equivalent to planarity for graphs [11]. In this section we generalise one direction of Mac Lane's result. Before doing so, we need to show the following topological invariance property.

Lemma 1 *Let Σ and Γ be two triangulations of a d-dimensional topological space X. Then Σ has an m-complete basis if and only if Γ has a m-complete basis.*

Proof Let $H_d(X; \mathbb{Z}_2)$ be the singular homology group of X (this is the only place in this paper where we use singular homology theory). We refer to the book [16] by Munkres for the definition and properties of the singular homology.

Let $H_d(X; \mathbb{Z}_2) = \mathbb{Z}_2^r$. We can always assume that there are d-dimensional pseudomanifolds $M_1, \dots, M_r$ and continuous maps $f^i : M_i \to X$, for $1 \le i \le r$, so that the d-dimensional homology classes of X are $f^i_\sharp[M_i]$, where $[M_i]$ is the fundamental class of M_i. We claim that a triangulation of X has an m-complete basis if and only if there is a choice of M_i and f^i such that for any subset I of $\{1, 2, \dots, r\}$ of size greater than m one has

$$\dim \left(\bigcap_{i \in I} f^i(M_i) \right) < d.$$

Observe that once the claim is verified the desired statement is immediate. However, the verification of the claim is standard and we leave it to the reader. □

Remark 1 Since we are working with $\mathbb{Z}_2$ coefficients, it follows from a result by Thom that, $M_1, \dots, M_r$ in the proof of Lemma 1 can be taken to be closed manifolds. See, for instance, [19, p. 343].

Theorem 1 *Let Σ be a d-dimensional simplicial complex that admits a PL embedding into $\mathbb{R}^{d+1}$. Then $H_d(\Sigma; \mathbb{Z}_2)$ has a 2-complete basis.*

Proof First notice that, by Lemma 1, Σ has a 2-complete basis if and only if any subdivision of Σ has this property. This allows us to replace Σ by a subdivision of Σ if needed. Also, observe that since Σ is PL embeddable into $\mathbb{R}^{d+1}$, then Σ is PL embeddable into the $(d+1)$-simplex Δ_{d+1}. Thus there is a subdivision Σ' of Σ and a subdivision $\mathbf{B}$ of Δ_{d+1} such that Σ' is a subcomplex of $\mathbf{B}$. So, we may assume that Σ' is a subcomplex of a simplicial $(d+1)$-sphere $\mathbf{S}$, say by embedding $\mathbf{B}$ into a hyperplane H of $\mathbb{R}^{d+2}$ and taking $\mathbf{S} = \{p\} * \partial\mathbf{B} \cup \mathbf{B}$, where p is a point outside H and $*$ denotes the simplicial cone.

Now, set $r := \beta_d(\Sigma'; \mathbb{Z}_2) + 1$. There is nothing to prove if $r = 1$. So, we may assume that $r > 1$. It follows from Alexander duality [16, Theorem 71.1] that $\|\mathbf{S}\| - \|\Sigma'\|$ has r connected components, say $K_1, \dots, K_r$. For $1 \le j \le r$, let c_j be the formal sum (modulo 2) of all facets F of $\mathbf{S}$ such that the barycenter of F lies in K_j. Let b_j be the boundary $\partial_{d+1} c_j$ of c_j. Notice that $b_j \neq 0$, since $r > 1$ and therefore, c_j cannot be a $(d+1)$-cycle.

We will show that $b_1, \dots, b_{r-1}$ form a 2-complete basis for $H_d(\Sigma'; \mathbb{Z}_2)$.

Let $\sigma \in \operatorname{supp}(b_j)$ for some $1 \le j \le r$. Then σ is a facet of Σ'. Otherwise, the facets F_σ and F'_σ of $\mathbf{S}$ that contain σ lie in the same connected component K_j. This implies that F_σ and F'_σ are in the support of c_j. Hence, $\sigma \notin \operatorname{supp}(b_j)$, which is a contradiction. Also, observe that there exists exactly one $i \neq j$ such that $\sigma \in \operatorname{supp}(b_i)$, since every codimension one face of $\mathbf{S}$ is in exactly two facets.

It is immediate that $\partial_d b_j = \partial_d \partial_{d+1} c_j = 0$, hence every b_j is a d-cycle in $\mathbf{S}$. However, since $\operatorname{supp}(b_j)$ is a subset of the set of faces of Σ', then every b_j is a d-cycle in Σ'.

Finally, we have that $\sum_{i \in A} b_i \neq 0$ for all proper subsets A of $\{1, \dots, r\}$. Otherwise,

$$\partial_{d+1}\left(\sum_{i \in A} c_i\right) = \left(\sum_{i \in A} \partial_{d+1} c_i\right) = \sum_{i \in A} b_i = 0,$$

that is, the subcomplex of $\mathbf{S}$ whose set of facets are c_i, $i \in A$, has non-trivial $(d+1)$-dimensional homology. However, this cannot happen, since every proper subcomplex of $\mathbf{S}$ has trivial $(d+1)$-dimensional homology. Therefore, $b_1, \dots, b_{r-1}$ is a 2-complete basis for $H_d(\Sigma'; \mathbb{Z}_2)$, as promised. □

Remark 2 It might be possible that the conclusion of Theorem 1 is still valid if we consider the more general case of topological embedding. However, since we use Alexander duality, our method would not be directly applicable in that general setting.
Notice that the converse to Theorem 1 is obviously false for all $d > 1$. For instance, there are d-manifolds that do not admit an embedding into $\mathbb{R}^{d+1}$; non-orientable manifolds for example. In fact, it follows from Alexander duality that if Σ is embeddable into the $(d+1)$-sphere $\mathbb{S}^{d+1}$, then the cohomology $H^d(\Sigma; \mathbb{Z})$ is isomorphic to $\widetilde{H}_0(\mathbb{S}^{d+1} \setminus \Sigma; \mathbb{Z})$ and, thus, is torsion-free.

Having Theorem 1 in mind it is tempting to conjecture that if a d-dimensional simplicial complex Σ embeds into $\mathbb{R}^{d+m-1}$, then $H_d(\Sigma; \mathbb{Z}_2)$ has an m-complete basis. The following example shows that this is not the case.

Example 2 Let n be an integer and let Δ be the 2-dimensional complex obtained by suspending the complete bipartite graph $K_{n,n}$. Clearly, $f(\Delta) = (n+2, n^2+2n, 2n^2)$ and $\beta_2(\Delta) = n^2 - 2n + 1$. On the other hand, Δ (being a suspension of a complex embeddable in 3-space) is embeddable into $\mathbb{R}^4$. However, we show that for large enough n, $H_2(\Delta; \mathbb{Z}_2)$ does not have a 3-complete basis. First observe that if Ω is a minimal cycle in Δ, then Ω has at least 8 triangles. Now, let $\mathcal{B}$ be a basis for $H_2(\Delta; \mathbb{Z}_2)$ and let M be the $n^2 - 2n + 1$ by $2n^2$ $\{0, 1\}$-matrix whose rows are labeled by the elements Ω of $\mathcal{B}$ and whose columns are labeled by the facets of Δ, and for which the entry (F, Ω) is the coefficient of F in Ω. Since the number of facets with non-zero coefficient in each element of $\mathcal{B}$ is at least 8, the minimum number of 1s in M is $8(n^2 - 2n + 1)$. On the other hand, if $H_2(\Delta; \mathbb{Z}_2)$ has a 3-complete basis, then the maximum number of 1 s in M must be 3 times the number of facets, that is, $6n^2$. Therefore, if n is large enough then $H_2(\Delta; \mathbb{Z}_2)$ does not have a 3-complete basis.

5 Face Numbers

In this section we provide upper bounds for the number of top dimensional faces of complexes that admit a codimension one embedding in terms of the lower dimensional face numbers and Betti numbers.

For a d-dimensional simplicial complex Σ, with non-trivial top Betti number, let us define the *girth* of Σ, denoted $\mathrm{g}(\Sigma)$, to be the minimum number of d-dimensional faces of a subcomplex with non-zero d-dimensional Betti number. This notion extends the graph theoretic notion of girth as the minimal size of a circuit. If $\beta_d(\Sigma) = 0$ we define the girth to be $d + 2$. Note that the girth of a d-dimensional complex satisfies $\mathrm{g}(\Sigma) \geq d + 2$.

Theorem 2 *Let Σ be a d-dimensional simplicial complex such that $H_d(\Sigma; \mathbb{Z}_2)$ admits a 2-complete basis. Then*

$$\mathrm{g}(\Sigma)\,(\beta_d(\Sigma; \mathbb{Z}_2) + 1) \leq 2f_d(\Sigma). \tag{1}$$

Proof Let r and $b_1, \ldots, b_r$ be as defined in the proof of Theorem 1. On the one hand, for $1 \leq j \leq r$, $\operatorname{supp}(b_j)$ has at least $\mathrm{g}(\Sigma)$ elements. On the other hand, a d-dimensional face of Σ appears, if at all, in the support of two of the b_j's. Therefore, $\mathrm{g}(\Sigma) r \leq 2 f_d(\Sigma)$, as desired. □

To help simplify the notation, let $\delta_j = f_j(\Sigma) - \beta_j(\Sigma; \mathbb{Z}_2)$, for all j. Then, let

$$\chi_{j-1}(\Sigma) = \sum_{i=1}^{j} (-1)^{i-1} \delta_{j-i}$$

It follows from the rank-nullity theorem that $\chi_{j-1}(\Sigma) \geq 0$ for all j. These inequalities, sometimes called the *strong Morse inequalities*, are discussed in Milnor [15], and appear in slightly sharper form in [2].

Theorem 3 *Let Σ be a d-dimensional simplicial complex that admits a PL embedding into $\mathbb{R}^{d+1}$. Then,*

$$f_d(\Sigma) \leq \frac{\mathrm{g}(\Sigma)}{\mathrm{g}(\Sigma) - 2} (\delta_{d-1} - \delta_{d-2} + \delta_{d-3} - \cdots + \delta_{d-k} - 1) \tag{2}$$

for all odd $k \geq 1$.

Proof Our point of departure is the inequality (1) of Theorem 2. Replace $\beta_d(\Sigma; \mathbb{Z}_2)$ in the left hand side of the inequality by the right hand side of the following form of the Euler-Poincaré formula:

$$\beta_d(\Sigma; \mathbb{Z}_2) = f_d(\Sigma) - \chi_{d-1}(\Sigma),$$

and then simplify and use $\chi_{d-k-1} \geq 0$ to get the desired inequality. □

Corollary 4 *Let Σ be a d-dimensional simplicial complex that admits a PL embedding into $\mathbb{R}^{d+1}$. Then,*

$$f_d(\Sigma) \leq \frac{d+2}{d} (f_{d-1} - \beta_{d-1} - 1).$$

Proof This is the $k = 1$ case of Theorem 3, using that $\mathrm{g}(\Sigma) \geq d + 2$. □

Next, we focus on balanced simplicial complexes. Recall that a d-dimensional simplicial complex is said to be *balanced* if its underlying graph (1-skeleton) is $(d+1)$-colorable in the graph theoretic sense.

Theorem 5 *Let Σ be a balanced d-dimensional simplicial complex that admits a PL embedding into $\mathbb{R}^{d+1}$. Then the following hold true:*

(a) $2^d (\beta_d(\Sigma; \mathbb{Z}_2) + 1) \leq f_d(\Sigma)$;
(b) $f_d(\Sigma) \leq \frac{2^d}{2^d - 1} (\chi_{d-1} - 1)$.

Proof It suffices to show that the girth of a balanced d-dimensional simplicial complex is at least 2^{d+1}. The crucial point is that a balanced d-dimensional complex with non-zero top dimensional homology has at least 2^{d+1} faces of dimension d. To see this one can observe that such a complex must contain a balanced d-dimensional pseudomanifold without boundary; the pure complex whose facets are support of a d-cycle. The claim then can be proved easily for pseudomanifolds, say by induction on the dimension. We leave it to the reader to fill in the details. □

Our method is applicable also to complexes that admit a codimension zero embedding. For this, we first need to prove an auxiliary result.

Lemma 2 *Let Σ be a d-dimensional simplicial complex and let $\Sigma^{(-1)}$ denote its codimension one skeleton. Then one has*

$$f_d(\Sigma) = \beta_d(\Sigma; \mathbb{Z}_2) - \beta_{d-1}(\Sigma; \mathbb{Z}_2) + \beta_{d-1}(\Sigma^{(-1)}; \mathbb{Z}_2).$$

Proof We have that $f_i(\Sigma^{(-1)}) = f_i(\Sigma)$ for all $i \le d-1$, and $\beta_i(\Sigma^{(-1)}) = \beta_i(\Sigma)$ for all $i \le d-2$. Hence, by the Euler-Poincaré formula

$$(-1)^d f_d(\Sigma) = \chi(\Sigma) - \chi(\Sigma^{(-1)}) = (-1)^d \left(\beta_d(\Sigma) - \beta_{d-1}(\Sigma) + \beta_{d-1}(\Sigma^{(-1)}\right)$$

□

Corollary 6 *Let Σ be a d-dimensional simplicial complex that admits a PL embedding into $\mathbb{R}^d$. Then $f_d(\Sigma) \le \frac{2}{d+1} f_{d-1}(\Sigma) - 1$.*

Proof It can easily be shown, say by using Alexander duality, that the top dimensional homology of Σ must be zero. Thus, it follows from Lemma 2 that $\beta_{d-1}(\Sigma^{(-1)}; \mathbb{Z}_2) \ge f_d(\Sigma)$. Now, applying Theorem 2 to $\Sigma^{(-1)}$ we get

$$(d+1)(f_d(\Sigma)+1) \le (d+1)(\beta_{d-1}(\Sigma^{(-1)}; \mathbb{Z}_2)+1) \le 2f_{d-1}(\Sigma^{(-1)}) = 2f_{d-1}(\Sigma).$$

□

6 Corollaries in Low Dimensions

In this section we summarise direct consequences of the main results for embeddings into dimensions 2, 3 and 4. Throughout, the number of vertices of a simplicial complex is denoted by n (rather than f_0).

Proposition 7 *Let Σ be a 2-dimensional complex that PL embeds into $\mathbb{R}^2$. Then $f_2(\Sigma) \le \frac{2}{3} f_1(\Sigma) - 1$. In particular, $f_2(\Sigma) \le 2n - 5$.*

Proof The first inequality follows from Corollary 6. The second inequality follows from the fact that the underlying graph of Σ is planar. □

Proposition 8 *Let Σ be a connected 2-dimensional complex that PL embeds into $\mathbb{R}^3$. Then $f_2(\Sigma) \le 2(f_1(\Sigma) - \beta_1(\Sigma) - n)$.*

Proof This follows easily from Theorem 3. □

Corollary 9 (Dey–Edelsbrunner [5]) *Let Σ be a 2-dimensional complex that PL embeds into $\mathbb{R}^3$. Then $f_2(\Sigma) \le n(n-3)$.*

Proof Without loss of generality, we may assume that Σ is connected. The inequality is an immediate consequence of Proposition 8 and the trivial $\binom{n}{2}$ upper bound for $f_1(\Sigma)$. □

Corollary 10 *Let Σ be a 3-dimensional complex that PL embeds into $\mathbb{R}^3$. Then $f_3(\Sigma) \le n(n-3)/2 - 1$.*

Proof This follows from Corollaries 6 and 9. □

Proposition 11 *Let Σ be a connected balanced 2-dimensional complex that embeds into $\mathbb{R}^3$. Then $f_2(\Sigma) \le \frac{4}{9}(n^2 - 3n)$.*

Proof It follows from Theorem 5 that $f_2(\Sigma) \le \frac{4}{3}(f_1(\Sigma) - n)$. Now, since the underlying graph of Σ is 3-colorable, one has $f_1(\Sigma) \le 3(\frac{n}{3})^2$. The conclusion now follows easily. □

For embeddings into dimension 4 much less is known. It was conjectured by Kalai and Sarkaria (see Kalai's blog [9], for instance) that if a 2-dimensional complex is embedded into $\mathbb{R}^4$, then it has at most $2n(n-1)$ triangles. This conjecture is wide open. Currently, the best known bound [17] is $C \cdot n^{8/3}$, where C is a constant. Here is what our method yields in the case of embeddings into dimension four.

Proposition 12 *Let Σ be a connected 3-dimensional complex that PL embeds into $\mathbb{R}^4$. Then $f_3(\Sigma) \le \frac{5}{3}\,(f_2(\Sigma) - f_1(\Sigma) - \beta_2(\Sigma) + \beta_1(\Sigma) + n - 2)$.*

Proof This follows from Theorem 3. □

Corollary 13 *Let Σ be a connected 3-dimensional complex that PL embeds into $\mathbb{R}^4$. Then,*

$$f_3(\Sigma) \le \frac{5}{3}(f_2(\Sigma) + \beta_1(\Sigma) - 1) \ \textit{and}\ f_3(\Sigma) \le \frac{5}{3}\left(\binom{n}{3} + n - 2\right).$$

If Σ is simply connected, then $f_3(\Sigma) \le \frac{5}{3}(f_2(\Sigma) - 1)$.

Proof The inequalities are immediate consequences of Proposition 12 and the trivial $\binom{n}{3}$ upper bound for $f_2(\Sigma)$. □

7 Estimates

In the following we give an upper bound for the number of top dimensional faces of a d-dimensional simplicial complex embedded into $\mathbb{R}^{d+1}$ in terms of the number of its vertices. Let us begin by observing that for a d-complex Σ on n vertices one has $f_{d-1}(\Sigma) \leq \binom{n}{d}$. Hence, it follows from Theorem 3 that $f_d(\Sigma) < (1 + \frac{2}{d})\binom{n}{d}$. Therefore, we can easily obtain the upper bound $f_d(\Sigma) = \mathcal{O}(n^d)$ due to Dey and Pach [6, Theorem 3.1], where $\mathcal{O}$ is the big O notation.

Below we present a slightly better upper bound by using our Theorem 3 and a combination of an idea due to Gundert [10] and Sperner's Lemma [7, Lemma 4.5]. Recall that Sperner's Lemma asserts that for a simplicial complex Σ on n vertices the quantity $f_i(\Sigma)/f_{i-1}(\Sigma)$ is at most $\binom{n}{i+1}/\binom{n}{i}$. Notice that Sperner's Lemma can easily be strengthened to

$$f_i(\Sigma)/f_j(\Sigma) \leq \binom{n}{i+1}\Big/\binom{n}{j+1} = \mathcal{O}(n^{i-j}),$$

for all $i > j$.

Theorem 14 *Let Σ be a d-dimensional simplicial complex that admits a PL embedding into $\mathbb{R}^{d+1}$. Then $f_d(\Sigma) = \mathcal{O}(n^{d-\epsilon})$, where $\epsilon = 3^{-\lceil \frac{d+1}{2} \rceil}$.*

Proof Let us, to simplify notation, put $\ell = \lceil \frac{d+1}{2} \rceil$. Let Δ be the ℓ-dimensional skeleton of Σ. Since Δ is embeddable into $\mathbb{R}^d$, it follows from [10, Proposition 3.3.5] that

$$f_\ell(\Sigma) = f_\ell(\Delta) = \mathcal{O}(n^{\ell+1-3^{-\ell}}).$$

Now, it follows from Sperner's Lemma that $f_{d-1}(\Sigma) = \mathcal{O}(n^{d-\ell-1})f_\ell(\Sigma)$. Therefore, one obtains that $f_{d-1}(\Sigma) = \mathcal{O}(n^{d-3^{-\ell}})$. Finally, the conclusion follows from Theorem 3. □

We remark that the upper bound provided in Theorem 14 is probably far from the true upper bound. Actually, it was shown by Dey and Pach [6] that if a k-dimensional complex Σ embeds into $\mathbb{R}^k$ then $f_k(\Sigma) = \mathcal{O}(n^{\lceil \frac{k}{2} \rceil})$. Indeed, for $k \geq 4$ it is an open problem to show that if a simplicial complex embeds into $\mathbb{R}^k$, then the total number of its faces is bounded above by $\mathcal{O}(n^{\lceil \frac{k}{2} \rceil})$.

8 An Upper Bound by Grünbaum

In the 1970 paper [8] Branko Grünbaum shows that if a d-dimensional complex Σ embeds into $\mathbb{R}^{d+1}$, then $f_d(\Sigma) \leq \frac{6}{d+1} f_{d-1}(\Sigma)$. He also proves slightly sharper versions of this result for pure complexes, see Proposition 15 below.

How do the different bounds compare? Due to their different structure it is hard to make a general comparison. In view of having leading constant $\frac{6}{d+1}$, it is clear that Grünbaum's upper bound is better than ours in several cases, particularly when one has only some partial f-vector information. However, our bound is tighter in other cases, especially if much structural information, expressed in terms of f- and β-vectors, is available. In this section, we present one such case.

Let us begin with the following result, which extends the validity of Grünbaum's inequality [8, 5(iii)] to embeddability into manifolds.

Proposition 15 *Let Σ be a pure d-dimensional simplicial complex that is PL embeddable into a $(d+1)$-dimensional PL manifold. Then*

$$f_d(\Sigma) \leq \frac{6}{d+1} f_{d-1}(\Sigma) - \frac{10}{d(d+1)} f_{d-2}(\Sigma). \tag{3}$$

Proof We know that if Σ is a planar graph which contains at least one edge, then[1] $f_1(\Sigma) \leq 3f_0(\Sigma) - 5$. This verifies the first step $d = 1$ of an inductive argument.

Now assume that the statement is valid for every $1 \leq k < d$ and Σ is a pure d-dimensional simplicial complex that is PL embeddable into a $(d+1)$-dimensional PL manifold. Let V denote the vertex set of Σ and for $v \in V$, let L_v be the link of v in Σ. Since Σ is embeddable into a $(d+1)$-manifold, L_v must be embeddable into a d-sphere and we have

$$d!f_{d-1}(L_v) \leq 6(d-1)!f_{d-2}(L_v) - 10(d-2)!f_{d-3}(L_v).$$

Summing over all vertices $v \in V$ and using the equation $\sum_v f_i(L_v) = (i+2)f_{i+1}(\Sigma)$ yields the desired conclusion. □

Say we are interested in the question whether the d-skeleton of a $(d+1)$-manifold is embeddable into $\mathbb{R}^{d+1}$. If the manifold in question has non-vanishing homology in dimension d (or equivalently in dimension one) our inequalities turn out to be sharp enough to provide a negative answer, while Grünbaum's inequality (3) is not.

Proposition 16 *Let Σ be the d-skeleton of a triangulated $(d+1)$-dimensional PL manifold with non-zero d-dimensional Betti number. Then Σ is not PL-embeddable into $\mathbb{R}^{d+1}$.*

Proof Let M denote the $(d+1)$-dimensional manifold in question. We know from Lemma 2 that

$$f_{d+1}(M) = \beta_{d+1}(M;\mathbb{Z}_2) - \beta_d(M;\mathbb{Z}_2) + \beta_d(\Sigma;\mathbb{Z}_2).$$

[1] Note that "-5" is needed here, instead of "-6", in order to include the case when $f_0 = 2$ for the inductive argument.

Since M is a manifold, one has $(d+2)f_{d+1}(M) = 2f_d(M)$ and $\beta_{d+1}(M) = 1$. This, together with the assumption $\beta_d(M) \geq 1$ imply that

$$(d+2)\,(\beta_d(\Sigma)+1) = (d+2)\,(\beta_{d+1}(M) + \beta_d(\Sigma)) > (d+2)f_{d+1}(M) = 2f_d(M),$$

which violates the inequality (1) of Theorem 5. Therefore, Σ is not PL embeddable into $\mathbb{R}^{d+1}$. Also the inequality (2) is violated.

Observe, however, that Grünbaum's inequality (3) is satisfied by $f(\Sigma)$. This follows from Proposition 15, since Σ is PL embeddable into a $(d+1)$-dimensional manifold, namely M. □

Example 3 As a concrete example of this type, one may take T to be a triangulation of the 3-torus with $f(T) = (15, 105, 180, 90)$. Such a triangulation exists and happens to be the smallest (w.r.t. the f-vector) known triangulation of the 3-torus $S^1 \times S^1 \times S^1$. See [12, Table 7] for instance. Let Σ be the 2-skeleton of T. Then one has $f(\Sigma) = (15, 105, 180)$ and $\beta(\Sigma) = (1, 3, 92)$.

Acknowledgements We thank the referees for very useful remarks. Also, we thank Uli Wagner for valuable comments and for bringing Grünbaum's paper [8] to our attention.

This research was supported by the grant VR-2015-05308 from Vetenskapsrådet. Part of the research of the second author has been made possible by the grant KAW-stipendiet 2015.0360 from the Knut and Alice Wallenberg's Fondation.

References

1. N.L. Biggs, E. Keith Lloyd, R.J. Wilson, *Graph Theory. 1736–1936*, 2nd edn. (The Clarendon Press/Oxford University Press, New York, 1986)
2. A. Björner, G. Kalai, An extended Euler-Poincaré theorem. Acta Math. **161**(3–4), 279–303 (1988)
3. U. Brehm, A nonpolyhedral triangulated Möbius strip. Proc. Am. Math. Soc. **89**(3), 519–522 (1983)
4. J.W. Cannon, Shrinking cell-like decompositions of manifolds. Codimension three. Ann. Math. (2) **110**(1), 83–112 (1979)
5. T.K. Dey, H. Edelsbrunner, Counting triangle crossings and halving planes. Discret. Comput. Geom. **12**(3), 281–289 (1994). ACM Symposium on Computational Geometry, San Diego, 1993
6. T.K. Dey, J. Pach, Extremal problems for geometric hypergraphs. Discret. Comput. Geom. **19**(4), 473–484 (1998)
7. C. Greene, D.J. Kleitman, Proof techniques in the theory of finite sets, in *Studies in Combinatorics*. MAA Studies in Mathematics, vol. 17 (Mathematical Association of America, Washington, DC, 1978), pp. 22–79
8. B. Grünbaum, Higher-dimensional analogs of the four-color problem and some inequalities for simplicial complexes. J. Comb. Theory **8**, 147–153 (1970)
9. A. Gundert, Extermal combinatorics II: some geometry and number theory (2008), https://gilkalai.wordpress.com/2008/07/17/extermal-combinatorics-ii-some-geometry-and-number-theory/. Combinatorics and More
10. A. Gundert, On the complexity of embeddable simplicial complexes, Master's thesis, Freie Universität Berlin, 2009

11. S. Mac Lane, A structural characterization of planar combinatorial graphs. Duke Math. J. **3**(3), 460–472 (1937)
12. F.H. Lutz, T. Sulanke, E. Swartz, f-vectors of 3-manifolds. Electron. J. Comb. **16**(2), R13 (2009)
13. J. Matoušek, *Using the Borsuk-Ulam Theorem. Lectures on Topological Methods in Combinatorics and Geometry*. Universitext (Springer, Berlin, 2003)
14. J. Matoušek, M. Tancer, U. Wagner, Hardness of embedding simplicial complexes in $\mathbb{R}^d$. J. Eur. Math. Soc. (JEMS) **13**(2), 259–295 (2011)
15. J. Milnor, *Morse Theory*. Annals of Mathematics Studies AM-51 (Princeton University Press, Princeton, 1963)
16. J.R. Munkres, *Elements of Algebraic Topology* (Addison-Wesley Publishing Company, Menlo Park, 1984)
17. S. Parsa, *On links of vertices in simplicial d-complexes embeddable in Euclidean $2d$-space* (2015, preprint). arXiv:1512.05164
18. A. Shapiro, Obstructions to the imbedding of a complex in a Euclidean space. I. The first obstruction. Ann. Math. (2) **66**, 256–269 (1957)
19. D. Sullivan, René Thom's work on geometric homology and bordism. Bull. Am. Math. Soc. **41**, 341–350 (2004)
20. W.T. Tutte, Toward a theory of crossing numbers. J. Comb. Theory **8**, 45–53 (1970)
21. E.R. van Kampen, Komplexe in euklidischen Räumen. Abh. Math. Sem. Univ. Hamburg **9**(1), 72–78 (1933)
22. U. Wagner, Minors in random and expanding hypergraphs, in *Proceedings of the Twenty-Seventh Annual Symposium on Computational Geometry* (ACM, 2011)
23. H. Whitney, Non-separable and planar graphs. Trans. Am. Math. Soc. **34**(2), 339–362 (1932)
24. J. Zaks, On minimal complexes. Pac. J. Math. **28** 721–727 (1969)
25. G.M. Ziegler, *Lectures on Polytopes*. Graduate Texts in Mathematics, vol. 152 (Springer, New York, 1995)

Using Brouwer's Fixed Point Theorem

Anders Björner, Jiří Matoušek, and Günter M. Ziegler

Abstract Brouwer's fixed point theorem from 1911 is a basic result in topology—with a wealth of combinatorial and geometric consequences. In these lecture notes we present some of them, related to the game of HEX and to the piercing of multiple intervals. We also sketch stronger theorems, due to Oliver and others, and explain their applications to the fascinating (and still not fully solved) evasiveness problem.

1 Introduction

The fixed point theorem of Brouwer is one of the most widely known results of topology. It says that every continuous map $f : B^d \to B^d$ of the d-dimensional closed unit ball to itself has a fixed point, that is, a point $x_0 \in B^d$ such that $f(x_0) = x_0$.

This result was established by Luitzen Egbertus Jan Brouwer (1881–1960) at the end of his important 1911 paper [20], in which he also introduced the fundamental concept (and proof technique) of the mapping degree. It has many striking and famous applications to problems in Geometry, Analysis, Game Theory and Combinatorics.

Brouwer's fixed point theorem is in several ways similar to the Borsuk–Ulam theorem from 1933, which has gotten a lot of attention and appreciation for being unusually rich in applications. For example, the 1978 proofs of the 1956 Kneser conjecture by Lovász and by Bárány employed the Borsuk–Ulam Theorem in order to solve a problem about partitioning a set system, or equivalently, bounding the chromatic numbers for a certain class of graphs. This unexpected use of a result

A. Björner
Department of Mathematics, Royal Institute of Technology (KTH), 100 44 Stockholm, Sweden

J. Matoušek
Department of Applied Mathematics, Charles University, Malostranské nám. 25, 118 00 Prague, Czech Republic

Institute of Theoretical Computer Science, ETH Zurich, 8092 Zurich, Switzerland

G.M. Ziegler (✉)
Institute of Mathematics, Freie Universität Berlin, Arnimallee 2, 14195 Berlin, Germany
e-mail: ziegler@math.fu.berlin.de

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_10

from equivariant topology is one of the starting points (probably the most famous one) for the field of "Topological Combinatorics" [11, 46]. We refer to the detailed, elementary exposition in Matoušek's book "Using the Borsuk–Ulam Theorem" [52]. Current research continues this line of work, using more advanced methods from Equivariant Algebraic Combinatorics; see for example the text "Beyond the Borsuk–Ulam Theorem: The Topological Tverberg Story" [14] in this volume.

In various respects, Brouwer's theorem is a simpler result than the Borsuk–Ulam theorem: For example, it is very easy to state (as it does not involve symmetry, or a group action), and it is quite easy to prove (see below). It can also easily be derived from the Borsuk–Ulam theorem (see [73]), while indeed it is not as straightforward to obtain "Borsuk–Ulam from Brouwer."

Just like the Borsuk–Ulam theorem, Brouwer's theorem has many equivalent versions, as well as powerful and useful extensions. For instance, the Lefschetz fixed point theorem that works for spaces much more general than a ball, the Schauder fixed point theorem that works also for compact balls in infinite-dimensional Banach spaces, the Kakutani fixed point theorem for set-valued maps, *and so on*. See Shapiro [67] for a friendly introduction to fixed point theorems with Analysis applications in mind.

The striking applications of the Brouwer theorem in Combinatorics and Geometry seem not to be as well known as the applications of the Borsuk–Ulam theorem. In order to help to remedy this, we present three distinct areas of such applications in the three main sections of these lecture notes:

1. Brouwer's theorem can be invoked to prove that the game of HEX can never end without a winner. And indeed, the d-dimensional version of this claim turns out to be equivalent to Brouwer's theorem! This observation of David Gale in his award-winning 1979 paper [27] may also be counted among the starting points of Topological Combinatorics. In our presentation we not only use this to prove the HEX theorem, but we also give a combinatorial proof of the HEX theorem and derive Brouwer's theorem from this.
2. Some results about hypergraph matchings and transversals have a topological core, to be derived from the Brouwer theorem. Our presentation treats one striking instance, concerning the relation between packing and transversal numbers for systems of d-intervals.
3. The *Evasiveness conjecture* states that every non-trivial monotone graph property is evasive, that is, it does not allow for a query strategy that cannot be tricked into checking *all* potential edges of a graph in order to establish the property. This conjecture is still open in general, but the special case of a graph on a prime power number of vertices was proved using fixed point theorems of Smith and Oliver. These theorems may be seen as extensions of Brouwer's. The Appendix to this paper collects and sketches the necessary tools.

Further remarkable applications of Brouwer's fixed point theorem on geometric problems, not treated here, include the work by Bondarenko and Viazovska [17] on the construction of spherical designs, and the work on center points and regression depth by Amenta et al. [5].

Our presentation is based on lecture notes that were written about fifteen years ago, with a history that for some parts goes back nearly thirty years. These notes can be regarded as a companion or perhaps as a "prequel" to Matoušek's book [52].

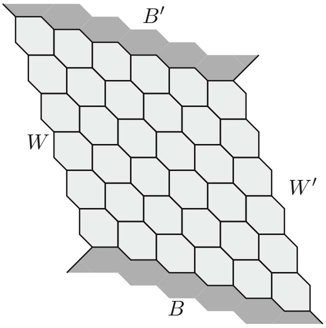
The three main parts do not depend on each other, so they can be read indepenently. We refer to [52] for notation and terminology not explained here.

2 A Game Model for Brouwer's Fixed Point Theorem

2.1 The Game of HEX

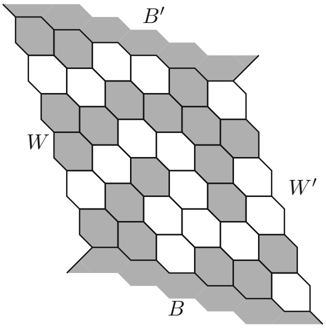
Let's start with a game: "HEX" is a board game for two players, invented by the ingenious Danish poet, designer and engineer Piet Hein in 1942 [29], and rediscovered in 1948 by the mathematician John Nash [57], who got a Nobel memorial prize in economics in 1994 (for his work on game theory, but not really for this game …).

HEX, in Hein's version, is played on a rhombical board, as depicted in the figure.



The rules of the game are simple: There are two players, whom we call White and Black. The players alternate, with White going first. Each move consists of coloring one "grey" hexagonal tile of the board white resp. black. White has to connect the white borders of the board (marked W and W') by a path of his white tiles, while Black tries to connect B and B' by a black path. They can't both win: Any winning path for white separates the two black borders, and conversely. (This isn't hard to prove—however, the statement is closely related to the Jordan curve theorem, which is trickier than it may seem when judged at first sight: see Exercise 13.)

However, here we concentrate on the opposite statement: There is no draw possible—when the whole board is covered by black and white tiles, then there always is a winner. (This is even true if one of the players has cheated badly and ends up with much more tiles than his/her opponent! It is also true if the board isn't really "square," that is, if it has sides of unequal lengths.) Our next figure depicts a final HEX position—sure enough one of the players has won, and the proof of the following "HEX theorem" will give us a systematic method to find out which one.



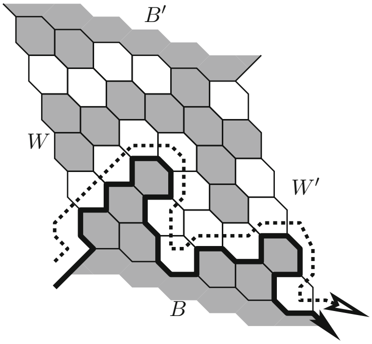
Theorem 2.1 (The HEX theorem) *If each tile of an* $(n \times m)$*-HEX board is colored black or white, then either there is a path of white tiles that connects the white borders W and* W'*, or there is a path of black tiles that connects the black borders B and* B'*.*

Our plan for this section is the following:

- We give a simple proof of the HEX theorem.
- We show that it implies the Brouwer fixed point theorem . . .
- . . . and conversely: The Brouwer fixed point theorem implies the HEX theorem.
- Then we prove that one of the players has a winning strategy.
- And then we see that on a square board, the first player can win, while on an uneven board, the player with the longer borders has a strategy to win.

All of this is really quite simple, but it nicely illustrates how a topological theorem enters the analysis of a discrete situation.

Proof of the HEX theorem For the proof we trace a certain path *between* the black and the white tiles. It starts in the lower left-hand corner of the HEX board on the edge that separates W and B. Whenever the path reaches a corner of degree 3, there will be both colors present at the corner (due to the edge we reach it from), and so there will be a unique edge to proceed on that does have different colors on its two sides.



Our path can never get stuck or branch or turn back onto itself, otherwise we would have found a vertex that has one or three edges that separate colors, whereas this number clearly has to be even at each vertex. Thus the path can be continued until it leaves the board—that is, until it reaches W' or B'. But that means that we find a path that connects W to W', or B to B', and on its sides keeps a white path of tiles resp. a black path. That is, one of White and Black has won! □

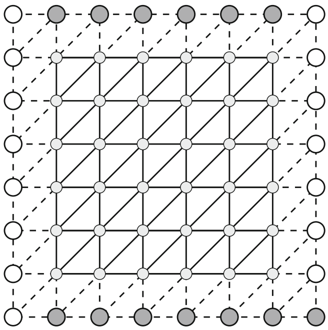
Now this was easy, and (hopefully) fun. We continue with a re-interpretation of the HEX board—in Nash's version—that buys us two drinks for the price of one:

(i) a d-dimensional version of the HEX theorem, and
(ii) the connection to the Brouwer fixed point theorem.

Definition 2.2 (The *d*-dimensional HEX board) The d-dimensional *HEX board* is the graph $H(n,d)$ on the vertex set $V = \{-1, 0, 1, \dots, n, n+1\}^d$, in which two vertices $\boldsymbol{v}, \boldsymbol{w} \in V$ are connected by an edge if and only if $\boldsymbol{v} - \boldsymbol{w} \in \{0,1\}^d \cup \{0,-1\}^d$.

The *colors* for the d-dimensional HEX game are $1, 2, \dots, d$, where we identify "1 $=$ white" and "2 $=$ black." The *interior* of the HEX board is given by $V' = \{0, 1, 2, \dots, n\}^d$. All the other vertices, in $V \setminus V'$, form the *boundary* of the board. The vertices in the boundary of $H(n,d)$ get preassigned colors

$$\kappa(\boldsymbol{v}) = \kappa(v_1, \dots, v_d) := \begin{cases} \min\{i : \; v_i = -1\} & \text{if this exists,} \\ \min\{i : \; v_i = n+1\} & \text{otherwise.} \end{cases}$$



Our drawing depicts the 2-dimensional HEX board $H(5,2)$, which represents a dual graph for the (6×6)-board that we used in our previous figures, with the preassigned colors on the boundary.

The d-dimensional HEX game is played between d players who take turns in coloring the interior vertices of $H(n,d)$. The i-th player *wins* if he[1] achieves a path of vertices of color i that connects a vertex whose i-th coordinate is -1 to a vertex whose i-th coordinate is $n+1$.

[1] Using "he" here is not politically correct.

Theorem 2.3 (The *d*-dimensional HEX theorem) *For d-dimensional HEX at least one of the players reaches his goal: When all interior vertices of $H(d,n)$ are colored, then at least one player has won.*

Proof The proof that we used for 2-dimensional HEX still works: It just has to be properly translated for the new setting. For this we first check that $H(n,d)$ is the graph of a triangulation $\Delta(n,d)$ of $[-1,n+1]^d$, which is given by the *clique complex* of $H(n,d)$. That is, a set of lattice points $S \subseteq \{-1,0,1,\dots,n+1\}^d$ forms a simplex in $\Delta(n,d)$ if and only if the points in S are pairwise connected by edges. To check this, verify that each point $x \in [-1,n+1]^d$ lies in the relative interior of a unique simplex, which is given by

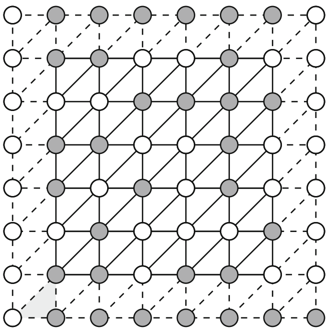
$$\begin{aligned}\Delta(\boldsymbol{x}) := \operatorname{conv}\big\{\boldsymbol{v} \in \{-1,\dots,n+1\}^d :\ & \\ & \lfloor x_i \rfloor \le v_i \le \lceil x_i \rceil \text{ for all } i, \\ & \lfloor x_i - x_j \rfloor \le v_i - v_j \le \lceil x_i - x_j \rceil \text{ for all } i \neq j\big\}.\end{aligned}$$

Every full-dimensional simplex in $\Delta(n,d)$ has $d+1$ vertices. A simplex S in $\Delta(n,d)$ is *completely colored* if it has all d colors on its vertices. Thus each completely colored d-simplex in Δ has exactly two completely colored facets, which are $(d-1)$-faces of the complex $\Delta(n,d)$. Conversely, every completely colored $(d-1)$-face is contained in exactly two completely colored d-simplices—if it is not on the boundary of $[-1,n+1]^d$.

With this the (constructive) proof that we gave before for the 2-dimensional HEX theorem generalizes to the following: We start at the d-simplex

$$\begin{aligned}\Delta_0 &:= \operatorname{conv}\{-\mathbf{1}, -\mathbf{1}+\boldsymbol{e}_1, -\mathbf{1}+\boldsymbol{e}_1+\boldsymbol{e}_2, \ \dots\ , -\mathbf{1}+\boldsymbol{e}_1+\dots+\boldsymbol{e}_{d-1}, -\mathbf{1}+\boldsymbol{e}_1+\dots+\boldsymbol{e}_d\} \\ &= \operatorname{conv}\{-\mathbf{1}, -\mathbf{1}+\boldsymbol{e}_1, -\mathbf{1}+\boldsymbol{e}_1+\boldsymbol{e}_2, \ \dots\ , -\boldsymbol{e}_d, \mathbf{0}\},\end{aligned}$$

whose facet ($(d-1)$-face) $\operatorname{conv}\{-\mathbf{1}, -\mathbf{1}+\boldsymbol{e}_1, \dots, -\boldsymbol{e}_{d-1}-\boldsymbol{e}_d, -\boldsymbol{e}_d\}$ is completely colored. (Verify this!) This simplex is shaded in the following figure for $H(5,2)$, which depicts the same final position that we considered before.

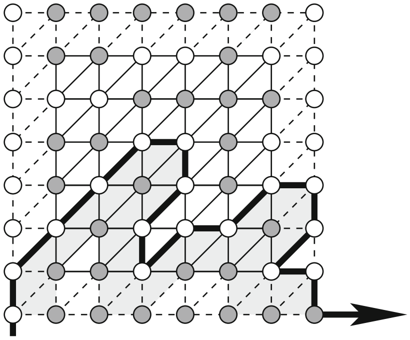


Now we construct a sequence of completely colored d-dimensional simplices that starts at Δ_0: We find the second completely colored $(d-1)$-face of Δ_0, find the second completely colored d-simplex it is contained in, etc. Thus we find a chain of completely colored d-simplices that ends on the boundary of $[-1, n+1]^d$—at a different simplex than the one we started from. In particular, the last d-simplex in the chain has a completely colored facet in the boundary, and by construction this facet has to lie in a hyperplane $H_i^+ = \{\boldsymbol{x} : x_i = n+1\}$. At this point we check that every completely colored $(d-1)$-simplex in the boundary of $H(n,d)$ is contained in one of the hyperplanes H_i^+, with the sole exception of the boundary facet of our starting d-simplex. The chain of d-simplices then provides us with an i-colored path from the i-colored vertex

$$-\boldsymbol{1} + \boldsymbol{e}_1 + \cdots + \boldsymbol{e}_{i-1} \in H_i^- = \{\boldsymbol{x} : x_i = -1\}$$

to the i-colored vertex in H_i^+: So the i-th player wins. □

Our drawing illustrates the chain of completely colored simplices (shaded) and the sequence of (white) vertices for the winning path that we get from it.

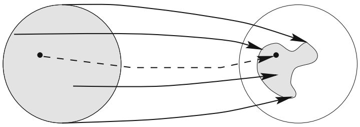


2.2 The Brouwer Fixed Point Theorem

Now we proceed from the discrete mathematics setting of the HEX game to the continuous world of topological fixed point theorems. Here are three versions of the Brouwer fixed point theorem.

Theorem 2.4 (Brouwer fixed point theorem) *The following are equivalent (and true):*

(Br1) *Every continuous map $f\colon B^d \longrightarrow B^d$ has a fixed point.*
(Br2) *Every continuous map $f\colon B^d \longrightarrow S^{d-1}$ has a fixed point.*
(Br3) *Every null-homotopic map $f\colon S^{d-1} \longrightarrow S^{d-1}$ has a fixed point.*

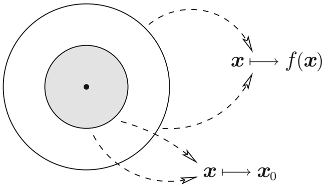


(The term *null-homotopic* that appears here refers to a map that can be deformed to a constant map.)

Proof of the equivalences (Br1)$\Longrightarrow$(Br2) is trivial, since $S^{d-1} \subseteq B^d$.

For (Br2)$\Longrightarrow$(Br3) let $h\colon\ S^{d-1} \times [0,1] \longrightarrow S^{d-1}$ be a null-homotopy for f, i.e., a continuous map that interpolates between our original map f and a constant map, with $h(\boldsymbol{x},0) = f(\boldsymbol{x})$ and $h(\boldsymbol{x},1) = \boldsymbol{x}_0$ for all $\boldsymbol{x} \in S^{d-1}$. From this we construct a continuous map $F\colon\ B^d \longrightarrow S^{d-1}$ that extends f, by

$$F(\boldsymbol{x}) := \begin{cases} h(\frac{\boldsymbol{x}}{|\boldsymbol{x}|}, 2 - 2|\boldsymbol{x}|) & \text{if } \frac{1}{2} \leq |\boldsymbol{x}| \leq 1, \\ \boldsymbol{x}_0 & \text{for } |\boldsymbol{x}| \leq \frac{1}{2}. \end{cases}$$



This map is continuous, and by (Br2) it has a fixed point, which must lie in the image, that is, in S^{d-1}.

For the converse, (Br3)$\Longrightarrow$(Br2), let $f\colon\ B^d \longrightarrow S^{d-1}$ be continuous. Then the restriction $f|_{S^{d-1}}$ is null-homotopic, since $h(\boldsymbol{x};t) := f((1-t)\boldsymbol{x})$ provides a null-homotopy. Thus, by (Br3) the map $f|_{S^{d-1}}$ has a fixed point, hence so does f.

Finally, we get (Br2)$\Longrightarrow$(Br1): If $f\colon\ B^d \longrightarrow B^d$ has no fixed point, then we set $g(\boldsymbol{x}) := \frac{f(\boldsymbol{x})-\boldsymbol{x}}{|f(\boldsymbol{x})-\boldsymbol{x}|}$. This defines a map $g\colon\ B^d \longrightarrow S^{d-1}$ that has a fixed point $\boldsymbol{x}_0 \in S^{d-1}$ by (Br2), with $\boldsymbol{x}_0 = \frac{f(\boldsymbol{x}_0)-\boldsymbol{x}_0}{|f(\boldsymbol{x}_0)-\boldsymbol{x}_0|}$. But this implies $f(\boldsymbol{x}_0) = \boldsymbol{x}_0(1+t)$ for $t := |f(\boldsymbol{x}_0) - \boldsymbol{x}_0| > 0$, and this is impossible for $\boldsymbol{x}_0 \in S^{d-1}$. □

In the following we use the unit cube $[0,1]^d$ in place of the ball B^d: It should be clear that the Brouwer fixed point theorem equally applies to self-maps of any domain D that is homeomorphic to the ball B^d, resp. of the boundary ∂D of such a domain.

Proof of the Brouwer fixed point theorem ("HEX $\Longrightarrow$ (Br1)"). If $f\colon\ [0,1]^d \longrightarrow [0,1]^d$ has no fixed point, then for some $\varepsilon > 0$ we have that $|f(\boldsymbol{x}) - \boldsymbol{x}|_\infty \geq \varepsilon$ for all $\boldsymbol{x} \in [0,1]^d$ (namely, one can take $\varepsilon := \min\{|f(\boldsymbol{x}) - \boldsymbol{x}|_\infty : \boldsymbol{x} \in [0,1]^d\}$, which exists since $[0,1]^d$ is compact).

Furthermore, any continuous function on the compact set $[0,1]^d$ is uniformly continuous (see e.g. Munkres [59, §27]), hence there exists some $\delta > 0$ such that

$|\boldsymbol{x}-\boldsymbol{x}'|_\infty < \delta$ implies $|f(\boldsymbol{x})-f(\boldsymbol{x}')|_\infty < \varepsilon$. We take $\delta < \varepsilon$ (without loss of generality), and then choose n with $\frac{1}{n} < \delta$.

From f, we now define a d-coloring of $H(n, d)$, by setting

$$\kappa(\boldsymbol{v}) := \min\{i : \ |f_i(\tfrac{\boldsymbol{v}}{n}) - \tfrac{v_i}{n}| \geq \varepsilon\}$$

for the interior vertices $\boldsymbol{v} \in H(n, d)$, where f_i denotes the ith component of f. This is well-defined, since $\frac{\boldsymbol{v}}{n} \in [0, 1]^d$, and thus the absolute value of at least one component of $f(\frac{\boldsymbol{v}}{n}) - \frac{\boldsymbol{v}}{n}$ has to be at least ε. Now, the d-dimensional HEX theorem guarantees a chain $\boldsymbol{v}^0, \boldsymbol{v}^1, \dots, \boldsymbol{v}^N$ of vertices of color i, for some i, where $v_i^0 = 0$ and $v_i^N = n$. Furthermore, we know that $|f_i(\frac{\boldsymbol{v}^k}{n}) - \frac{v_i^k}{n}| \geq \varepsilon$ for $0 \leq k \leq N$. Also, at the ends of the chain we know the signs:

$$f(\tfrac{\boldsymbol{v}^0}{n}) \in [0, 1]^d \text{ implies } f_i(\tfrac{\boldsymbol{v}^0}{n}) \geq 0 \text{ and hence } f_i(\tfrac{\boldsymbol{v}^0}{n}) - \tfrac{v_i^0}{n} \geq \varepsilon, \text{ and}$$
$$f(\tfrac{\boldsymbol{v}^N}{n}) \in [0, 1]^d \text{ implies } f_i(\tfrac{\boldsymbol{v}^N}{n}) \leq 1 \text{ and hence } f_i(\tfrac{\boldsymbol{v}^N}{n}) - \tfrac{v_i^N}{n} \leq -\varepsilon.$$

It follows that for some $k \in \{1, 2, \dots, N\}$ we must have a sign change:

$$f_i(\tfrac{\boldsymbol{v}^{k-1}}{n}) - \tfrac{v_i^{k-1}}{n} \geq \varepsilon \text{ and } f_i(\tfrac{\boldsymbol{v}^k}{n}) - \tfrac{v_i^k}{n} \leq -\varepsilon.$$

All these facts taken together provide a contradiction, since

$$|\tfrac{\boldsymbol{v}^{k-1}}{n} - \tfrac{\boldsymbol{v}^k}{n}|_\infty = \tfrac{1}{n} < \delta,$$

whereas

$$|f(\tfrac{\boldsymbol{v}^{k-1}}{n}) - f(\tfrac{\boldsymbol{v}^k}{n})|_\infty \geq |f_i(\tfrac{\boldsymbol{v}^{k-1}}{n}) - f_i(\tfrac{\boldsymbol{v}^k}{n})| \geq 2\varepsilon - |\tfrac{v_i^{k-1}}{n} - \tfrac{v_i^k}{n}| \geq 2\varepsilon - \tfrac{1}{n} > 2\varepsilon - \delta > \varepsilon.$$

□

Proof that the Brouwer fixed point theorem implies the HEX theorem ("Br1 $\Longrightarrow$ HEX"). Assume we have a coloring of $H(n, d)$. We use it to define a map $[0, n]^d \longrightarrow [0, n]^d$, as follows: On the points in $\{0, 1, \dots, n\}^d$ we define

$$f(\boldsymbol{v}) = \begin{cases} \boldsymbol{v} + \boldsymbol{e}_i & \text{if } \boldsymbol{v} \text{ has color } i \text{, and there is a path on vertices of color } i \\ & \text{that connects } \boldsymbol{v} \text{ to a vertex } \boldsymbol{w} \text{ with } w_i = 0 \\ \boldsymbol{v} - \boldsymbol{e}_i & \text{if } \boldsymbol{v} \text{ has color } i \text{, but there is no such path.} \end{cases}$$

If for the given coloring there is no winning path for HEX, then these definitions do not map any point $\boldsymbol{v}$ outside $[0, n]^d$. Hence this by linear extension defines a simplicial map $f\colon\ [0, n]^d \longrightarrow [0, n]^d$ on the simplices of the triangulation $\Delta(n, d)$ that we have considered before.

The following two observations now give us a contradiction, showing that this f cannot have a fixed point:

- If $\Delta = \operatorname{conv}\{\boldsymbol{v}^0, \boldsymbol{v}^1, \boldsymbol{v}^2, \dots, \boldsymbol{v}^d\} \subseteq \mathbb{R}^d$ is a simplex and $f\colon\ \Delta \longrightarrow \mathbb{R}^d$ is a linear map defined by $f(\boldsymbol{v}^i) = \boldsymbol{v}^i + \boldsymbol{w}^i$, then f has a fixed point on Δ if and only if $\boldsymbol{0} \in \operatorname{conv}\{\boldsymbol{w}^0, \dots, \boldsymbol{w}^d\}$.
- If $\boldsymbol{v}, \boldsymbol{v}'$ are adjacent vertices, then we cannot get $f(\boldsymbol{v}) = \boldsymbol{v} - \boldsymbol{e}_i$ and $f(\boldsymbol{v}') = \boldsymbol{v}' + \boldsymbol{e}_i$. Hence for each simplex of $\Delta(n, d)$, all the vectors $\boldsymbol{w}^i$ lie in one orthant of $\mathbb{R}^d$! □

2.3 The Joy of HEX: Who Wins?

So, who can win the 2-dimensional HEX game? A simple but ingenious argument due to John Nash, known as “stealing a strategy,” shows that on a square board the first player (“White”) always has a winning strategy. In the following we first define winning strategies, then show that one of the players has one, and finally conclude that the first player has one. Still: The proof will be non-constructive, and we don’t know how to win HEX. So, the game still remains interesting ...

Definition 2.5 A *strategy* is a set of rules that tells one of the players which move to choose (i.e., which tile to color) for every legal position on the board. A *winning strategy* here guarantees to lead to a win, starting from an empty board, for all possible moves of the opponent.

A *position* of the HEX game is a board on which some tiles may have been colored white or black, together with the information who moves next (unless all tiles are colored). A position is *legal* if it can occur in a HEX game: That is, if either White moves next, and the numbers of white and black tiles agree, or if Black moves next, and White has one more tile.

A *winning position for White* is a position such that White has a *winning strategy* that tells him how to proceed (for arbitrary moves of Black) and guarantees a win. Similarly, *a winning position for Black* has a *winning strategy* that guarantees to lead Black to a win.

Lemma 2.6 *Every (legal) position for HEX is either a winning position for White or a winning position for Black.*

Proof Here we proceed by induction on the number g of “grey” tiles (i.e., “free” positions on the board). If no grey tiles are present ($g = 0$), then one of the players has won—by the HEX theorem.

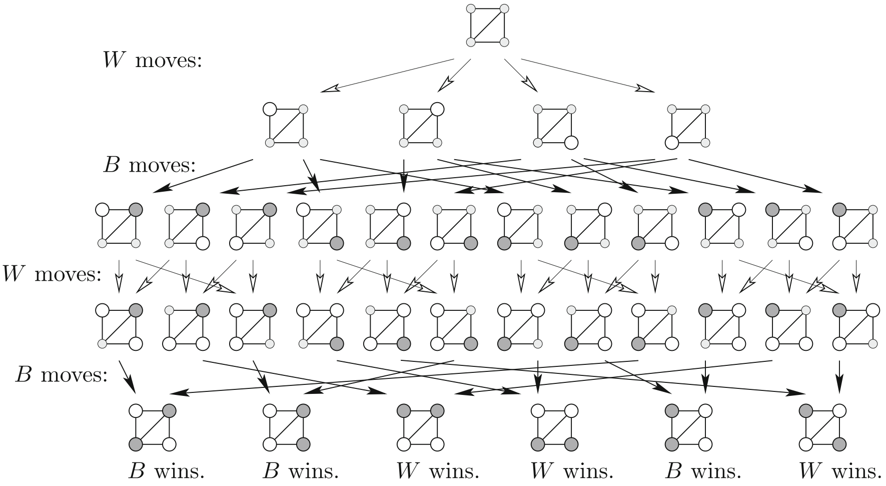
If $g > 0$ and White is to move, then any move that White could choose reduces g, and thus (by induction) produces a winning position for one of the players. If there is a move that leads to a winning position for White, then this is really nice and great for White: This makes the present position into a winning position for White, and any such move can be used for a winning position for White. Otherwise—too bad: If every possible move for White produces a winning position for Black, then we are at a winning position for Black already.

And the same argument applies for $g > 0$ if Black is to move. □

Of course, the argument given here is *much* more general: Essentially we have proved that for any finite deterministic 2-person game without a draw and with “complete information” there is a winning strategy for one of the players. (This is a theorem of Zermelo, which was rediscovered by von Neumann and Morgenstern). Furthermore, for games where a draw is possible either one player has a winning strategy, or *both* players can force a draw. We refer to Exercise 12, and to Blackwell and Girshick [13, p. 21].

For HEX, Lemma 2.6 shows that at the beginning (for the starting position, where all tiles are grey, and White is to move), there is a winning strategy either for White or for Black. But who is the winner?

Our first attempt might be to follow the proof of Lemma 2.6. Only for the 2×2 board this can be done:



In this drawing, you can decide for every position whether it is a winning position for White or for Black, starting with the bottom row ($g = 0$) that has three winning positions for each player, ending at the top node ($g = 4$), which turns out to be a winning position for White.

For larger boards, this approach is hopeless—after all, there are $\binom{n^2}{\lfloor n^2/2 \rfloor}$ final positions to classify for "$g = 0$," and from this one would have to work one's way up to the top node of a huge tree (of height n^2). Nevertheless, people have worked out winning strategies for White on the $n \times n$ boards for $n \leq 5$ (see Gardner [28]).

Theorem 2.7 *For the HEX game played on a HEX board with equal side lengths, White (the first player) has a winning strategy.*

Proof Assume not. Then by Lemma 2.6 Black has a winning strategy. But then White can start with an arbitrary move, and then—using the symmetry of the board and of the rules—just ignore his first tile, and follow Black's winning strategy "for the second player." This strategy will tell White always which move to take. Here the "extra" white tiles cannot hurt White: If the move for White asks to occupy a tile that is already white, then an arbitrary move is fine for White. But this "stealing a strategy" argument produces a winning strategy for White, contradicting our assumption! □

Notes Gale's beautiful paper [27] was the source and inspiration for our treatment of Brouwer's fixed point theorem in terms of the HEX game. Nash's analysis for the winning strategies for HEX is from Gardner's classical account in [28], some of which reappears in Milnor's [57]. See also the accounts in Jensen and Toft [37, Sect. 17.14], and in Berlekamp, Conway and Guy [9, p. 680], where other cases of "strategy stealing" are discussed. (A theoretical set-up for this is in Hales and Jewett [33, Sect. 3].)

The traditional combinatorial approach to the Brouwer fixed point theorem is via Sperner's lemma [71]; see e.g. Exercise 4 below and the presentation in [1]. Lovász's [48] matroid version of Sperner's lemma in Exercise 5 was further generalized by Lindström [45]. Kryński [44], however, showed that these results can easily be derived from earlier results.

A more geometric version of the combinatorial lemmas is given by Mani [50].

Exercises

1. Stir your coffee cup. Show that the (moving, but flat) surface has at every moment at least one point that stands still (has velocity zero).
2. Prove that if you tear a sheet of paper from your notebook, crumble it into a small ball, and put that down on your notebook, then at least one point of the sheet comes to rest exactly on top of its original position.

   Could it happen that there are exactly two such points?
3. In the proof of the Brouwer fixed point theorem (Theorem 2.4, (Br2)$\Longrightarrow$(Br3)), we could have tried to simply put $F(\boldsymbol{x}) := h(\frac{\boldsymbol{x}}{|\boldsymbol{x}|}, 1 - |\boldsymbol{x}|)$. Is this continuous?
4. (a) Prove "Sperner's Lemma" [71]: Let Δ be a triangulation of the d-dimensional sphere and let us color the vertices of Δ using $d + 1$ colors. Then Δ has an even number of colorful facets (meaning d-faces containing vertices of all colors).

   (b) Show that Sperner's Lemma implies the Brouwer fixed point theorem.
5. (a) Let Δ be a triangulation of a d-dimensional manifold with vertex set V. Assume that a matroid M of rank $d + 1$ without loops is defined on V. If Δ has a facet that is a basis of M then it has at least two such facets. (Lovász [48])

   (b) Show that part (a) implies Sperner's Lemma, and hence also Brouwer's theorem.
6. Let $B_E = 2^E \setminus \{\emptyset, E\}$ be the poset of all proper subsets of a finite set E, ordered by containment. Show that if an order-preserving map $f\colon\ B_E \to B_E$ does not have a fixed point then it is surjective, and hence an automorphism.
7. Let $P = B_E \setminus \{A\}$, for some proper subset A.

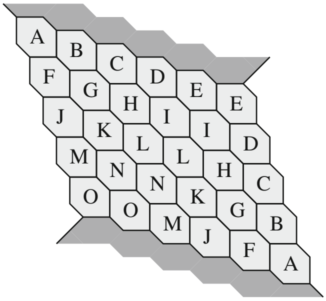
   (a) Give a quick proof that P has the *fixed point property*, meaning that any order-preserving self-map has a fixed point.

   (b) Give a slow proof, not using topology, that P has the fixed point property.

8. For HEX on a 3 × 3 board, how large is the tree of possible positions?
9. Can you write a computer program that plays HEX and wins (sometimes) [22]?
10. For d-dimensional HEX, is there always some "short" winning path? Show that for every $d \geq 2$ there is a constant c_d such that for all n there is a final configuration such that only one player wins, but his shortest path uses more than $c_d \cdot n^d$ tiles.
11. Construct an algorithm that, for given $\varepsilon > 0$ and $f\colon\ [0,1]^2 \longrightarrow [0,1]^2$, calculates a point $x_0 \in [0,1]^2$ with $|f(x_0) - x_0| < \varepsilon$. [27, p. 827]
12. If in a complete information two player game a draw is possible, argue why either one of the players has a winning strategy, or *both* can force at least a draw.
13. Prove that for 2-dimensional HEX, not both players can win! For this, prove and use the "polygonal Jordan curve theorem": any simple closed polygon in the plane uniquely divides the plane into an "inside" region and an "outside" region.

    (The general Jordan curve theorem for simple "Jordan arcs" in the plane has extensive discussions in many books; see for example Munkres [59], Stillwell [72, Sect. 0.3], or Thomassen [75].)
14. On an $(m \times n)$-board that is not square (that is, $m \neq n$), the player who gets the longer sides, and hence the shorter distance to bridge by a winning path, has a winning strategy. Our figure illustrates the case of a (6×5)-board, where the claim is that Black has a winning strategy.

    (i) Show that for this, it is sufficient to consider the case where $m = n + 1$ (i.e., the second player Black, who gets the longer side, has a sure win).



    (ii) Show that in the situation of (i), Black has the following winning strategy. Label the tiles in the "symmetric" way that is indicated by the figure, such that there are two tiles of each label. The strategy for Black is to always take the second tile that has the same label as the one taken by White. Why will this strategy always win for Black? (Hint: You will need the Jordan curve theorem.)

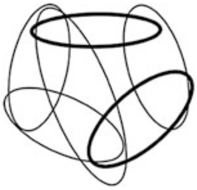
    (This is in Gardner [28] and in Milnor [57], but neither source gives the proof. You'll have to work yourself!)

3 Piercing Multiple Intervals

3.1 Packing Number and Transversal Number

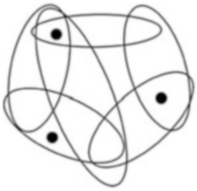
Let $\mathcal{S}$ be a system of subsets of a ground set X; both $\mathcal{S}$ and X may generally be infinite. The *packing number* of $\mathcal{S}$, usually denoted by $\nu(\mathcal{S})$ and often also called the *matching number*, is the maximum cardinality of a system of pairwise disjoint sets in $\mathcal{S}$:

$$\nu(\mathcal{S}) = \sup\{|\mathcal{M}| : \mathcal{M} \subseteq \mathcal{S},\ M_1 \cap M_2 = \varnothing \text{ for all } M_1, M_2 \in \mathcal{M}, M_1 \neq M_2\}.$$



The *transversal number* or *piercing number* of $\mathcal{S}$ is the smallest number of points of X that capture all the sets in $\mathcal{S}$:

$$\tau(\mathcal{S}) = \min\{|T| : T \subseteq X,\ S \cap T \neq \varnothing \text{ for all } S \in \mathcal{S}\}.$$



A subsystem $\mathcal{M} \subseteq \mathcal{S}$ of pairwise disjoint sets is usually called a *matching* (this refers to the graph-theoretical matching, which is a system of pairwise disjoint edges), and a set $T \subseteq X$ intersecting all sets of $\mathcal{S}$ is referred to as a *transversal* of $\mathcal{S}$. Clearly, any transversal is at least as large as any matching, and so always

$$\nu(\mathcal{S}) \leq \tau(\mathcal{S}).$$

In the reverse direction, very little can be said in general, since $\tau(\mathcal{S})$ can be arbitrarily large even if $\nu(\mathcal{S}) = 1$. As a simple geometric example, we can take the plane as the ground set of $\mathcal{S}$ and let the sets of $\mathcal{S}$ be lines in general position. Then $\nu = 1$, since every two lines intersect, but $\tau \geq \frac{1}{2}|\mathcal{S}|$, because no point is contained in more than two of the lines.

One of the basic general questions in combinatorics asks for interesting special classes of set systems where the transversal number can be bounded in terms of the matching number.[2] Many such examples come from geometry. Here we restrict our attention to one particular type of systems, the *d-intervals*, where the best results have been obtained by topological methods.

Fractional packing and transversal numbers Before introducing d-intervals, we mention another important parameter of a set system, which always lies between ν and τ and often provides useful estimates for ν or τ. This parameter can be introduced in two seemingly different ways. For simplicity, we restrict ourselves to finite set systems (on possibly infinite ground sets). A *fractional packing* for a finite set system $\mathcal{S}$ on a ground set X is a function $w\colon\ \mathcal{S} \longrightarrow [0,1]$ such that for each $x \in X$, we have $\sum_{S\in\mathcal{S}:\,x\in S} w(S) \leq 1$. The *size* of a fractional packing w is $\sum_{S\in\mathcal{S}} w(S)$, and the *fractional packing number* $\nu^*(\mathcal{S})$ is the supremum of the sizes of all fractional packings for $\mathcal{S}$. So in a fractional packing, we can take, say, one-third of one set and two-thirds of another, but at each point, the fractions for the sets containing that point must add up to at most 1. We always have $\nu(\mathcal{S}) \leq \nu^*(\mathcal{S})$, since a packing $\mathcal{M}$ defines a fractional packing w by setting $w(S) = 1$ for $S \in \mathcal{M}$ and $w(S) = 0$ otherwise.

Similar to the fractional packing, one can also introduce a fractional version of a transversal. A *fractional transversal* for a (finite) set system $\mathcal{S}$ on a ground set X is a function $\varphi\colon\ X \longrightarrow [0,1]$ attaining only finitely many nonzero values such that for each $S \in \mathcal{S}$, we have $\sum_{x\in S} \varphi(x) \geq 1$. The size of a fractional transversal φ is $\sum_{x\in X} \varphi(x)$, and the *fractional transversal number* $\tau^*(\mathcal{S})$ is the infimum of the sizes of fractional transversals.

By the duality theorem of linear programming (or by the theorem about separation of disjoint convex sets by a hyperplane), it follows that $\nu^*(\mathcal{S}) = \tau^*(\mathcal{S})$ and thus that

$$\nu(\mathcal{S}) \leq \nu^*(\mathcal{S}) = \tau^*(\mathcal{S}) \leq \tau(\mathcal{S})$$

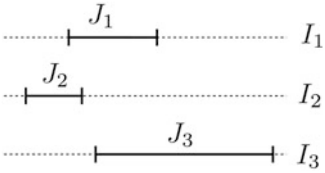
for any finite set system $\mathcal{S}$.

When trying to bound τ in terms of ν, in many instances it proved very useful to bound ν^* as a function of ν first, and then τ in terms of τ^*. The proof presented below follows a somewhat similar approach.

[2] This kind of problem is certainly not restricted to combinatorics. For example, if $\mathcal{S}$ is the system of all open sets in a topological space, $\tau(\mathcal{S})$ is the minimum size of a dense set and is called the *density*, while $\nu(\mathcal{S})$ is known as the *Souslin number* or *cellularity* of the space. In 1920, Souslin asked whether a linearly ordered topological space exists (the open sets are unions of open intervals) with countable ν but uncountable τ. It turned out in the 1970s that the answer depends on the axioms one is willing to assume beyond the usual (ZFC) axioms of set theory. For example, it is yes if one assumes the continuum hypothesis; see e.g. [23].

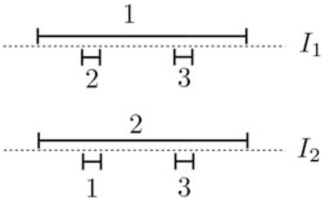
3.2 The d-Intervals

Let $I_1, I_2, \dots, I_d$ be disjoint parallel segments in the plane. (We may assume without loss of generality that they are horizontal unit length intervals at distinct heights/y-coordinates.) A set $J \subset \bigcup_{i=1}^{d} I_i$ is a *d-interval* if it intersects each I_i in a closed interval. We denote this intersection by J_i and call it the *ith component* of J. The following drawing shows a 3-interval:



Intersection and piercing for d-intervals are taken in the set-theoretical sense: Two d-intervals intersect if, for some i, their ith components intersect.

The 1-intervals, which are just intervals in the usual sense, behave nicely with respect to packing and piercing, as for any family $\mathcal{F}$ of intervals, we have $\nu(\mathcal{F}) = \tau(\mathcal{F})$. (This is well-known and easy to prove: Exercise 1!) This, however, does not extend to d-intervals. For example, the family $\mathcal{F}$ of three 2-intervals



has $\nu(\mathcal{F}) = 1$ while $\tau(\mathcal{F}) = 2$. By taking multiple copies of this family, one obtains families with $\tau = 2\nu$ for all values of ν.

Gyárfás and Lehel [31] showed by elementary methods that for any d and any family $\mathcal{F}$ of d-intervals, $\tau(\mathcal{F})$ can be bounded by a function of $\nu(\mathcal{F})$ (also see [32]). Their function was rather large (about $\nu^{d!}$ for d fixed). After an initial breakthrough by Tardos [74], who proved $\tau(\mathcal{F}) \leq 2\nu(\mathcal{F})$ for any family of 2-intervals, Kaiser [39] obtained the following result:

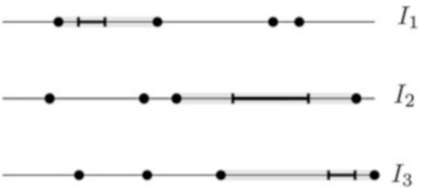
Theorem 3.1 (The Tardos–Kaiser theorem on *d*-intervals) *Every family $\mathcal{F}$ of d-intervals, $d \geq 2$, has a transversal of size at most $(d^2 - d) \cdot \nu(\mathcal{F})$.*

Here we present a proof using the Brouwer fixed point theorem. Alon [2] found a short non-topological proof of the slightly weaker bound $\tau(\mathcal{F}) \leq 2d^2\nu(\mathcal{F})$.

Proof Let $\mathcal{F}$ be a fixed system of d-intervals with $\nu(\mathcal{F}) = k$, and let $t = t(d, k)$ be a suitable (yet undetermined) integer. The general plan of the proof is this: Assuming that there is no transversal of $\mathcal{F}$ of size dt, we show by a topological method that the fractional packing number $\nu^*(\mathcal{F})$ is at least $t + 1$. Then a simple combinatorial argument proves that the packing number $\nu(\mathcal{F})$ is at least $\frac{t+1}{d}$, which leads to $t < d^2 \cdot \nu(\mathcal{F})$. Sharper combinatorial reasoning in this step leads to the slightly better bound in the theorem.

Our candidates for a transversal of $\mathcal{F}$ are all sets T with each $T_i = T \cap I_i$ having exactly t points; so $|T| = td$. For technical reasons, we also permit that some of the t points in I_i coincide, so T can be a multiset.

The letter T could also abbreviate a *trap*. The trap is set to catch all the d-intervals in $\mathcal{F}$, but if it is not set well enough, some of the d-intervals can escape. Each of them escapes through a hole in the trap, namely through a *d-hole*. The points of T_i cut the segment I_i into $t + 1$ open intervals (some of them may be empty), and these are the *holes in I_i*; they are numbered 1 through $t+1$ from left to right. A d-hole consists of d holes, one in each I_i. The *type* of a d-hole H is the set $\{(1,j_1), (2,j_2), \ldots, (d,j_d)\}$, where $j_i \in [t+1]$ is the number of the hole in I_i contained in H. A d-interval $J \in \mathcal{F}$ *escapes* through a d-hole H if it is contained in the union of its holes. The drawing shows a 3-hole, of type $\{(1, 2), (2, 4), (3, 4)\}$, and a 3-interval escaping through it:



Let $\mathcal{H}_0$ be the hypergraph with vertex set $[d] \times [t+1]$ and with edges being all possible types of d-holes; for example, the hole in the picture yields the edge $\{(1, 2), (2, 4), (3, 4)\}$. So $\mathcal{H}_0$ is a complete d-partite d-uniform hypergraph. By saying that a $J \in \mathcal{F}$ escapes through an edge H of $\mathcal{H}_0$, we mean that J escapes through the d-hole (uniquely) corresponding to H.

Next, we define weights on the edges of $\mathcal{H}_0$; these weights depend on the set T (and also on $\mathcal{F}$, but this is considered fixed). The weight of an edge $H \in \mathcal{H}_0$ is

$$q_H = \sup\{\text{dist}(J, T) : J \in \mathcal{F}, J \text{ escapes through } H\}.$$

Here $\text{dist}(J, T) := \min_{1 \le i \le d}\{\text{dist}(J_i, T_i)\}$ and $\text{dist}(J_i, T_i)$ is the distance of the ith component of J to the closest point of T_i. Thus q_H can be interpreted as the largest margin by which some d-interval from $\mathcal{F}$ escapes through H. If no members of $\mathcal{F}$ escape through H, we define q_H as 0. Note that this is the only case where $q_H = 0$. Otherwise, if anything escapes, it does so by a positive margin, since we are dealing with closed intervals.

From the edge weights, we derive weights of vertices: The weight w_v of a vertex $v = (i,j)$ is the sum of the weights of the edges of $\mathcal{H}_0$ containing v. These weights, too, are functions of T; to emphasize this, we write $w_v = w_v(T)$.

Lemma 3.2 *For any $d \geq 1$, $t \geq 1$, and any $\mathcal{F}$, there is a choice of T such that all the vertex weights $w_v(T)$, $v \in [d] \times [t{+}1]$, coincide.*

It is this lemma whose proof is topological. We postpone that proof and finish the combinatorial part first.

Let us suppose that a trap T was chosen as in the lemma, with $w_v(T) = W$ for all v. If $W = 0$ then T is a transversal, since all edge weights are 0 and no $J \in \mathcal{F}$ escapes. So suppose that $W > 0$.

Let $\mathcal{H} = \mathcal{H}(T) \subseteq \mathcal{H}_0$, the *escape hypergraph* of T, consist of the edges of $\mathcal{H}_0$ with nonzero weights. Note that

$$\nu(\mathcal{H}) \leq \nu(\mathcal{F}). \tag{1}$$

Indeed, given a matching $\mathcal{M}$ in $\mathcal{H}$, for each edge $H \in \mathcal{M}$ choose a $J \in \mathcal{F}$ escaping through H—this gives a matching in $\mathcal{F}$.

We note that the re-normalized edge weights $\tilde{q}_H = \frac{1}{W}\, q_H$ determine a fractional packing in $\mathcal{H}$ (since the weights at each vertex sum up to 1). For the size of this fractional packing, which is the total weight of all vertices, we find by double counting

$$\sum_{H\in\mathcal{H}} \tilde{q}_H = \frac{1}{d}\sum_{H\in\mathcal{H}}\sum_{v\in H}\tilde{q}_H = \frac{1}{d}\sum_{v\in[d]\times[t+1]}\frac{w_v}{W} = \frac{1}{d}\sum_{v} 1 = t+1.$$

As $\nu^*(\mathcal{H})$ is the supremum of the weights of all fractional packings, and $\tilde{q}_H$ is a particular fractional packing, this yields $\nu^*(\mathcal{H}) \geq \sum_{H\in\mathcal{H}}\tilde{q}_H = t+1$.

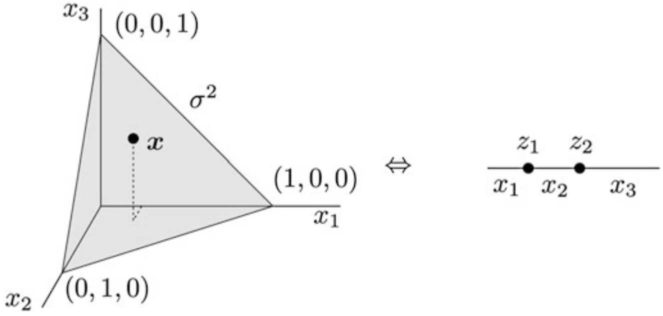
The last step is to show that $\nu(\mathcal{H})$ cannot be small if $\nu^*(\mathcal{H})$ is large. Here is a simple argument leading to a slightly suboptimal bound, namely $\nu(\mathcal{H}) \geq \frac{1}{d}\,\nu^*(\mathcal{H})$.

Given a fractional matching $\tilde{q}$ of size $t+1$ in $\mathcal{H}$, a matching can be obtained by the following greedy procedure: Pick an edge H_1 and discard all edges intersecting it, pick H_2 among the remaining edges, etc., until all edges are exhausted. The $\tilde{q}$-weight of H_i plus all the edges discarded with it is at most $d = |H_i|$, while all edges together have weight $t+1$. Thus, the number of steps, and also the size of the matching $\{H_1, H_2, \ldots\}$, is at least $\lceil\frac{t+1}{d}\rceil$.

If we set $t = d \cdot \nu(\mathcal{F})$, we get $\nu(\mathcal{H}) > \nu(\mathcal{F})$, which contradicts (1). Therefore, for this choice of t, all the vertex weights must be 0, and T as in Lemma 3.2 is a transversal of $\mathcal{F}$ of size at most $d^2\nu(\mathcal{F})$.

The improved bound $\tau(\mathcal{F}) \leq (d^2 - d)\cdot\nu(\mathcal{F})$ for $d \geq 3$ follows similarly using a theorem of Füredi [26], which implies that the matching number of any d-uniform d-partite hypergraph $\mathcal{H}$ satisfies $\nu^*(\mathcal{H}) \leq (d-1)\nu(\mathcal{H})$. (For $d = 2$, a separate argument needs to be used, based on a theorem of Lovász stating that $\nu^*(G) \leq \frac{3}{2}\nu(G)$ for all graphs G.) The Tardos–Kaiser Theorem 3.1 is proved. □

Proof of Lemma 3.2 Let σ^t denote the standard t-dimensional simplex in $\mathbb{R}^{t+1}$, i.e. the set $\{\boldsymbol{x} \in \mathbb{R}^{t+1} : x_j \geq 0, x_1 + \cdots + x_{t+1} = 1\}$. A point $\boldsymbol{x} \in \sigma^t$ defines a t-point multiset $\{z_1, z_2, \ldots, z_t\} \subset [0, 1]$, $z_1 \leq z_2 \leq \cdots \leq z_t$, by setting $z_k = \sum_{j=1}^{k} x_j$. Here is a picture for $t = 2$:



A candidate transversal T with t points in each I_i can thus be defined by an ordered d-tuple $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)$ of points, $\boldsymbol{x}_i \in \sigma^t$, where $\boldsymbol{x}_i$ determines T_i. Such an ordered d-tuple can be regarded as a single point $\boldsymbol{x}$ in the Cartesian product $P = \sigma^t \times \sigma^t \times \cdots \times \sigma^t = (\sigma^t)^d$. To each $\boldsymbol{x} \in P$, we have thus assigned a candidate transversal $T(\boldsymbol{x})$.

For each vertex $v = (i, j)$ of the hypergraph $\mathcal{H}_0$, we define the function g_{ij}: $P \to \mathbb{R}$ by $g_{ij}(\boldsymbol{x}) = w_{(i,j)}(T(\boldsymbol{x}))$, where $w_v(T)$ is the vertex weight. This is a *continuous* function of $\boldsymbol{x}$, since the edge weights q_H and hence the vertex weights $w_{(i,j)}(T(\boldsymbol{x}))$ change continuously when $T(\boldsymbol{x})$ moves—even if by this move new edges from $\mathcal{F}$ escape, or fail to escape, through a hole: If this is due to a small change of $T(\boldsymbol{x})$, then they escape, or fail to escape, by a narrow margin.

We note that for each $\boldsymbol{x}$, the sum

$$S_i(\boldsymbol{x}) = \sum_{j=1}^{t+1} g_{ij}(\boldsymbol{x})$$

is independent of i; this is because $S_i(\boldsymbol{x})$ equals the sum of the weights of all edges. So we can write just $S(\boldsymbol{x})$ instead of $S_i(\boldsymbol{x})$.

If there is an $\boldsymbol{x} \in P$ with $S(\boldsymbol{x}) = 0$, then all the vertex weights $w_{(i,j)}(T(\boldsymbol{x}))$ are 0 and we are done. Otherwise, we define the normalized functions

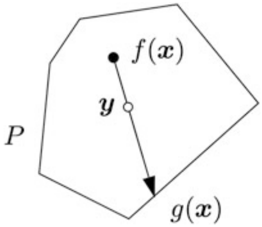
$$f_{ij}(\boldsymbol{x}) = \frac{1}{S(\boldsymbol{x})}\, g_{ij}(\boldsymbol{x}).$$

For each $i, f_{i1}(\boldsymbol{x}), \ldots, f_{i(t+1)}(\boldsymbol{x})$ are nonnegative and sum up to 1, and so they are the coordinates of a point in the standard simplex σ^t. All the maps f_{ij} together can be regarded as a map f: $P \to P$. To prove the lemma, we need to show that the image of f contains the point of P with all the $d(t + 1)$ coordinates equal to $\frac{1}{t+1}$.

The product P is a convex polytope, and its nonempty faces are exactly all Cartesian products $F_1 \times F_2 \times \cdots \times F_d$, where the $F_1, \ldots, F_d$ are nonempty faces of the factors $\sigma^t, \ldots, \sigma^t$ of $P = \sigma^t \times \sigma^t \times \cdots \times \sigma^t$ (Exercise 2). We note that for any face F of P, we have $f(F) \subseteq F$: Indeed, any face G of σ^t has the form $G = \{\boldsymbol{x} \in \sigma^t : x_i = 0 \text{ for all } i \in I\}$, for some index set I, and the faces of P are products of faces G of this form. So it suffices to know that $f_{ij}(\boldsymbol{x}) = 0$ whenever $(\boldsymbol{x}_i)_j = 0$. This holds, since $(\boldsymbol{x}_i)_j = 0$ means that the jth hole in I_i is empty, so nothing can escape through that hole, and thus $f_{ij}(\boldsymbol{x}) = 0$. The proof of Lemma 3.2 is now reduced to the following statement.

Lemma 3.3 *Let P be a convex polytope and let $f\colon P \to P$ be a continuous map satisfying $f(F) \subseteq F$ for each face*[3] *F of P. Then f is surjective.*

Proof Since the condition is hereditary for faces, it suffices to show that each point $\boldsymbol{y}$ in the interior of P has a preimage. For contradiction, suppose that some $\boldsymbol{y} \in \operatorname{int} P$ is not in the image of f. For $\boldsymbol{x} \in P$, consider the ray that starts at $f(\boldsymbol{x})$ and passes through $\boldsymbol{y}$, and let $g(\boldsymbol{x})$ be the unique intersection of that ray with the boundary of P.



This g is a well-defined and continuous map $P \to P$, and by Brouwer's fixed point theorem, there is an $\boldsymbol{x}_0 \in P$ with $g(\boldsymbol{x}_0) = \boldsymbol{x}_0$. The point $\boldsymbol{x}_0$ lies on the boundary of P, in some proper face F. But $f(\boldsymbol{x}_0)$ cannot lie in F, because the segment $\boldsymbol{x}_0 f(\boldsymbol{x}_0)$ passes through the point $\boldsymbol{y}$ outside F—a contradiction. □

3.3 Lower Bounds

It turns out that the bound in Theorem 3.1 is not far from being the best possible. In particular, for $\nu(\mathcal{F}) = 1$ and d large, the transversal number can be near-quadratic in d, which is rather surprising. For all k and d, systems $\mathcal{F}$ of d-intervals can be constructed with $\nu(\mathcal{F}) = k$ and

$$\tau(\mathcal{F}) \geq c \frac{d^2}{(\log d)^2} k$$

[3] In fact, it suffices to require $f(F) \subseteq F$ for each facet of P (that is, for each face of dimension $\dim(P) - 1$), since each face is the intersection of some facets.

for a suitable constant $c > 0$ (Matoušek [51]). The construction involves an extension of a construction due to Sgall [66] of certain systems of set pairs. Here we outline a (non-topological!) proof of a somewhat simpler result concerning families of *homogeneous* d-intervals, which are unions of at most d closed intervals on the real line. These are more general than the d-intervals, but an upper bound only slightly weaker than Theorem 3.1 can be proved for them along the same lines (Exercise 4): $\tau \leq (d^2 - d + 1)\nu$.

Proposition 3.4 *There is a constant $c > 0$ such that for every $d \geq 2$ and $k \geq 1$, there exists a system $\mathcal{F}$ of homogeneous d-intervals with $\nu(\mathcal{F}) = k$ and*

$$\tau(\mathcal{F}) \geq c\,\frac{d^2}{\log d}\,k.$$

Proof Given d and k, we want to construct a system $\mathcal{F}$ of homogeneous d-intervals. Clearly, it suffices to consider the case $k = 1$, since for larger k, we can take k disjoint copies of the $\mathcal{F}$ constructed for $k = 1$. Thus, we want an $\mathcal{F}$ in which every two d-intervals intersect and with $\tau(\mathcal{F})$ large.

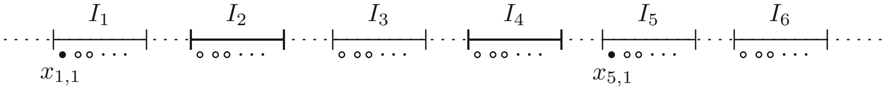
In the construction, we will use homogeneous d-intervals of a quite special form: Each component is either a single point or a unit-length interval. First, it is instructive to see why we cannot get a good example if all the components are only points. In that case, the family $\mathcal{F}$ is simply a d-uniform hypergraph (whose vertices happen to be points of the real line). We require that any two edges intersect, and thus any edge is a transversal and we have $\tau(\mathcal{F}) \leq d$.

For the actual construction, let n and N be integer parameters (whose value will be set later). Let $V = [n]$ be an index set, and I_v, for $v \in V$, be auxiliary pairwise disjoint unit intervals on the real line. In each I_v, we choose N distinct points $x_{v,i}$, $i = 1, 2, \ldots, N$.

The constructed system $\mathcal{F}$ will consist of homogeneous d-intervals $J^1, J^2, \ldots, J^N$. For each $i = 1, 2, \ldots, N$, we choose auxiliary sets $\emptyset \subset B_i \subseteq A_i \subseteq V$ and then construct J^i as follows:

$$J^i = \Big(\bigcup_{v \in B_i} I_v \Big) \cup \{x_{u,i} : u \in A_i \setminus B_i\}.$$

The picture shows an example of J^1 for $n = 6, A_1 = \{1, 2, 4, 5\}$ and $B_1 = \{2, 4\}$:



The heart of the proof is the construction of suitable sets A_i and B_i on the ground set V. Since the J^i should be homogeneous d-intervals, we obviously require

(C1) For all $i = 1, 2, \ldots, N$, $\varnothing \subset B_i \subseteq A_i$ and $|A_i| \leq d$.

The condition that every two members of $\mathcal{F}$ intersect is implied by the following:

(C2) For all i_1, i_2, $1 \leq i_1 < i_2 \leq N$, we have $A_{i_1} \cap B_{i_2} \neq \emptyset$ or $A_{i_2} \cap B_{i_1} \neq \emptyset$ (or both).

Finally, we want $\mathcal{F}$ to have no small transversal. Since no two d-intervals of $\mathcal{F}$ have a point component in common, a transversal of size t intersects no more than t members of $\mathcal{F}$ in their point components, and all the other members of $\mathcal{F}$ must be intersected in their interval components. Therefore, the transversal condition translates to

(C3) Put $t = cd^2/\log d$ for a sufficiently small constant $c > 0$, and let $\mathcal{B} = \{B_1, B_2, \dots, B_N\}$. Then $\tau(\mathcal{B}) \geq 2t$, and consequently $\tau(\mathcal{B}') \geq t$ for any $\mathcal{B}'$ arising from $\mathcal{B}$ by removing at most t sets.

A construction of sets $A_1, \dots, A_N$ and $B_1, \dots, B_N$ as above was provided by Sgall [66]. His results give the following:

Proposition 3.5 *Let b be a given integer, let $n \leq cb^2/\log b$ for a sufficiently small constant $c > 0$, and let $B_1, B_2, \dots, B_N$ be b-element subsets of $V = [n]$. Then there exist sets $A_1, A_2, \dots, A_N$, with $B_i \subseteq A_i$, $|A_i| \leq 3b$, and such that (C2) is satisfied.*

With this proposition, the proof of Proposition 3.4 is easily finished. We set $b = \lfloor \frac{d}{3} \rfloor$, $n = cb^2/\log b$, and we let $B_1, B_2, \dots, B_N$ be all the $N = \binom{n}{b}$ subsets of V of size b. We have $\tau(\{B_1, \dots, B_n\}) = n - b + 1$ and condition (C3) holds. It remains to construct the sets A_i according to Proposition 3.5; then (C1) and (C2) are satisfied too. The proof of Proposition 3.4 is concluded by passing from the A_i and B_i to the system $\mathcal{F}$ of homogeneous d-intervals as was described above. □

Sketch of proof of Proposition 3.5 Let $G = (V, E)$ be a graph on n vertices of maximum degree b with the following expander-type property: For any two disjoint b-element subsets $A, B \subseteq V$, there is at least one edge $e \in E$ connecting a vertex of A to a vertex of B. (The existence of such a graph can be easily shown by the probabilistic method; the constant c arises in this argument. See [66] for references.)

For each i, let v_i be an (arbitrary) element of the set B_i, and let

$$A_i = B_i \cup N(v_i) \cup \Big(V \setminus \bigcup_{u \in B_i} N(u)\Big),$$

where $N(v)$ denotes the set of neighbors in G of a vertex $v \in V$. It is easy to check that $|A_i| \leq 3b$, and some thought reveals that the condition (C2) is satisfied. □

3.4 A Helly-Type Problem for d-Intervals

Kaiser and Rabinovich [41] investigated conditions on a family $\mathcal{F}$ of d-intervals guaranteeing that $\mathcal{F}$ can be pierced by a "multipoint," that is, $\tau(\mathcal{F}) \leq d$ and there is a transversal using one point of each I_i. They proved the following.

Theorem 3.6 (The Kaiser–Rabinovich theorem on *d*-intervals)
Let $k = \lceil \log_2(d+2) \rceil$ and let $\mathcal{F}$ be a family of d-intervals such that any k or fewer members of $\mathcal{F}$ have a common point. Then $\mathcal{F}$ can be pierced by a multipoint.

Let's put this result into context: The proof of the Kaiser–Tardos Theorem 3.1 sets out to show that there exists a transversal consisting of exactly t points in each of the intervals I_i, for a suitable t. We eventually get that if every two d-intervals meet (that is, $\nu(\mathcal{F}) = 1$), then we can take $t < d$. The Kaiser–Rabinovich theorem says that if every $\lceil \log_2(d+2) \rceil$ meet then $t < 2$ suffices. The upcoming proof of Theorem 3.6 can be extended to yield an interpolation between this result and the Kaiser–Tardos theorem: If every $\lceil \log_b(d+2) \rceil$ edges meet, then we can take $t < b$. For $b = d$ this yields the result of Kaiser–Tardos for $\nu(\mathcal{F}) = 1$.

Proof We use notation from the proof of Theorem 3.1. We apply Lemma 3.2 with $t = 1$, obtaining a set T with one point in each T_i such that all the $2d$ vertices of the escape hypergraph $\mathcal{H} = \mathcal{H}(T)$ have the same weight W. If $W = 0$ we are done, so let us assume $W > 0$.

By the assumption on $\mathcal{F}$, every k edges of $\mathcal{H}$ share a common vertex. We will prove the following claim for every ℓ:

> If every $\ell + 1$ edges of $\mathcal{H}$ have at least m common vertices, then every ℓ edges of $\mathcal{H}$ have at least $2m + 1$ common vertices.

For $\ell = k$, the assumption holds with $m = 1$, and so by $(k-1)$-fold application of this claim, we get that every edge of $\mathcal{H}$ "intersects itself" in at least $2^k - 1$ vertices, i.e. $d > 2^k - 2$. The claim thus implies the theorem.

The claim is proved by contradiction. Suppose that $\mathcal{A} \subseteq \mathcal{H}$ is a set of ℓ edges such that $C = \bigcap \mathcal{A}$ has at most $2m$ vertices, and let $\bar{C} := \{(i, 3-j) : (i,j) \in C\}$. No edge $H \in \mathcal{H}$ contains both $(i,1)$ and $(i,2)$, thus also C does not contain both $(i,1)$ and $(i,2)$, and thus $\bar{C}$ is a subset of the complement of C; it is matched to C by $(i, 3-j) \leftrightarrow (i,j)$, and thus $|C| = |\bar{C}|$.

By the assumption, $\mathcal{A}$ plus any other edge together intersect in at least m vertices. Thus, any $H \in \mathcal{H} \setminus \mathcal{A}$ contains at least m vertices of C, and consequently no more than m vertices of $\bar{C}$.

Let U be the total weight of the vertices in C, and $\bar{U}$ the total weight of the vertices in $\bar{C}$. The edges in $\mathcal{A}$ contribute solely to U, while any other edge H contributes at least as much to U as to $\bar{U}$, and so $U > \bar{U}$. But this is impossible since all vertex weights are identical and $|C| = |\bar{C}|$. The claim, and Theorem 3.6 too, are proved. □

An interesting open problem is whether $k = \lceil \log_2(d+2) \rceil$ in Theorem 3.6 could be replaced by $k = k_0$ for some constant k_0 independent of d. The best known lower bound is $k_0 \geq 3$.

Notes Tardos [74] proved the optimal bound $\tau \leq 2\nu$ for 2-intervals by a topological argument using the homology of suitable simplicial complexes. Kaiser's argument [39] is similar to the presented one, but he proves Lemma 3.2 using a rather advanced Borsuk–Ulam-type theorem of Ramos [64] concerning continuous maps defined on products of spheres. The method with Brouwer's theorem was used by Kaiser and Rabinovich [41] for a proof of Theorem 3.6.

Lemma 3.3 seems to be new in the version that we give here, but it relates to a vast literature of "KKM-type lemmas," which starts with a paper by Knaster, Kuratowski, and Mazurkiewicz [43] from 1929. We refer to Bárány and Grinberg [7] and the references given there, such as `mathoverflow.net/questions/67318`.

Alon's short proof [2] of the bound $\tau \leq 2d^2\nu$ for families of d-intervals applies a powerful technique developed in Alon and Kleitman [4]. For the so-called Hadwiger–Debrunner (p,q)-problem solved in the latter paper, the quantitative bounds are probably quite far from the truth. It would be interesting to find an alternative topological approach to that problem, which could perhaps lead to better bounds. See, for example, Hell [34].

The variant of the piercing problem for families of homogeneous d-intervals has been considered simultaneously with d-intervals; see [2, 32, 39, 74]. The upper bounds obtained for the homogeneous case are slightly worse: $\tau \leq 3\nu$ for homogeneous 2-intervals, which is tight, and $\tau \leq (d^2-d+1)\nu$ for homogeneous d-intervals, $d \geq 3$ [39]. The reason for the worse bounds is that the escape hypergraph needs no longer be d-partite, and so Füredi's theorem [26] relating ν to ν^* gives a little worse bound (for $d = 2$, one uses a theorem of Lovász instead, asserting that $\nu^* \leq \frac{3}{2}\nu$ for any graph).

Sgall's construction [66] answered a problem raised by Wigderson in 1985. The title of Sgall's paper refers to a different, but essentially equivalent, formulation of the problem dealing with labeled tournaments.

Alon [3] proved by the method of [2] that if T is a tree and $\mathcal{F}$ is a family of subgraphs of T with at most d connected components, then $\tau(\mathcal{F}) \leq 2d^2\nu(\mathcal{F})$. More generally, he established a similar bound for the situation where T is a graph of bounded tree-width (on the other hand, if the tree-width of T is sufficiently large, then one can find a system of connected subgraps of T with $\nu = 1$ and τ arbitrarily large, and so the tree-width condition is also necessary in this sense). A somewhat weaker bound for trees has been obtained independently by Kaiser [40].

Strong results for piercing of d-trees, improving on Alon's results, were obtained by Berger [8], based on a topological approach via KKM-type lemmas. (For these see the references given above.)

Exercises

1. We have claimed that for any family $\mathcal{F}$ of intervals, it is well-known and easy to prove that $\nu(\mathcal{F}) = \tau(\mathcal{F})$. Prove this!
2. Let P and Q be convex polytopes. Show that there is a bijection between the nonempty faces of the Cartesian product $P \times Q$ and all the products $F \times G$, where F is a nonempty face of P and G is a nonempty face of Q.

3. Show that the following "Brouwer-like" claim resembling Lemma 3.3 is *not* true: If $f\colon B^n \longrightarrow B^n$ is a continuous map of the n-ball such that the boundary of B^n is mapped surjectively onto itself, then f is surjective.
4. Prove the bound $\tau(\mathcal{F}) \leq d^2\nu(\mathcal{F})$ for any family of *homogeneous* d-intervals (unions of d intervals on a single line). Hint: Follow the proof for d-intervals above, but encode a candidate transversal T by a point of a simplex (rather than a product of simplices).

4 Evasiveness

4.1 A General Model

The idea of evasiveness comes from the theory of complexity of algorithms. Evasiveness appears in different versions for graphs, digraphs and bipartite graphs. We start with a general model that contains them all.

Definition 4.1 (Argument complexity of a set system; evasiveness) In the following, we are concerned with a fixed and known set system $\mathcal{F} \subseteq 2^E$, and with the complexity of deciding whether some unknown set $A \subseteq E$ is in the set system. Here our "model of computation" is such that

> **given, and known**, is a set system $\mathcal{F} \subseteq 2^E$, where E is fixed, $|E| = m$.
> On the other hand, there is a
> **fixed, but unknown** subset $A \subseteq E$.
> We have to
> **decide** whether $A \in \mathcal{F}$, using only
> **questions** of the type "Is $e \in A$?"

(It is assumed that we always get correct answers YES or NO. We only count the *number* of questions that are needed in order to reach the correct conclusion: It is assumed that it is not difficult to decide whether $e \in A$. You can assume that some "oracle" that knows both A and $\mathcal{F}$ is answering.)

The *argument complexity* $c(\mathcal{F})$ of the set system $\mathcal{F}$ is the number of elements of the ground set E that we have to test in the worst case—with the optimal strategy.

Clearly $0 \leq c(\mathcal{F}) \leq m$. The set system $\mathcal{F}$ is *trivial* if $c(\mathcal{F}) = 0$: then no questions need to be asked; this can only be the case if $\mathcal{F} = \{\}$ or if $\mathcal{F} = 2^E$. Otherwise $\mathcal{F}$ is *non-trivial.*

The set system $\mathcal{F}$ is *evasive* if $c(\mathcal{F}) = m$, that is, if even with an optimal strategy one has to test all the elements of E in the worst case.

For example, if $\mathcal{F} = \{\emptyset\}$, then $c(\mathcal{F}) = m$: If we again and again get the answer NO, then we have to test all the elements to be sure that $A = \emptyset$. So $\mathcal{F} = \{\emptyset\}$ is an evasive set system: "being empty" is an evasive set property.

4.2 Complexity of Graph Properties

Definition 4.2 (Graph properties) Here we consider graphs on a fixed vertex set $V = [n]$. Loops and multiple edges are excluded. Thus any graph $G = (V, A)$ is determined by its edge set A, which is a subset of the set $E = \binom{n}{2}$ of "potential edges."

We identify a *property* $\mathcal{P}$ of graphs with the family of graphs that have the property $\mathcal{P}$, and thus with the set family $\mathcal{F}(\mathcal{P}) \subseteq 2^E$ given by

$$\mathcal{F}(\mathcal{P}) := \{A \subseteq E : \ ([n], A) \text{ has property } \mathcal{P}\}.$$

Furthermore, we will consider only graph properties that are isomorphism invariant; that is, properties of abstract graphs that are preserved under renumbering the vertices.

A graph property is *evasive* if the associated set system is evasive, and otherwise it is *non-evasive*.

With the symmetry condition of Definition 4.2, we would accept "being connected", "being planar," "having no isolated vertices," and "having even vertex degrees" as graph properties. However, "vertex 1 is not isolated," "123 is a triangle," and "there are no edges between odd-numbered vertices" are not graph properties.

Example 4.3 (Graph properties) For the following properties of graphs on n vertices we can easily determine the argument complexity.

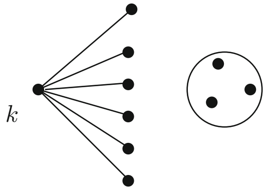
Having no edge: Clearly we have to check every single $e \in E$ in order to be sure that it is not contained in A, so this property is evasive: Its argument complexity is $c(\mathcal{F}) = m = \binom{n}{2}$.

Having at most *k* edges: Let us assume that we ask questions, and the answer we get is YES for the first k questions, and then we get NO-answers for all further questions, except for possibly the last one. Assuming that $k < m$, this implies that the property is evasive. Otherwise, for $k \geq m$, the property is trivial.

Being connected: This property is evasive for $n \geq 2$. Convince yourself that for any strategy, a sequence of "bad" answers can force you to ask all the questions.

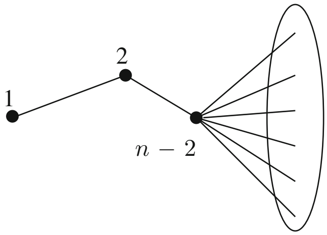
Being planar: This property is trivial for $n \leq 4$ but evasive for $n \geq 5$. In fact, for $n = 5$ one has to ask all the questions (in arbitrary order), and the answer will be $A \in \mathcal{F}$ unless we get a YES answer for all the questions—including the last one. This is, however, not at all obvious for $n > 5$: It was claimed by Hopcroft and Tarjan [35], and proved by Best, Van Emde Boas and Lenstra [10, Example 2] [15, p. 408].

A large star: Let $\mathcal{P}$ be the property of being a disjoint union of a star $\Delta_{1,n-4}$ and an arbitrary graph on 3 vertices, and let $\mathcal{F}$ be the corresponding set system.



Then $c(\mathcal{F}) < \binom{n}{2}$ for $n \geq 7$. For $n \geq 12$ we can easily see this, as follows. Test all the $\lfloor\frac{n}{2}\rfloor\lceil\frac{n}{2}\rceil$ edges $\{i, j\}$ with $i \leq \lfloor\frac{n}{2}\rfloor < j$. That way we will find exactly one vertex k with at least $\lfloor\frac{n}{2}\rfloor - 3 \geq 3$ neighbors (otherwise property $\mathcal{P}$ cannot be satisfied): That vertex k has to be the center of the star. We test all other edges adjacent to k: We must find that k has exactly $n - 4$ neighbors. Thus we have identified three vertices that are not neighbors of k: At least one of the edges between those three has not been tested. We test all other edges to check that $([n], A)$ has property $\mathcal{P}$. (This property was found by L. Carter [10, Example 16].)

Being a scorpion graph: A *scorpion graph* is an n-vertex graph that has one vertex of degree 1 adjacent to a vertex of degree 2 whose other neighbor has degree $n - 2$. We leave it as an (instructive!) exercise to check that "being a scorpion graph" is not evasive if n is large: In fact, Best, van Emde Boas and Lenstra [10, Example 18] [15, p. 410] have shown that $c(\mathcal{F}) \leq 6n$.



From these examples it may seem that most "interesting" graph properties are evasive. In fact, many more examples of evasive graph properties can be found in Bollobás [15, Sect. VIII.1], alongside with techniques to establish that graph properties are evasive, such as Milner and Welsh's "simple strategy" [15, p. 406].

Why is this model of interest? Finite graphs (similarly for digraphs and bipartite graphs) can be represented in different types of *data structures* that are not at all equivalent for algorithmic applications. For example, if a finite graph is given by an *adjacency list*, which for every vertex lists the neighbors in some order, then one can decide fast ("in linear time") whether the graph is planar, e.g. using an old algorithm of Hopcroft and Tarjan [35]; see also Mehlhorn [53, Sect. IV.10] and [54]. Note that such a planar graph has at most $3n - 6$ edges (for $n \geq 3$).

However, assume that a graph is given in terms of its adjacency matrix

$$M(G) = \big(m_{ij}\big)_{1 \leq i,j \leq n} \in \{0, 1\}^{n \times n},$$

where $m_{ij} = 1$ means that $\{i,j\}$ is an edge of G, and $m_{ij} = 0$ says that $\{i,j\}$ is not an edge. Here G is faithfully represented by the set of all $\binom{n}{2}$ superdiagonal entries (with $i < j$). Then one possibly has to inspect a large part of the matrix until one has enough information to decide whether the graph in question is planar. In fact, if $\mathcal{F} \subseteq 2^E$ is the set system corresponding to all planar graphs, then $c(\mathcal{F})$ is exactly the number of superdiagonal matrix entries that every algorithm for planarity testing has to inspect in the worst case.

The statement that "being planar" is evasive (for $n \geq 5$) thus translates into the fact that every planarity testing algorithm that starts from an adjacency matrix needs to read at least $\binom{n}{2}$ bits of the input, and hence its running time is bounded from below by $\binom{n}{2} = \Omega(n^2)$. This means that such an algorithm—such as the one considered by Fisher [24]—cannot run in linear time, and thus cannot be efficient.

Definition 4.4 (Digraph properties; bipartite graph properties)

(1) For digraph properties we again use the fixed vertex set $V = [n]$. Loops and parallel edges are excluded, but anti-parallel edges are allowed. Thus any digraph $G = (V, A)$ is determined by its arc set A, which is a subset of the set E' of all $m := n^2 - n$ "potential arcs" (corresponding to the off-diagonal entries of an $n \times n$ adjacency matrix).

A *digraph property* is a property of digraphs $([n], A)$ that is invariant under relabelling of the vertex set. Equivalently, a digraph property is a family of arc sets $\mathcal{F} \subseteq 2^{E'}$ that is symmetric under the action of $\mathfrak{S}_n$ that acts by renumbering the vertices (and renumbering all arcs correspondingly). A digraph property is *evasive* if the associated set system is evasive, otherwise it is *non-evasive*.

(2) For bipartite graph properties we use a fixed vertex set $V \uplus W$ of size $m+n$, and use $E'' := V \times W$ as the set of potential edges. A *bipartite graph property* is a property of graphs $(V \cup W, A)$ with $A \subseteq E''$ that is preserved under renumbering the vertices in V, and also under permuting the vertices in W. Equivalently, a bipartite graph property on $V \times W$ is a set system $\mathcal{F} \subseteq 2^{V \times W}$ that is stable under the action of the automorphism group $\mathfrak{S}_n \times \mathfrak{S}_m$ that acts transitively on $V \times W$.

Example 4.5 (Digraph properties) For the following digraph properties on n vertices we can determine the argument complexity.

Having at most k arcs: Again, this is clearly evasive with $c(\mathcal{F}) = m$ if $k < m = n^2 - n$, and trivial otherwise.

Having a sink: A *sink* in a digraph on n vertices is a vertex k for which all arcs going into k are present, but no arc leaves k, that is, a vertex of out-degree $\delta^+(v) = 0$, and in-degree $\delta^-(v) = n-1$. Let $\mathcal{F}$ be the set system of all digraphs on n vertices that have a sink. It is easy to see that $c(\mathcal{F}) \leq 3n - 4$. In particular, for $n \geq 3$ "having a sink" is a non-trivial but non-evasive digraph property.

In fact, if we test whether $(i,j) \in A$, then either we get the answer YES, then i is not a sink, or we get the answer NO, then j is not a sink. So, by testing arcs between pairs of vertices that "could be sinks," after $n-1$ questions we are down to one single "candidate sink" k. At this point at least one arc adjacent to k has been tested. So we need at most $2n - 3$ further questions to test whether it is a sink.

In the early 1970s Arnold L. Rosenberg conjectured that all non-trivial digraph properties have quadratic argument complexity, that is, that there is a constant $\gamma > 0$ such that for all non-trivial properties of digraphs on n vertices one has $c(\mathcal{F}) \geq \gamma n^2$. However, Stål Aanderaa found the counter-example (for digraphs) of "having a sink" [10, Example 15] [63, p. 372]. We have also seen that "being a scorpion graph" is a counter-example for graphs.

Hence Rosenberg modified the conjecture: At least all *monotone* graph properties, that is, properties that are preserved under deletion of edges, should have quadratic argument complexity. This is the statement of the *Aanderaa–Rosenberg conjecture* [65]. Richard Karp considerably sharpened the statement, as follows.

Conjecture 4.6 (The evasiveness conjecture) *Every non-trivial monotone graph property or digraph property is evasive.*

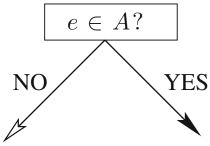
We will prove this below for graphs and digraphs in the special case when n is a prime power; from this one can derive the Aanderaa–Rosenberg conjecture, with $\gamma \approx \frac{1}{4}$. Similarly, we will prove that monotone properties of bipartite graphs on a fixed ground set $V \cup W$ are evasive (without any restriction on $|V| = m$ and $|W| = n$). However, we first return to the more general setting of set systems.

4.3 Decision Trees

Any strategy to determine whether an (unknown) set A is contained in a (known) set system $\mathcal{F}$—as in Definition 4.1—can be represented in terms of a decision tree of the following form.

Definition 4.7 A *decision tree* is a rooted, planar, binary tree whose leaves are labelled "YES" or "NO," and whose internal nodes are labelled by questions (here they are of the type "$e \in A$?"). Its edges are labelled by answers: We will represent them so that the edges labelled "YES" point to the right child, and the "NO" edges point to the left child.

A *decision tree for* $\mathcal{F} \subseteq 2^E$ is a decision tree such that starting at the root with an arbitrary $A \subseteq E$, and going to the right resp. left child depending on whether the question at an internal node we reach has answer YES or NO, we always reach a leaf that correctly answers the question "$A \in \mathcal{F}$?".

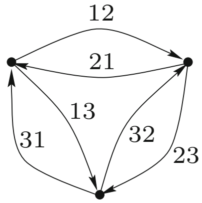


The root of a decision tree is at *level* 0, and the children of a node at level i have level $i + 1$. The *depth* of a tree is the greatest k such that the tree has a vertex at level k (a leaf).

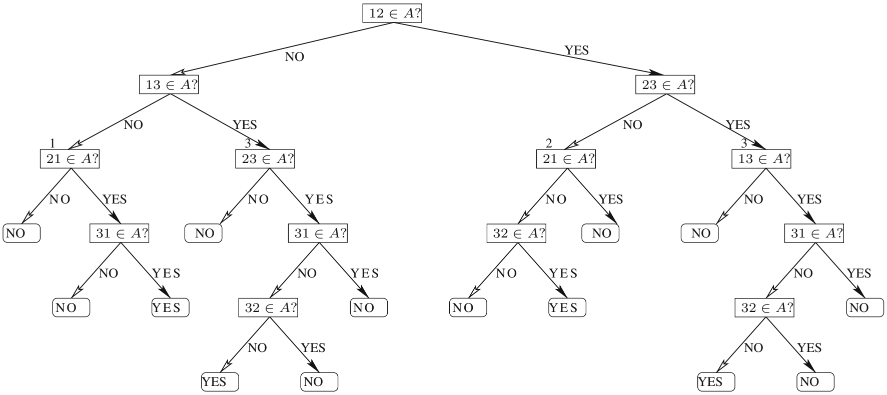
We assume (without loss of generality) that the trees we consider correspond to strategies where we never ask the same question twice.

A decision tree for $\mathcal{F}$ is *optimal* if it has the smallest depth among all decision trees for $\mathcal{F}$, that is, if it leads us to ask the smallest number of questions for the worst possible input.

Let us consider an explicit example.



The following figure represents an optimal algorithm for the "sink" problem on digraphs with $n = 3$ vertices. This has a ground set $E = \{12, 21, 13, 31, 23, 32\}$ of size $m = 6$.



The algorithm first asks, in the root node at level 0, whether $12 \in A$. In case the answer is YES (so we know that 1 is not a sink), it branches to the right, leading to a question node at level 1 that asks whether $23 \in A$?, etc. In case the answer to the question $12 \in A$? is NO (so we know that 2 is not a sink), it branches to the left, leading to a question node at level 1 that asks whether $13 \in A$?, etc.

For every possible input A (there are $2^6 = 32$ different ones), after two questions we have identified a unique "candidate sink"; after not more than 5 question nodes one arrives at a leaf node that correctly answers the question whether the graph (V, A) has a sink node: YES or NO. (The number of the unique candidate is noted next to each node at level 2.)

For each node (leaf or inner) of level k, there are exactly 2^{m-k} different inputs that lead to this node. This proves the following lemma.

Lemma 4.8 *The following are equivalent:*

- *$\mathcal{F}$ is non-evasive.*
- *The optimal decision trees $T_{\mathcal{F}}$ for $\mathcal{F}$ have depth smaller than m.*
- *Every leaf of an optimal decision tree $T_{\mathcal{F}}$ is reached by at least two distinct inputs.*

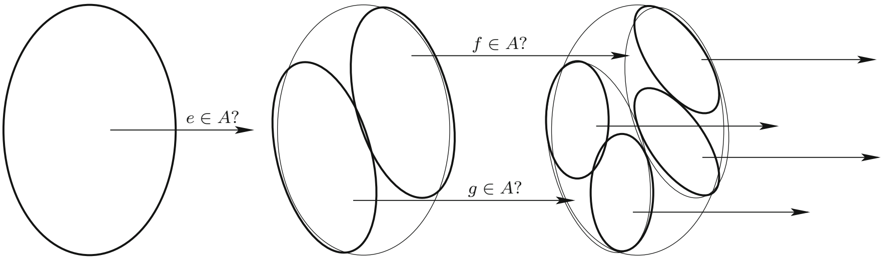
Corollary 4.9 *If $\mathcal{F}$ is non-evasive, then $|\mathcal{F}|$ is even.*

This can be used to show, for example, that the directed graph property "has a directed cycle" is evasive [10, Example 4].

Another way to view a (binary) decision tree algorithm is as follows. In the beginning, we do not know anything about the set A, so we can view the collection of possible sets as the complete boolean algebra of all 2^m subsets of E.

In the first node (at "level 0") we ask a question of the type "$e \in A$?"; this induces a subdivision of the boolean algebra into two halves, depending on whether we get answer YES or NO. If you think of the boolean algebra as a partially ordered set (indeed, a lattice), then each of the halves is an interval of length $m - 1$ of the boolean algebra $(2^E, \subseteq)$. If you prefer to think of it as a rendition of the m-dimensional hypercube, then the halves are subcubes of codimension 1, containing all the vertices of two opposite facets.

At level 1 we ask a new question, depending on the outcome of the first question. Thus we *independently* bisect the two halves of level 0, getting four pieces of the boolean algebra, all of the same size.



This process is iterated. It stops—as we do not need to ask a further question—on parts that we create that either contain only sets that are in $\mathcal{F}$ (this yields a YES-leaf) or that contain only sets not in $\mathcal{F}$ (corresponding to NO-leaves).

Thus the final result is a special type of partition of the boolean algebra into intervals. Some of them are YES intervals, containing only sets of $\mathcal{F}$, all the others are NO-intervals, containing no sets from $\mathcal{F}$. If the property in question is monotone, then the union of the YES intervals (i.e., the set system $\mathcal{F}$) forms an *ideal* in the boolean algebra, that is, a "down-closed" set such that with any set that it contains it must also contain all its subsets.

Let $p_{\mathcal{F}}(t)$ be the generating function for the set system $\mathcal{F}$, that is, the polynomial

$$p_{\mathcal{F}}(t) := \sum_{A \in \mathcal{F}} t^{|A|} \;=\; f_{-1} + tf_0 + t^2 f_1 + t^3 f_2 + \dots .$$

where $f_i = |\{A \in \mathcal{F} : |A| = i + 1\}|$.

Proposition 4.10

$$(1 + t)^{m - c(\mathcal{F})} \;\Big|\; p_{\mathcal{F}}(t).$$

Proof Consider one interval $\mathcal{I}$ in the partition of 2^E that is induced by any optimal algorithm for $\mathcal{F}$. If the leaf, at level k, corresponding to the interval is reached through a sequence of k_Y YES-answers and k_N NO-answers (with $k_Y + k_N = k$), then this means that there are sets $A_Y \subseteq E$ with $|A_Y| = k_Y$ and $A_N \subseteq E$ with $|A_N| = k_N$, such that

$$\mathcal{I} \;=\; \{A \subseteq E : A_Y \subseteq A \subseteq E \backslash A_N\}.$$

In other words, the interval $\mathcal{I}$ contains all sets that give YES-answers when asked about any of the k_Y elements of A_Y, NO-answers when asked about any of the k_N elements of A_N, while the $m - k_Y - k_N$ elements of $E \backslash (A_Y \cup A_N)$ may or may not be contained in A. Thus the interval $\mathcal{I}$ has size $2^{m - k_Y - k_N}$, and its counting polynomial is

$$p_{\mathcal{I}}(t) := \sum_{A \in \mathcal{I}} t^{|A|} \;=\; t^{k_Y} (1 + t)^{m - k_Y - k_N}.$$

Now the complete set system $\mathcal{F}$ is a disjoint union of the intervals $\mathcal{I}$, and we get

$$p_{\mathcal{F}}(t) \;=\; \sum_{\mathcal{I}} p_{\mathcal{I}}(t).$$

In particular, for an optimal decision tree we have $k_Y + k_N = k \leq c(\mathcal{F})$ and thus $m - c(\mathcal{F}) \leq m - k_Y - k_N$ at every leaf of level k, which means that all the summands $p_{\mathcal{I}}(t)$ have a common factor of $(1 + t)^{m - c(\mathcal{F})}$. □

Corollary 4.11 *If $\mathcal{F}$ is non-evasive, then $|\mathcal{F}^{even}| = |\mathcal{F}^{odd}|$, that is,*

$$-f_{-1} + f_0 - f_1 + f_2 \mp \dots = 0.$$

Proof Use Proposition 4.10, and put $t = -1$. □

We can now draw the conclusion, based only on simple counting, that most set families are evasive. This cannot of course be used to settle any specific cases, but it can at least make the various evasiveness conjectures seem more plausible.

Corollary 4.12 *Asymptotically, almost all set families $\mathcal{F}$ are evasive.*

Proof The number of set families $\mathcal{F} \subseteq 2^E$ such that

$$\#\{A \in \mathcal{F} \mid \#A \text{ odd}\} = \#\{A \in \mathcal{F} \mid \#A \text{ even}\} = k$$

is $\binom{2^{m-1}}{k}^2$. Hence, using Stirling's estimate of factorials,

$$\text{Prob}\,(\mathcal{F} \text{ non-evasive}) \leq \frac{\sum_{k=0}^{2^{m-1}} \binom{2^{m-1}}{k}^2}{2^{2^m}} = \frac{\binom{2^m}{2^{m-1}}}{2^{2^m}} \sim \frac{1}{\sqrt{\pi 2^{m-1}}} \to 0,$$

as $m \to \infty$. □

Conjecture 4.13 (The "Generalized Aanderaa–Rosenberg Conjecture", Rivest and Vuillemin [62]) *If $\mathcal{F} \subseteq 2^E$, with symmetry group $G \subseteq \mathfrak{S}_E$ that is transitive on the ground set E, and if $\emptyset \in \mathcal{F}$ but $E \notin \mathcal{F}$, then $\mathcal{F}$ is evasive.*

Note that for this it is *not* assumed that $\mathcal{F}$ is monotone. However, the assumption that $\emptyset \in \mathcal{F}$ but $E \notin \mathcal{F}$ is satisfied neither by "being a scorpion" nor by "having a sink."

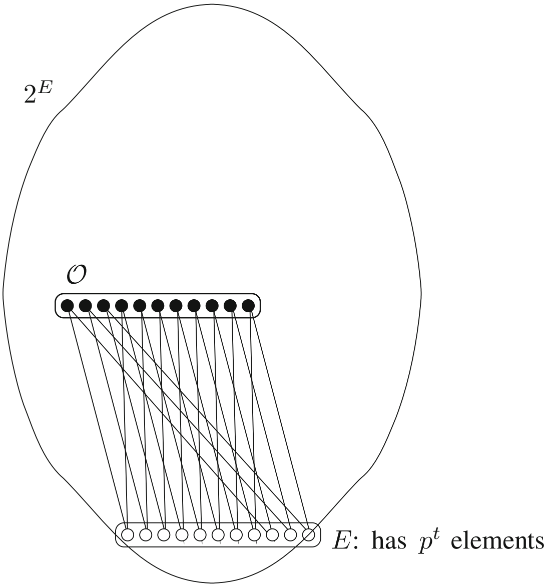
Proposition 4.14 (Rivest and Vuillemin [62]) *The Generalized Aanderaa–Rosenberg Conjecture 4.13 holds if the size of the ground set is a prime power, $|E| = p^t$.*

Proof Let $\mathcal{O}$ be any k-orbit of G, that is, a collection of k-sets $\mathcal{O} \subseteq \mathcal{F}$ on which G acts transitively. While every set in $\mathcal{O}$ contains k elements $e \in E$, we know from transitivity that every element of E is contained in the same number, say d, of sets of the orbit $\mathcal{O}$. Thus, double-counting the edges of the bipartite graph on the vertex set $E \uplus \mathcal{O}$ defined by "$e \in A$" (displayed in the figure below) we find that $k|\mathcal{O}| = d|E| = dp^t$. Thus for $0 < k < p^t$ we have that p divides $|\mathcal{O}|$, while $\{\varnothing\}$ is one single "trivial" orbit of size 1, and $k = p^t$ doesn't appear. Hence we have

$$-f_{-1} + f_0 - f_1 + f_2 \mp \cdots \equiv -1 \bmod p,$$

which implies evasiveness by Corollary 4.11.

□



Proposition 4.15 (Illies [36]) *The Generalized Aanderaa–Rosenberg Conjecture 4.13 fails for $n = 12$.*

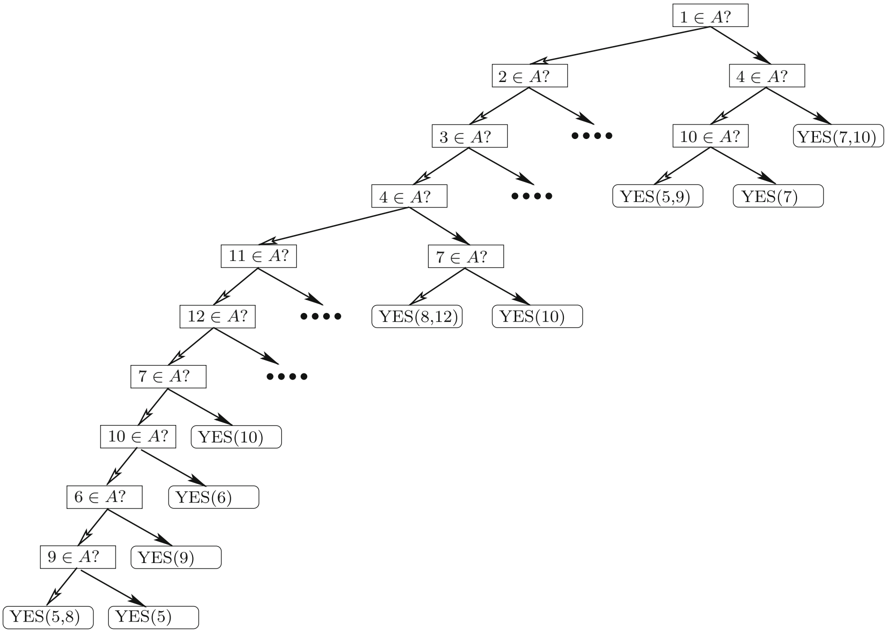
Proof Here is Illies' counterexample: Take $E = \{1, 2, 3, \ldots, 12\}$, and let the cyclic group $G = \mathbb{Z}_{12}$ permute the elements of E with the obvious cyclic action.

Take $\mathcal{F}_I \subseteq 2^E$ to be the following system of sets

- $\emptyset$, so we have $f_{-1} = 1$
- $\{1\}$ and all images under $\mathbb{Z}_{12}$, that is, all singleton sets: $f_0 = 12$,
- $\{1, 4\}$ and $\{1, 5\}$ and all images under $\mathbb{Z}_{12}$, so $f_1 = 12 + 12 = 24$,
- $\{1, 4, 7\}$ and $\{1, 5, 9\}$ and all their $\mathbb{Z}_{12}$-images, so $f_2 = 12 + 4 = 16$,
- $\{1, 4, 7, 10\}$ and their $\mathbb{Z}_{12}$-images, so $f_3 = 3$.

An explicit decision tree of depth 11 for this $\mathcal{F}_I$ is given in our figure below. Here the *pseudo-leaf* "YES(7,10)" denotes a decision tree where we check all elements $e \in E$ that have not been checked before, other than the elements 7 and 10. If none of them is contained in A, then the answer is YES (irrespective of whether $7 \in A$ or $10 \in A$), otherwise the answer is NO. The fact that two elements need not be checked means that this branch of the decision tree denoted by this "pseudo-leaf" does not go beyond depth 10. Similarly, a pseudo-leaf of the type "YES(7)" represents a subtree of depth 11.

Thus the following figure completes the proof. Here dots denote subtrees that are analogous to the ones just above. □



Note, however, that Illies' example is not monotone: For example, we have $\{1, 4, 7\} \in \mathcal{F}_I$, whereas $\{1, 7\} \notin \mathcal{F}_I$.

4.4 Monotone Systems

We now concentrate on the case where $\mathcal{F}$ is closed under taking subsets, that is, $\mathcal{F}$ is an abstract simplicial complex, which we also denote by $\Delta := \mathcal{F}$. In this setting, the symmetry group acts on Δ as a group of simplicial homeomorphisms. If $\mathcal{F}$ is a graph or digraph property, then this means that the action of G is transitive on the vertex set E of Δ, which corresponds to the edge set of the graph in question. Again we denote the cardinality of the ground set (the vertex set of Δ) by $|E| = m$.

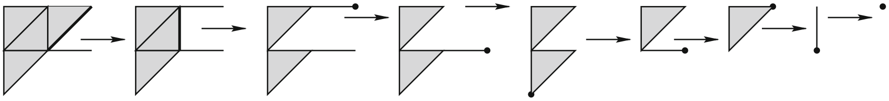
A complex $\Delta \subseteq 2^E$ is a *cone* if it has a vertex v such that $A \cup \{v\}$ is a face of Δ for any face $A \in \Delta$. For example, every simplex $\Delta = 2^E$ is a cone, but also every star graph $K_{m,1}$, considered as a simplicial complex of dimension 1, is a cone.

A complex $\Delta \subseteq 2^E$ is *collapsible* if it can be reduced to a one-point complex (equivalently, to a simplex) by steps of the form

$$\Delta \longrightarrow \Delta \backslash \{A \in \Delta : A_0 \subseteq A \subseteq A_1\},$$

where $A_0 \subset A_1$ are faces of Δ with $\varnothing \neq A_0 \neq A_1$, and A_1 is the *unique* maximal element of Δ that contains A_0. For example, every tree, considered as a simplicial complex of dimension 1, is collapsible.

Our figure illustrates a sequence of collapses that reduce a 2-dimensional complex to a point. In each case the face A_0 that is contained in a unique maximal face is drawn fattened.



Theorem 4.16 *We have the following implications:*

Δ *is a cone* $\Longrightarrow$ Δ *is non-evasive* $\Longrightarrow$ Δ *is collapsible* $\Longrightarrow$ Δ *is contractible.*

Proof The first implication is clear: For a cone we don't have to test the apex e_0 in order to see whether a set A is a face of Δ, since $A \in \Delta$ if and only if $A \cup \{e_0\} \in \Delta$. The third implication is easy topology: One can write down explicit deformation retractions. The middle implication we will derive from the following claim, which uses the notion of a *link* of a vertex e in a simplicial complex Δ: This is the complex Δ/e formed by all faces $A \in \Delta$ such that $e \notin A$ but $A \cup \{e\} \in \Delta$.

Claim Δ *is non-evasive if and only if either* Δ *is a simplex, or it is not a simplex but it has a vertex* e *such that both the deletion* $\Delta\backslash e$ *and the link* Δ/e *are non-evasive.*

Let us first verify this claim: If no questions need to be asked (that is, if $c(\Delta) = 0$), then Δ is a simplex. Otherwise we have some e that corresponds to the first question to be asked by an optimal algorithm. If one gets a YES answer, then the problem is reduced to the link Δ/e, since the faces $B \in \Delta/e$ correspond to the faces $A = B \cup \{e\}$ of Δ for which $e \in A$. In the case of a NO-answer the problem similarly reduces to the deletion $\Delta\backslash e$.

Now let us return to the proof of Theorem 4.16, where we still have to verify that "Δ is non-evasive $\Longrightarrow$ Δ is collapsible." We use induction on the number of faces of Δ.

If Δ is not a simplex, then by the Claim it has a vertex e such that the link Δ/e and the deletion $\Delta\backslash e$ are collapsible. If the link is a simplex, then deletion of e is a collapsing step $\Delta \to \Delta\backslash e$, where $\Delta\backslash e$ is collapsible, so we are done by induction.

If the link is not a simplex, then it has faces $\varnothing \subset A_0 \subset A_1$ such that A_1 is the unique maximal face in the link that contains A_0. This means that Δ has faces $\{e\} \subset A_0 \cup \{e\} \subset A_1 \cup \{e\}$ such that $A_1 \cup \{e\}$ is the unique maximal face in Δ that contains $A_0 \cup \{e\}$. In this way any collapsing step in the link Δ/e yields a collapsing step in Δ, and again we are done by induction. □

4.5 A Topological Approach

The following simple lemma provides the step from the topological fixed point theorems for complexes to combinatorial information.

Lemma 4.17 *If a (finite) group G acts vertex-transitively and with a fixed point on a finite complex Δ, then Δ is a simplex.*

Proof If $V := \{v_1, \dots, v_n\}$ is the vertex set of Δ, then any point $x \in \|\Delta\|$ has a unique representation of the form

$$x = \sum_{i=1}^{n} \lambda_i v_i,$$

with $\lambda_i \geq 0$ and $\sum_{i=1}^{n} \lambda_i = 1$. If the group action, with

$$gx = \sum_{i=1}^{n} \lambda_i gv_i,$$

is transitive, then this means that for every i, j there is some $g \in G$ with $gv_i = v_j$. Furthermore, if x is a fixed point, then we have $gx = x$ for all $g \in G$, and hence we get $\lambda_i = \lambda_j$ for all i, j. From this we derive $\lambda_i = \frac{1}{n}$ for all i. Hence we get

$$x = \frac{1}{n} \sum_{i=1}^{n} v_i$$

and this is a point in $\|\Delta\|$ only if Δ is the complete simplex with vertex set V.

Alternatively: The fixed point set of any group action is a subcomplex of the barycentric subdivision, by Lemma A.4. Thus a vertex x of the fixed point complex is the barycenter of a face A of Δ. Since x is fixed by the whole group, so is its support, the set A. Thus vertex transitivity implies that $A = E$, and $\Delta = 2^E$. □

Theorem 4.18 (The Evasiveness Conjecture for prime powers: Kahn, Saks and Sturtevant [38]) *All monontone non-trivial graph properties and digraph properties for graphs on a prime power number of vertices $|V| = q = p^t$ are evasive.*

Proof We identify the fixed vertex set V with $\mathrm{GF}(q)$. Corresponding to a non-evasive monotone non-trivial graph property we have a non-evasive complex Δ on a set $E = \binom{V}{2}$ of $\binom{q}{2}$ vertices. By Theorem 4.16 Δ is collapsible and hence $\mathbb{Z}_p$*-acyclic*, that is, all its reduced homology groups with $\mathbb{Z}_p$-coefficients vanish.

The symmetry group of Δ includes the symmetric group $\mathfrak{S}_q$, but we take only the subgroup of all "affine maps"

$$G := \{x \longmapsto ax + b : a, b \in \mathrm{GF}(q),\ a \neq 0\},$$

and its subgroup

$$P := \{x \longmapsto x + b : b \in \mathrm{GF}(q)\}$$

that permute the vertex set V, and (since we are considering graph properties) extend to an action on the vertex set $E = \binom{V}{2}$ of Δ. Then we can easily verify the following facts:

- G is doubly transitive on V, and hence induces a vertex transitive group of symmetries of the complex Δ on the vertex set $E = \binom{V}{2}$ (interpret GF(q) as a 1-dimensional vector space, then any (ordered) pair of distinct points can be mapped to any other such pair by an affine map on the line);
- P is a p-group (of order $p^t = q$);
- P is the kernel of the homomorphism that maps $(x \longmapsto ax + b)$ to $a \in \mathrm{GF}(q)^*$, the multiplicative group of GF(q), and thus a normal subgroup of G;
- $G/P \cong \mathrm{GF}(q)^*$ is cyclic (this is known from your algebra class).

Taking these facts together, we have verified all the requirements of Oliver's fixed point theorem, as provided in the Appendix as Theorem A.7. Hence G has a fixed point on Δ, and by Lemma 4.17 Δ is a simplex, and hence the corresponding (di)graph property is trivial. □

From this one can also deduce—with a lemma due to Kleitman and Kwiatowski [42, Thm. 2]—that every non-trivial monotone graph property on n vertices has complexity at least $n^2/4 + o(n^2) = m/2 + o(m)$. (For the proof see [38, Thm. 6].) This establishes the Aanderaa–Rosenberg Conjecture. On the other hand, the Evasiveness Conjecture is still an open problem for every $n \geq 10$ that is not a prime power. Kahn, Saks and Sturtevant [38, Sect. 4] report that they verified it for $n = 6$.

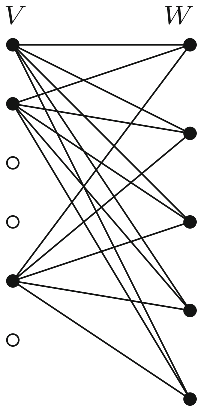
The following treats the bipartite version of the Evasiveness Conjecture. Note that in the case where mn is a prime power it follows from Proposition 4.14.

Theorem 4.19 (The Evasiveness Conjecture for bipartite graphs, Yao [76]) *All monotone non-trivial bipartite graph properties are evasive.*

Proof The ground set now is $E = V \times W$, where any monotone bipartite graph property is represented by a simplicial complex $\Delta \subseteq 2^E$.

An interesting aspect of Yao's proof is that it does not use a vertex transitive group. In fact, let the cyclic group $G := \mathbb{Z}_n$ act by cyclically permuting the vertices in W, while leaving the vertices in V fixed. The group G satisfies the assumptions of Oliver's Theorem A.7, with $P = \{0\}$. It acts on the complex Δ which is acyclic by Theorem 4.16. Thus we get from Oliver's Theorem that the fixed point set Δ^G is acyclic. This fixed point set is not a subcomplex of Δ (it does not contain any vertices of Δ), but it is a subcomplex of the order complex $\Delta(\Delta)$, which is the barycentric subdivision of Δ (Lemma A.4).

The bipartite graphs that are fixed under G are those for which every vertex in V is adjacent to none, or to all, of the vertices in W; thus they are complete bipartite graphs of the type $K_{k,n}$ for suitable k. Our figure illustrates this for the case where $m = 6$, $n = 5$, and $k = 3$.



Monotonicity now implies that the fixed graphs under G are *all* the complete bipartite graphs of type $K_{k,n}$ with $0 \leq k \leq r$ for some r with $0 \leq r < m$. (Here $r = m$ is impossible, since then Δ would be a simplex, corresponding to a trivial bipartite graph property.)

Now we observe that Δ^G is the order complex (the barycentric subdivision) of a different complex, namely of the complex whose vertices are the complete bipartite subgraphs $K_{1,n}$, and whose faces are *all* sets of at most r vertices.

Thus Δ^G is the barycentric subdivision of the $(r-1)$-dimensional skeleton of an $(m-1)$-dimensional simplex. In particular, this space is not acyclic. Even its reduced Euler characteristic, which can be computed to be $(-1)^{r-1}\binom{m-1}{r}$, does not vanish. □

We have the following sequence of implications:

$$\text{non-evasive}^{(1)} \Longrightarrow \text{collapsible}^{(2)} \Longrightarrow \text{contractible}^{(3)} \Longrightarrow \mathbb{Q}\text{-acyclic}^{(4)} \Longrightarrow \chi = 1^{(5)},$$

which corresponds to a sequence of conjectures:

Conjecture(*k*) *Every vertex-homogeneous simplicial complex with property* (k) *is a simplex.*

Here we call a simplicial complex *vertex-homogeneous* if its symmetry group acts transitively on the vertices.

The above implications show that

$$\text{Conj. (5)} \Longrightarrow \text{Conj. (4)} \Longrightarrow \text{Conj. (3)} \Longrightarrow \text{Conj. (2)} \Longrightarrow \text{Conj. (1)} \Longrightarrow \begin{array}{l}\text{Evasiveness}\\ \text{Conjecture}\end{array}$$

Here Conjecture (5) is *true* for a prime power number of vertices, by Theorem 4.14.

However, Conjectures (5) and (4) fail for $n = 6$: A counterexample is provided by the six-vertex triangulation of the real projective plane (see [52, Section 5.8]). Even Conjectures (3) and possibly (2) fail for $n = 60$: a counterexample by Oliver (unpublished), of dimension 11, is based on the group A_5; see Lutz [49].

So, it seems that Conjecture (1)—the monotone version of the Generalized Aanderaa–Rosenberg Conjecture 4.13—may be the right generality to prove, even though its non-monotone version fails by Proposition 4.15.

4.6 Quillen's Conjecture

In this final section we briefly comment on a well-known conjecture of Daniel Quillen from 1978 concerning finite groups. Upon first sight it seems very remote from the topic of evasiveness that we have just discussed, but under the surface one finds some surprising similarities.

In this section we assume familiarity with basic finite group theory, and with the topology of order complexes.

A finite group is a *p-group* if its order is a power of the prime number p. A subgroup of a finite group G is a *p-Sylow subgroup* if it is a maximal p-group. The number n_p of p-Sylow subgroups of G is called the *p-Sylow number* of G.

Let G be a finite group and p^e a prime power such that $|G| = p^e m$ and p does not divide m. Here are some well known properties.

1. There exists a p-Sylow subgroup of G of order p^e.
2. Any two p-Sylow subgroups of G are conjugate to each other.
3. $n_p(G) \equiv 1 \mod p$.

These statements are the familiar *Sylow theorems*, the first substantial results in most treatises on group theory.

For a finite group G and a prime number p dividing its order, let $L_p(G)$ denote the poset of non-trivial p-subgroups of G, ordered by inclusion. This is a ranked poset, the maximal elements of which are the p-Sylow subgroups. It becomes a lattice if one adds new bottom and top elements.

In 1978 Quillen published the following conjecture [61], which in a surprising way connects a topological condition with an algebraic one.

Conjecture 4.20 (Quillen's conjecture) *$L_p(G)$ is contractible if and only if G has a non-trivial normal p-subgroup.*

Here $L_p(G)$ refers to the order complex, whose simplices are the totally ordered chains $x_0 < x_1 < \cdots < x_d$ of $L_p(G)$. The "if" direction, which is very easy, was proved by Quillen, and he proved the "only if" direction for the case of solvable groups. The conjecture has since then been verified in many cases, but the general case is still wide open.

In the previous section we considered an array of conjectures, among them this one:

Conjecture (3) *Every vertex-homogeneous contractible simplicial complex is a simplex.*

This conjecture turns out to be relevant both for evasiveness and for p-subgroups:

Conjecture (3) $\Longrightarrow$ Evasiveness Conjecture,
Conjecture (3) $\Longrightarrow$ Quillen's Conjecture.

However, Conjecture (3) is false. It was mentioned in the previous section that counterexamples on 60 vertices are known. So, why spend time on discussing it? We believe that it is nevertheless instructive to see in which way Conjecture (3) is relevant for Quillen's Conjecture. It is conceivable that progress for one of the Evasiveness Conjecture and the Quillen Conjecture can lead to progress for the other.

Proposition 4.21 *Conjecture* (3) $\Longrightarrow$ *Quillen's Conjecture*

Proof Suppose that $L_p(G)$ is contractible. We are to prove that G has a non-trivial normal p-subgroup.

Define the auxiliary *Sylow complex* $\mathrm{Syl}_p(G)$ this way: The vertices are the p-Sylow subgroups of G. A collection of such subgroups form a simplex (or, face) of $\mathrm{Syl}_p(G)$ if their intersection is nontrivial (not just the identity). This is clearly a simplicial complex.

An application of the nerve theorem (or the crosscut theorem), see Björner [12, p. 1850], shows that these two complexes are of same homotopy type:

$$\mathrm{Syl}_p(G) \sim L_p(G)$$

The group G acts by conjugation on the vertex set of $\mathrm{Syl}_p(G)$, and by the second Sylow theorem this action is transitive. So, $\mathrm{Syl}_p(G)$ is a vertex-homogeneous and contractible complex. Conjecture (3) then implies that $\mathrm{Syl}_p(G)$ is a big simplex. This means precisely that the intersection of *all* p-Sylow subgroups is non-trivial and is a fixed point under the action. Hence this is a non-trivial normal p-subgroup.

Following along the reasoning in this proof can help to verify the Quillen conjecture in some special cases, such as this.

Proposition 4.22 *If $n_p = q^e$, that is, if the number of p-Sylow subgroups is the power of some prime number q, then G satisfies the Quillen conjecture.*

Here the Rivest–Vuillemin Theorem 4.14 is relevant. In fact, with this and Conjecture (5) a sharper version of the Quillen conjecture can be obtained in the case when $n_p = q^e$, using trivial Euler characteristic instead of contractibility. We leave further thoughts and experiments in this direction to the reader.

Notes The classical textbook account on evasiveness, from the Graph Theory point of view, is in Bollobas [15, Chap. VIII].

A textbook account from a Topological Combinatorics point-of-view was recently given in de Longueville [47, Chap. 3]. The appendices A–E to this book also provide a concise and user-friendly account of the Algebraic Topology tools employed. See also Miller [55].

Gorenstein [30] is a standard text on finite groups. The book by Smith [70] contains a wealth of material on subgroup lattices and can serve as our general reference for these.

Exercises

1. What kind of values of $c(\mathcal{F})$ are possible for graph properties of graphs on n vertices? For monotone properties, it is assumed that one has $c(\mathcal{F}) \in \{0, m\}$, and this is proved if n is a prime power. In general, it is known that $c(\mathcal{F}) \geq 2n-4$ unless $c(\mathcal{F}) = 0$, by Bollobás and Eldridge [16], see [15, Sect. VIII.5].
2. Show that the digraph property "has a sink" has complexity

$$c(\mathcal{F}_{sink}) \leq 3(n-1) - \lfloor \log_2(n) \rfloor .$$

Can you also prove that for any non-trivial digraph property one has $c(\mathcal{F}) \geq c(\mathcal{F}_{sink})$?

(This is stated in Best, van Emde Boas and Lenstra [10, p. 17]; there are analogous results by Bollobás and Eldridge [16] [15, Sect. VIII.5] in a different model for digraphs.)
3. Show that if a complex Δ corresponds to a non-evasive monotone graph property, then it has a complete 1-skeleton.
4. Give examples of simplicial complexes that are contractible, but not collapsible. (The "dunce hat" is a key word for a search in the literature …)
5. Assume that when testing some unknown set A with respect to a set system $\mathcal{F}$, you always get the answer YES if there is *any* set $A \in \mathcal{F}$ for which this YES and all the previous answers are correct, that is, unless this "YES" would allow you to conclude $A \notin \mathcal{F}$ at this point.

   (i) Show that with this type of answers you *always* need m questions for *any* algorithm (and thus $\mathcal{F}$ is evasive) if and only if $\mathcal{F}$ satisfies the following property:
   (*) for any $e \in A \in \mathcal{F}$ there is some $f \in E \backslash A$ such that $A \backslash \{e\} \cup \{f\} \in \mathcal{F}$.
   (ii) Show that for $n \geq 5$, the family $\mathcal{F}$ of edge sets of planar graphs satisfies property $(*)$.
   (iii) Give other examples of graph properties that satisfy $(*)$, and are thus evasive.

   (This is the "simple strategy" of Milner and Welsh [56]; see Bollobás [15, p. 406].)
6. Let Δ be a vertex-homogeneous simplicial complex with n vertices and Euler characteristic $\chi(\Delta) = -1$. Suppose that $n = p_1^{e1} \cdots p_k^{ek}$ is the prime factorization of n and let $m = \max\{p_1^{e1}, \ldots, p_k^{ek}\}$. Prove that $\dim \Delta \geq m - 1$.
7. Let W_n^q be the set of all words of length n in the alphabet $\{1, 2, \ldots, q\}$, $q \geq 2$. For subsets $\mathcal{F} \subseteq W_n^q$, let $c(\mathcal{F})$ be the least number of inspections of single letters

(or rather, positions) that the best algorithm needs in the worst case $s \in W_n^q$ in order to decide whether $s \in \mathcal{F}$.

Define the polynomial

$$p_{\mathcal{F}}(x_1, \ldots, x_q) = \sum_{s \in \mathcal{F}} x_1^{\mu_1} \cdots x_q^{\mu_q},$$

where $\mu_i = \#\{j : s_j = i\}$ for $s = s_1 \cdots s_q$.

Show that

$$(x_1 + \cdots + x_q)^{n-c(\mathcal{F})} \mid p_{\mathcal{F}}(x_1, \ldots, x_q).$$

Appendix: Fixed Point Theorems and Homology

Lefschetz' Theorem

Fixed point theorems are "global–local tools": From global information about a space (such as its homology) they derive local effects, such as the existence of special points where "something happens."

Of course, in applications to combinatorial problems we need to combine them with suitable "continuous–discrete tools": From continous effects, such as topological information about continuous maps of simplicial complexes, we have to find our way back to combinatorial information.

In this Appendix we assume familiarity with more Algebra and Algebraic Topology than in other parts of these lecture notes, including some basic finite group theory, chain complexes, etc. As this is meant to be a reference and survey section, no detailed proofs will be given. A main result we head for is Oliver's Theorem A.7, which is needed in Sect. 4. On the way to this, skim or skip, depending on your tastes and familiarity[4] with these notions.

A powerful tool on our agenda (which yields a classical proof for Brouwer's fixed point theorem and some of its extensions) is Hopf's trace theorem. For this let V be any finite-dimensional vector space, or a free abelian group of finite rank. When we consider an endomorphism $g\colon\ V \longrightarrow V$ then the *trace* $\operatorname{trace}(g)$ is the sum of the diagonal elements of the matrix that represents g. The trace is independent of the basis chosen for V. In the case when V is a free abelian group, then $\operatorname{trace}(g)$ is an integer.

Theorem A.1 (The Hopf trace theorem) *Let Δ be a finite simplicial complex, let $f\colon \|\Delta\| \longrightarrow \|\Delta\|$ be a self-map, and denote by $f_{\#i}$ resp. f_{*i} the maps that f induces on i-dimensional chain groups resp. homology groups.*

[4] See [52] for a detailed discussion of simplicial complexes, their geometric realizations, etc. In particular, we use the notation $\|K\|$ for the polyhedron (the geometric realization of a simplicial complex Δ).

Using an arbitrary field of coefficients **k**, *one has*

$$\sum_i (-1)^i \mathit{trace}(f_{\# i}) \;=\; \sum_i (-1)^i \mathit{trace}(f_{* i}).$$

The same identity holds if we use integer coefficients, and compute the traces for homology in the quotients $H_i(\Delta,\mathbb{Z})/T_i(\Delta,\mathbb{Z})$ of the homology groups modulo their torsion subgroups; these quotients are free abelian groups.

This theorem is remarkable as it allows to compute a topological invariant that depends solely on the homotopy class of f, by means of a simple combinatorial counting. The proof for this uses the definition of simplicial homology, and simple linear algebra; we refer to Munkres [58, Thm. 22.1] or Bredon [19, Sect. IV.23].

For an arbitrary coefficient field **k**, the *Lefschetz number* of the map $f\colon\ \|\Delta\| \longrightarrow \|\Delta\|$ is defined as

$$L_{\mathbf{k}}(f) := \sum_i (-1)^i \mathrm{trace}(f_{*i}) \quad \in \mathbf{k}.$$

Similarly, taking integral homology modulo torsion, the *integral Lefschetz number* is defined as

$$L(f) := \sum_i (-1)^i \mathrm{trace}(f_{*i}) \quad \in \mathbb{Z}.$$

The universal coefficient theorems imply that one always has $L_{\mathbb{Q}}(f) = L(f)$: Thus the integral Lefschetz number $L(f)$ can be computed in rational homology, but it is an integer.

The *Euler characteristic* of a complex Δ coincides with the Lefschetz number of the identity map $\mathrm{id}_\Delta\colon\ \|\Delta\| \longrightarrow \|\Delta\|$,

$$\chi(\Delta) = L(\mathrm{id}_\Delta), \quad \text{where } \mathrm{trace}((\mathrm{id}_\Delta)_{*i}) = \beta_i(\Delta).$$

Thus the Hopf trace theorem yields that the Euler characteristic of a finite simplicial complex Δ can be defined resp. computed without a reference to homology, simply as the alternating sum of the face numbers of the complex Δ, where $f_i = f_i(\Delta)$ denotes the number of i-dimensional faces of Δ:

$$\chi(\Delta) := f_0(\Delta) - f_1(\Delta) + f_2(\Delta) - \cdots .$$

This is then a finite sum that ends with $(-1)^d f_d(\Delta)$ if Δ has dimension d. Thus the Hopf trace theorem applied to the identity map just reproduces the Euler–Poincaré formula. This proves, for example, the d-dimensional Euler polyhedron formula, not only for polytopes, but also for general spheres, shellable or not (as discussed in Ziegler [77]). The Hopf trace formula also has powerful combinatorial applications, see Ziegler [78]. However, for us its main consequence is the following theorem, which is a vast generalization of the Brouwer fixed point theorem.

Theorem A.2 (The Lefschetz fixed point theorem) *Let Δ be a finite simplicial complex, and* $\mathbf{k}$ *an arbitrary field. If a self-map* $f\colon \|\Delta\| \longrightarrow \|\Delta\|$ *has Lefschetz number* $L_{\mathbf{k}}(f) \neq 0$*, then* f *and every map homotopic to* f *have a fixed point.*

In particular, if Δ *is* $\mathbb{Z}_p$*-acyclic for some prime* p*, then every continuous map* $f\colon \|\Delta\| \longrightarrow \|\Delta\|$ *has a fixed point.*

(A complex is $\mathbb{Z}_p$-acyclic if its reduced homology with $\mathbb{Z}_p$-coefficients vanishes. That is, in terms of homology it looks like a contractible space, say a d-ball.)

Proof (Sketch) For a finite simplicial complex Δ, the polyhedron $\|\Delta\|$ is compact. So if f does not have a fixed point, there is some $\varepsilon > 0$ such that $|f(x)-x| > \varepsilon$ for all $x \in \Delta$. Now take a subdivision Δ' of Δ into simplices of diameter smaller than ε, and a simplicial approximation of error smaller than $\varepsilon/2$, so that the simplicial approximation $f' : \Delta' \to \Delta'$, which is homotopic to f, does not have a fixed point, either.

Now apply the trace theorem to see that $L_{\mathbf{k}}(f)$ is zero, contrary to the assumption, where the induced map $f'_{*0} = f_{*0}$ in 0-dimensional homology is the identity. □

Note that Brouwer's fixed point Theorem 2.4 is the special case of Theorem A.2 when Δ triangulates a ball.

For a reasonably large class of spaces, a converse to the Lefschetz fixed point theorem is also true: If $L(f) = 0$, then f is homotopic to a map without fixed points. See Brown [21, Chap. VIII].

The Theorems of Smith and Oliver

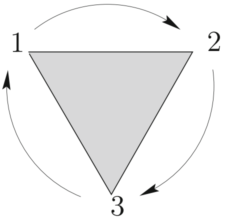
In addition to the usual game of connections between graphs, posets, complexes and spaces, we will now add groups. Namely we will discuss some useful topological effects caused by symmetry, that is, by finite group actions.

A (finite) group G *acts* on a (finite) simplicial complex Δ if each group element corresponds to a permutation of the vertices of Δ, where composition of group elements corresponds to composition of permutations, in such a way that $g(A) := \{gv : v \in A\}$ is a face of Δ for all $g \in G$ and for all $A \in \Delta$. This action on the vertices is extended to the geometric realization of the complex Δ, so that G acts as a group of simplicial homeomorphisms $g\colon \|\Delta\| \longrightarrow \|\Delta\|$.

The action is *faithful* if only the identity element in G acts as the identity permutation. In general, the set $G_0 := \{g \in G : gv = v \text{ for all } v \in \text{vert}(\Delta)\}$ is a normal subgroup of G. Hence we get that the quotient group G/G_0 acts faithfully on Δ, and we usually only consider faithful actions. In this case, we can interpret G as a subgroup of the *symmetry group* of the complex Δ. The action is *vertex transitive* if for any two vertices v, w of Δ there is a group element $g \in G$ with $gv = w$.

A *fixed point* (also known as *stable point*) of a group action is a point $x \in \|\Delta\|$ that satisfies $gx = x$ for all $g \in G$. We denote the set of all fixed points by Δ^G. Note that Δ^G is in general not a subcomplex of Δ.

Example A.3 Let $\Delta = 2^{[3]}$ be the complex of a triangle, and let $G = \mathbb{Z}_3$ be the cyclic group (a proper subgroup of the symmetry group $\mathfrak{S}_3$), acting such that a generator cyclically permutes the vertices, $1 \longmapsto 2 \longmapsto 3 \longmapsto 1$.



This is a faithful action; its fixed point set consists of the center of the triangle only—this is not a subcomplex of Δ, although it corresponds to a subcomplex of the barycentric subdivision $\mathrm{sd}(\Delta)$.

Lemma A.4 (Two barycentric subdivisions)

(1) *After replacing Δ by its barycentric subdivision (informally, let $\Delta := \mathrm{sd}(\Delta)$), we get that the fixed point set Δ^G is a subcomplex of Δ.*

(2) *After replacing Δ once again by its barycentric subdivision (so now $\Delta := \mathrm{sd}^2(\Delta)$), we even get that the quotient space $\|\Delta\|/G$ can be constructed from Δ by identifying all faces with their images under the action of G. That is, the equivalence classes of faces of Δ, with the induced partial order, form a simplicial complex that is homeomorphic to the quotient space $\|\Delta\|/G$.*

We leave the proof as an exercise. It is not difficult; for details and further discussion see Bredon [18, Sect. III.1].

"Smith Theory" was started by P. A. Smith [69] in the thirties. It analyzes finite group actions on compact spaces (such as finite simplicial complexes), providing relations between the structure of the group to its possible fixed point sets. Here is one key result.

Theorem A.5 (Smith [68]) *If P is a p-group (that is, a finite group of order $|P| = p^t$ for a prime p and some $t > 0$), acting on a complex Δ that is $\mathbb{Z}_p$-acyclic, then the fixed point set Δ^P is $\mathbb{Z}_p$-acyclic as well. In particular, it is not empty.*

Proof (Sketch) The key is that, with the preparations of Lemma A.4, the maps that f induces on the chain groups (with $\mathbb{Z}_p$ coefficients) nicely restrict to the chain groups on the fixed point set Δ^P. Passing to traces and using the Hopf trace theorem, one can derive that Δ^P is non-empty. A more detailed analysis leads to the "transfer isomorphism" in homology, which proves that Δ^P must be acyclic.

See Bredon [18, Thm. III.5.2] and Oliver [60, p. 157], and also de Longueville [47, Appendix D and E]. □

On the combinatorial side, one has an Euler characteristic relation due to Floyd [25] [18, Sect. III.4]:

$$\chi(\Delta) \;+\; (p-1)\chi(\Delta^{\mathbb{Z}_p}) \;=\; p\,\chi(\Delta/\mathbb{Z}_p).$$

If P is a p-group (in particular for $P = \mathbb{Z}_p$), then this implies that

$$\chi(\Delta^P) \equiv \chi(\Delta) \pmod{p},$$

using induction on t, where $|P| = p^t$.

Theorem A.6 (Oliver [60, Lemma I]) *If $G = \mathbb{Z}_n$ is a cyclic group, acting on a $\mathbb{Q}$-acyclic complex Δ, then the action has a fixed point.*

In this case the fixed point set Δ^G has the Euler characteristic of a point, $\chi(\Delta^G) = 1$.

Proof The first statement follows directly from the Lefschetz fixed point theorem: Any cyclic group is generated by a single element g, this element has a fixed point, this fixed point of g is also a fixed point of all powers of g, and hence of the whole group G.

For the second part, take p^t to be a maximal prime power that divides n, consider the corresponding subgroup isomorphic to $\mathbb{Z}_{p^t}$, and use induction on t and the transfer homomorphism, as for the previous proof. □

Unfortunately, results like these may give an overly optimistic impression of the generality of fixed point theorems for acyclic complexes. There are fixed point free finite group actions on balls: Examples were constructed by Floyd and Richardson and others; see Bredon [18, Sect. I.8].

On the positive side we have the following result due to Oliver, which plays a central role in Sect. 4.5.

Theorem A.7 (Oliver's Theorem I [60, Prop. I]) *If G has a normal subgroup $P \lhd G$ that is a p-group, such that the quotient G/P is cyclic, acting on a complex Δ that is $\mathbb{Z}_p$-acyclic, then the fixed point set Δ^G is $\mathbb{Z}_p$-acyclic as well. In particular, it is not empty.*

This is as much as we will need in this chapter. Oliver proved, in fact, a more general and complete theorem that includes a converse.

Theorem A.8 (Oliver's Theorem II [60]) *Let G be a finite group. Every action of G on a $\mathbb{Z}_p$-acyclic complex Δ has a fixed point if and only if G has the following structure:*

> *G has normal subgroups $P \lhd Q \lhd G$ such that P is a p-group, G/Q is a q-group (for a prime q that need not be distinct from p), and the quotient Q/P is cyclic.*

In this situation one always has $\chi(\Delta^G) \equiv 1 \bmod q$.

Notes The Lefschetz–Hopf fixed point theorem was announced by Lefschetz for a restriced class of complexes in 1923, with details appearing three years later. The first proof for the general version was by Hopf in 1929. There are generalizations, for example to Absolute Neighborhood Retracts; see Bredon [19, Cor. IV.23.5] and Brown [21, Chap. IIII]. We refer to Brown's book [21].

We refer to Bredon [18, Chapter III] for a nice textbook treatment of Smith Theory. The book by de Longueville [47, Appendix E] also has a very accessible discussion of the fixed point theorems of Smith and Oliver. The exercises concerning fixed point sets of poset maps $P \to P$ are drawn from Baclawski and Björner [6].

Exercises

1. Verify directly that if f maps $\|T\|$ to $\|T\|$, where T is a graph-theoretic tree, then f has a fixed point.

   How would you derive this from the Lefschetz fixed point theorem?
2. Let P be a poset (finite partially ordered set), and denote by $\Delta(P)$ its order complex (whose faces are the totally ordered subsets). Suppose that $f\colon P \to P$ is an order-preserving mapping with fixed point set $P^f := \{x \in P \mid f(x) = x\}$.

   (a) Show that if $\Delta(P)$ is acyclic over some field, then
   $$\mu(P^f) = 0,$$
   where $\mu(P^f)$ denotes the *Möbius function* (reduced Euler characteristic) of $\Delta(P^f)$. In particular, P^f is not empty.

   (b) Does it follow also that P^f itself is acyclic?
3. Suppose now that $f\colon P \to P$ is order-reversing and let $P_f := \{x \in P \mid x = f^2(x) \le f(x)\}$. Show that if $\Delta(P)$ is acyclic over some field, then
   $$\mu(P_f) = 0.$$
   In particular, if f has no fixed edge (i.e., no x such that $x = f^2(x) < f(x)$) then f has a unique fixed point.

Acknowledgements We are grateful to Marie-Sophie Litz and to the referees for very careful reading and a great number of very valuable comments and suggestions on the manuscript. Thanks to Moritz Firsching and Stephen D. Smith, and in particular to Penny Haxell, for additional references and very helpful explanations.

References

1. M. Aigner, G.M. Ziegler, *Proofs from THE BOOK*, 5th edn. (Springer, Heidelberg, 2014)
2. N. Alon, Piercing d-intervals. Discret. Comput. Geometry **19**, 333–334 (1998)
3. N. Alon, Covering a hypergraph of subgraphs. Discret. Math. **257**, 249–254 (2002)
4. N. Alon, D. Kleitman, Piercing convex sets and the Hadwiger Debrunner (p, q)-problem. Adv. Math. **96**, 103–112 (1992)

5. N. Amenta, M. Bern, D. Eppstein, S.-H. Teng, Regression depth and center points. Discret. Comput. Geomet. **23**, 305–323 (2000)
6. K. Baclawski, A. Björner, Fixed points in partially ordered sets Adv. Math. **31**, 263–287 (1979)
7. I. Bárány, V.S. Grinberg, Block partitions of sequences. Israel J. Math. **206**, 155–164 (2015)
8. E. Berger, KKM – a topological approach for trees. Combinatorica **25**, 1–18 (2004)
9. E.R. Berlekamp, J.H. Conway, R.K. Guy, *Winning Ways. Vol. 2: Games in Particular* (Academic Press, London, 1982)
10. M.R. Best, P. van Emde Boas, H.W. Lenstra Jr., A sharpened version of the Anderaa–Rosenberg conjecture. Technical Report ZW 30/74, Mathematisch Centrum Amsterdam, Afd. Zuivere Wisk., 1974, 20 pp.
11. A. Björner, Combinatorics and topology. Not. Am. Math. Soc. **32**, 339–345 (1985)
12. A. Björner, Topological methods, Chap. 34, in *Handbook of Combinatorics*, vol. II, ed. by R. Graham, M. Grötschel, L. Lovász (North Holland, Amsterdam), pp. 1819–1872
13. D. Blackwell, M.A. Girshick, *Theory of Games and Statistical Decisions* (Wiley, New York, 1954)
14. P.V.M. Blagojević, G.M. Ziegler, Beyond the Borsuk-Ulam theorem: the topological tverberg story. In this volume
15. B. Bollobás, *Extremal Graph Theory* (Academic Press, London, 1978)
16. B. Bollobás, S.E. Eldridge, Packings of graphs and applications to computational complexity. J. Combin. Theory Ser. B **25**, 105–124 (1978)
17. A.V. Bondarenko, M.S. Viazovska, Spherical designs via Brouwer fixed point theorem. SIAM J. Discret. Math. **24**, 207–217 (2010)
18. G.E. Bredon, *Introduction to Compact Transformation Groups* (Academic Press, New York, 1972)
19. G. Bredon, *Topology and Geometry*. Graduate Texts in Mathematics, vol. 139 (Springer, New York 1993)
20. L.E.J. Brouwer, Über Abbildungen von Mannigfaltigkeiten, Math. Annalen **71**, 97–115 (1911)
21. R.F. Brown, *The Lefschetz Fixed Point Theorem* (Scott, Foresman and Co., Glenview, 1971)
22. C. Browne, *HEX Strategy* (A K Peters, Wellesley, 2000)
23. R. Engelking, *General Topology* (PWN, Warszawa, 1977)
24. G.J. Fisher, Computer recognition and extraction of planar graphs from their incidence matrix. IEEE Trans. Circuit Theory **2**(CT-17), 154–163 (1966)
25. E.E. Floyd, On periodic maps and the euler characteristics of the associated spaces. Trans. Am. Math. Soc. **72**, 138–147 (1952)
26. Z. Füredi, Maximum degree and fractional matching in uniform hypergraphs. Combinatorica **1**, 155–162 (1981)
27. D. Gale, The game of Hex and the Brouwer fixed-point theorem. Am. Math. Mon. **86**, 818–827 (1979)
28. M. Gardner. *The Scientific American Book of Mathematical Games and Diversions* (Simon and Schuster, New York, 1958). Reprinted in "Hexaflexagons, Probability Paradoxes, and the Tower of Hanoi" (Mathematical Association of America, Cambridge University Press, 2008)
29. M. Gardner, *Mathematical Carnival*, updated and revised edn. (Mathematical Association of America, Washington, DC, 1989)
30. D. Gorenstein, *Finite Groups*, 2nd edn. (Chelsea Publishing Company, New York, 1980). Reprint by AMS Chelsea Publishing, 2007
31. A. Gyárfás, J. Lehel, A Helly-type problem in trees, in *Combinatorial Theory and Applications*, ed. by P. Erdős et al. Colloquia Mathematica Societatis Janos Bolyai, vol. 4 (North-Holland, Amsterdam, 1970), pp. 57–1584
32. A. Gyárfás, J. Lehel, Covering and coloring problems for relatives of intervals. Discret. Math. **55**, 167–180 (1985)
33. A.W. Hales, R.I. Jewett, Regularity and positional games. Trans. Am. Math. Soc. **106**, 222–229 (1963)
34. S. Hell, On a topological fractional Helly theorem. Preprint, June 2005, 11 pp. arXiv:math/0506399

35. J. Hopcroft, R. Tarjan, Efficient planarity testing. J. ACM **21**, 549–568 (1974)
36. N. Illies, A counterexample to the generalized Aanderaa–Rosenberg conjecture. Inf. Process. Lett. **7**, 154–155 (1978)
37. T.R. Jensen, B. Toft, *Graph Coloring Problems* (Wiley-Interscience, New York, 1995)
38. J. Kahn, M. Saks, D. Sturtevant, A topological approach to evasiveness. Combinatorica **4**, 297–306 (1984)
39. T. Kaiser, Transversals of d-intervals. Discret. Comput. Geomet. **18**, 195–203 (1997)
40. T. Kaiser, Piercing problems and topological methods. Doctoral dissertation, Department of Applied Mathematics, Charles University, Prague, 1998
41. T. Kaiser, Y. Rabinovich, Intersection properties of families of convex (n, d)-bodies. Discret. Comput. Geomet. **21**, 275–287 (1999)
42. D.J. Kleitman, D.J. Kwiatowski, Further results on the Aanderaa–Rosenberg conjecture. J. Combin. Theory Ser. B **28**, 85–95 (1980)
43. B. Knaster, C. Kuratowski, S. Mazurkiewicz, Ein Beweis des Fixpunktsatzes für n-dimensionale Simplexe. Fundamenta Mathematicae **14**, 132–137 (1929)
44. S. Kryński, Remarks on matroids and Sperner's lemma. Europ. J. Comb. **11**, 485–488 (1990)
45. B. Lindström, On Matroids and Sperner's Lemma Europ. J. Comb. **2**, 65–66 (1981)
46. M. de Longueville, 25 years proof of the Kneser conjecture – the advent of topological combinatorics. EMS-Newsletter **53**, 16–19 (2004)
47. M. de Longueville, *A Course in Topological Combinatorics.* Universitext (Springer, New York, 2013)
48. L. Lovász, Matroids and Sperner's Lemma Europ. J. Comb. **1**, 65–66 (1980)
49. F.H. Lutz, Examples of $\mathbb{Z}$-acyclic and contractible vertex-homogeneous simplicial complexes. Discret. Comput. Geom. **27**, 137–154 (2002)
50. P. Mani, Zwei kombinatorische Sätze vom Typ Sperner–Tucker–Ky Fan. Monatshefte Math. Physik **71**, 427–435 (1967)
51. J. Matoušek, Lower bounds on the transversal numbers of d-intervals. Discret. Comput. Geom. **26**, 283–287 (2001)
52. J. Matoušek, *Using the Borsuk–Ulam Theorem*, revised second printing 2008 (Springer, Berlin/Heidelberg, 2003)
53. K. Mehlhorn, *Data Structures and Efficient Algorithms. Vol. 2: Graph Algorithms and NP-Completeness* (Springer, Berlin, 1984)
54. K. Mehlhorn, P. Mutzel, On the embedding phase of the Hopcroft and Tarjan planarity testing algorithm. Algorithmica **16**, 233–242 (1996)
55. C.A. Miller, Evasiveness of graph properties and topological fixed point properties. Found. Trends Theor. Comput. Sci. **7**(4), 337–415 (2011)
56. E.C. Milner, D.J.A. Welsh, On the computational complexity of graph theoretical properties, in *Proceedings of the Fifth British Combinatorial Conference*, Aberdeen, 1975, ed. by C.S.J.A. Nash-Williams, J. Sheehan (Utilitas Mathematica, Winnipeg, 1976), pp. 471–487
57. J. Milnor, A Nobel prize for John Nash. Math. Intell. **17**(3), 11–17 (1995)
58. J.R. Munkres, *Elements of Algebraic Topology* (Addison-Wesley, Reading, 1984)
59. J.R. Munkres, *Topology. A First Course*, 2nd edn. (Prentice-Hall, Englewood Cliffs, 2000)
60. R. Oliver, Fixed-point sets of group actions on finite acyclic complexes. Commentarii Math. Helvetii **50**, 155–177 (1975)
61. D. Quillen, Homotopy properties of the poset of non-trivial p-subgroups of a group. Adv. Math. **28**, 101–128 (1978)
62. R. Rivest, S. Vuillemin, A generalization and proof of the Aanderaa–Rosenberg conjecture, in *Proceedings of the 7th Annual Symposium on Theory of Computing*, Albuquerque, 1975 (ACM, 1976), pp. 6–11
63. R. Rivest, S. Vuillemin, On recognizing graph properties from adjacency matrices. Theor. Comput. Sci. **3**, 371–384 (1978)
64. E.A. Ramos, Equipartition of mass distributions by hyperplanes. Discret. Comput. Geomet. **15**, 147–167 (1996)

65. A.L. Rosenberg, On the time required to recognize properties of graphs: a problem. SIGACT News **5**, 15–16 (1973)
66. J. Sgall, Solution of a covering problem related to labelled tournaments. J. Graph Theory **23**, 111–118 (1996)
67. J.H. Shapiro, *A Fixed-Point Farrago*. Undergraduate Texts in Mathematics (Springer, New York, 2016)
68. P.A. Smith, Transformations of finite period. Ann. Math. **39**, 127–164 (1938)
69. P.A. Smith, Fixed point theorems for periodic transformations. Am. J. Math. **63**, 1–8 (1941)
70. S.D. Smith, *Subgroup Complexes*. Mathematical Surveys and Monographs, vol. 179 (American Mathematical Society, Providence, 2011)
71. E. Sperner, Neuer Beweis für die Invarianz der Dimensionszahl und des Gebietes. Abh. Math. Sem. Hamburg **VI**, 265–272 (1928)
72. J. Stillwell, *Classical Topology and Combinatorial Group Theory*. Graduate Texts in Mathematics, vol. 72, 2nd edn. (Springer, New York, 1993)
73. F.E. Su, Borsuk–Ulam implies Brouwer: a direct construction. Am. Math. Mon. **104**, 855–859 (1997)
74. G. Tardos, Transversals of 2-intervals, a topological approach. Combinatorica **15**, 123–134 (1995)
75. C. Thomassen, The Jordan–Schönflies theorem and the classification of surfaces. Am. Math. Mon. **99**, 116–130 (1992)
76. A.C.-C. Yao, Monotone bipartite graph properties are evasive. SIAM J. Comput. **17**, 517–520 (1988)
77. G.M. Ziegler, Shelling polyhedral 3-balls and 4-polytopes. Discret. Comput. Geomet. **19**, 159–174 (1998)
78. G.M. Ziegler, Generalized Kneser coloring theorems with combinatorial proofs. Inventiones Math. **147**, 671–691 (2002)

Beyond the Borsuk–Ulam Theorem: The Topological Tverberg Story

Pavle V.M. Blagojević and Günter M. Ziegler

Dedicated to the memory of Jiří Matoušek.

Abstract Bárány's "topological Tverberg conjecture" from 1976 states that any continuous map of an N-simplex Δ_N to $\mathbb{R}^d$, for $N \geq (d+1)(r-1)$, maps points from r disjoint faces in Δ_N to the same point in $\mathbb{R}^d$. The proof of this result for the case when r is a prime, as well as some colored version of the same result, using the results of Borsuk–Ulam and Dold on the non-existence of equivariant maps between spaces with a free group action, were main topics of Matoušek's 2003 book "Using the Borsuk–Ulam theorem."

In this paper we show how advanced equivariant topology methods allow one to go beyond the prime case of the topological Tverberg conjecture.

First we explain in detail how equivariant cohomology tools (employing the Borel construction, comparison of Serre spectral sequences, Fadell–Husseini index, etc.) can be used to prove the topological Tverberg conjecture whenever r is a prime power. Our presentation includes a number of improved proofs as well as

The research by Pavle V. M. Blagojević leading to these results has received funding from DFG via Collaborative Research Center TRR 109 "Discretization in Geometry and Dynamics." Also supported by the grant ON 174008 of the Serbian Ministry of Education and Science.

The research by Günter M. Ziegler received funding from DFG via the Research Training Group "Methods for Discrete Structures" and the Collaborative Research Center TRR 109 "Discretization in Geometry and Dynamics."

P.V.M. Blagojević
Institute of Mathematics, FU Berlin, Arnimallee 2, 14195 Berlin, Germany

Mathematical Institute SANU, Knez Mihailova 36, 11001 Beograd, Serbia
e-mail: blagojevic@math.fu-berlin.de

G.M. Ziegler (✉)
Institute of Mathematics, FU Berlin, Arnimallee 2, 14195 Berlin, Germany
e-mail: ziegler@math.fu-berlin.de

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_11

new results, such as a complete determination of the Fadell–Husseini index of chessboard complexes in the prime case.

Then, we introduce the "constraint method," which applied to suitable "unavoidable complexes" yields a great variety of variations and corollaries to the topological Tverberg theorem, such as the "colored" and the "dimension-restricted" (Van Kampen–Flores type) versions.

Both parts have provided crucial components to the recent spectacular counterexamples in high dimensions for the case when r is not a prime power.

1 Introduction

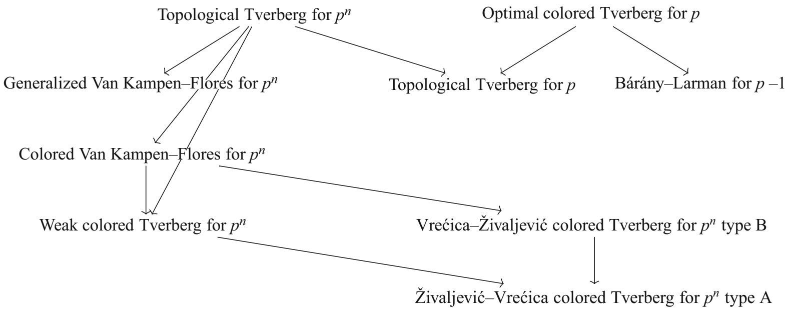
Jiří Matoušek's 2003 book "*Using the Borsuk–Ulam Theorem: Lectures on Topological Methods in Combinatorics and Geometry*" [34] is an inspiring introduction to the use of equivariant methods in Discrete Geometry. Its main tool is the Borsuk–Ulam theorem, and its generalization by Albrecht Dold, which says that there is no equivariant map from an n-connected space to an n-dimensional finite complex that is equivariant with respect to a non-trivial finite group acting freely. One of the main applications of this technology in Matoušek's book was a proof for Bárány's "topological Tverberg conjecture" on r-fold intersections in the case when r is a prime, originally due to Imre Bárány, Senya Shlosman and András Szűcs [8]. This conjecture claimed that for any continuous map $f : \Delta_N \to \mathbb{R}^d$, when $N \geq (d+1)(r-1)$, there are r points in disjoint faces of the simplex Δ_N that f maps to the same point in $\mathbb{R}^d$.

The topological Tverberg conjecture was extended to the case when r is a prime power by Murad Özaydin in an unpublished paper from 1987 [36]. This cannot, however, be achieved via the Dold theorem, since in the prime power case the group actions one could use on the codomain are not free. So more advanced methods are needed, such as the Serre spectral sequence for the Borel construction and the Fadell–Husseini index. In this paper we present the area about and around the topological Tverberg conjecture, with complete proofs for all of the results, which include the prime power case of the topological Tverberg conjecture.

Özaydin in 1987 not only proved the topological Tverberg theorem for prime power r, but he also showed, using equivariant obstruction theory, that the approach fails when r is not a prime power: In this case the equivariant map one looks for does exist.

In a spectacular recent development, Isaac Mabillard and Uli Wagner [31, 32] have developed an r-fold version of the classical "Whitney trick" (cf. [50]), which yields the failure of the generalized Van Kampen–Flores theorem when $r \geq 6$ is not a prime power. Then Florian Frick observed that this indeed implies the existence of counterexamples to the topological Tverberg conjecture [13, 25] by a lemma of Gromov [26, p. 445] that is an instance of the constraint method of Blagojević, Frick and Ziegler [12, Lemma 4.1(iii) and Lemma 4.2]. (See [5] for a popular rendition of the story.)

The Tverberg theorem from 1966 [45] and its conjectured extension to continuous maps (the topological Tverberg conjecture) have seen a great number of variations and extensions, among them "colored" variants as well as versions with restricted dimensions (known as generalized Van Kampen–Flores theorems). Although many of these were first obtained as independent results, sometimes with very similar proof patterns, our presentation shows that there are many easy implications between these results, using in particular the "constraint method" applied to "unavoidable complexes," as developed by the present authors with Florian Frick [12]. (Mikhail Gromov [26, p. 445] had sketched one particular instance: The topological Tverberg theorem for maps to $\mathbb{R}^{n+1}$ implies a generalized Van Kampen–Flores theorem for maps to $\mathbb{R}^n$.) Thus we can summarize the implications in the following scheme, which shows that all further main results follow from two sources, the topological Tverberg theorem for prime powers, and the optimal colored Tverberg theorem of the present authors with Benjamin Matschke [17], which up to now even for affine maps is available only for the prime case:


Topological Tverberg for p^n
Optimal colored Tverberg for p
Generalized Van Kampen–Flores for p^n
Topological Tverberg for p
Bárány–Larman for $p-1$
Colored Van Kampen–Flores for p^n
Weak colored Tverberg for p^n
Vrećica–Živaljević colored Tverberg for p^n type B
Živaljević–Vrećica colored Tverberg for p^n type A

Our journey in this paper starts with Radon's 1921 theorem and its topological version, in Sect. 2. Here the Borsuk–Ulam theorem is all that's needed. In Sect. 3 we state the topological Tverberg conjecture and first prove it in the prime case (with a proof that is close to the original argument by Bárány, Shlosman and Szűcs), and then for prime powers—this is where we go "beyond the Borsuk–Ulam theorem." Implications and corollaries of the topological Tverberg theorem are developed in Sect. 4—so that's where we put constraints, and "add color." In Sect. 5 we get to the counterexamples. And finally in Sect. 6 we discuss the "optimal colored Tverberg conjecture," which is a considerable strengthening of Tverberg's theorem, but up to now has been proven only in the prime case.

A summary of the main topological concepts and tools used in this paper is given at the end in the form of a dictionary, where a reference to the dictionary in the text is indicated by concept $^{\text{dict}}$.

2 The Beginning

2.1 Radon's Theorem

One of the first cornerstone results of convex geometry is a 1921 theorem of Johann Radon about overlapping convex hulls of points in a Euclidean space.

Let $\mathbb{R}^d$ be a d-dimensional Euclidean space. Let $\mathbf{x}_1, \dots, \mathbf{x}_m$ be points in $\mathbb{R}^d$ and let $\alpha_1, \dots, \alpha_m$ be non-negative real numbers that sum up to 1, that is, $\alpha_1 \geq 0, \dots, \alpha_m \geq 0$ and $\alpha_1 + \cdots + \alpha_m = 1$. The *convex combination* of the points $\mathbf{x}_1, \dots, \mathbf{x}_m$ determined by the scalars $\alpha_1, \dots, \alpha_m$ is the following point in $\mathbb{R}^d$:

$$\mathbf{x} = \alpha_1 \mathbf{x}_1 + \cdots + \alpha_m \mathbf{x}_m.$$

For a subset C of $\mathbb{R}^d$ we define the *convex hull* of C, denoted by $\operatorname{conv}(C)$, to be the set of all convex combinations of finitely many points in C:

$$\operatorname{conv}(C) := \{\alpha_1 \mathbf{x}_1 + \cdots + \alpha_m \mathbf{x}_m : m \in \mathbb{N},\, \mathbf{x}_i \in C,\, \alpha_i \in \mathbb{R}_{\geq 0},\, \alpha_1 + \cdots + \alpha_m = 1\}.$$

Now Radon's theorem can be stated as follows and proved using elementary linear algebra.

Theorem 2.1 (Radon's theorem, point configuration version [37]) *Let $\mathbb{R}^d$ be a d-dimensional Euclidean space, and let $X \subseteq \mathbb{R}^d$ be a subset with (at least) $d + 2$ elements. Then there are disjoint subsets P and N of X with the property that*

$$\operatorname{conv}(P) \cap \operatorname{conv}(N) \neq \emptyset.$$

Proof Let $X = \{\mathbf{x}_1, \dots, \mathbf{x}_{d+2}\} \subset \mathbb{R}^d$. The homogeneous system of $d + 1$ linear equations in $d + 2$ variables

$$\alpha_1 \mathbf{x}_1 + \cdots + \alpha_{d+2} \mathbf{x}_{d+2} = 0, \qquad \alpha_1 + \cdots + \alpha_{d+2} = 0$$

has a non-trivial solution, say $\alpha_1 = a_1, \dots, \alpha_{d+2} = a_{d+2}$. Denote

$$P := \{i : a_i > 0\} \qquad \text{and} \qquad N := \{i : a_i \leq 0\}.$$

Then $P \cap N = \emptyset$ while $P \neq \emptyset$ and $N \neq \emptyset$, and

$$a_1 \mathbf{x}_1 + \cdots + a_{d+2} \mathbf{x}_{d+2} = 0 \Rightarrow \sum_{i \in P} a_i \mathbf{x}_i = \sum_{i \in N} -a_i \mathbf{x}_i,$$

$$a_1 + \cdots + a_{d+2} = 0 \Rightarrow \sum_{i \in P} a_i = \sum_{i \in N} -a_i =: A,$$

where $A > 0$. Consequently, the following point is in the intersection of convex hulls of P and N:

$$\mathbf{x} := \sum_{i\in P} \frac{a_i}{A}\mathbf{x}_i = \sum_{i\in N} \frac{-a_i}{A}\mathbf{x}_i \in \operatorname{conv}(\{\mathbf{x}_i : i \in P\}) \cap \operatorname{conv}(\{\mathbf{x}_i : i \in N\}).$$

□

In order to reformulate Radon's theorem we recall the notion of an affine map. A map $f : D \to \mathbb{R}^d$ defined on a subset $D \subseteq \mathbb{R}^k$ is *affine* if for every $m \in \mathbb{N}$, $\mathbf{x}_1, \dots, \mathbf{x}_m \in D$, and $\alpha_1, \dots, \alpha_m \in \mathbb{R}$ with $\alpha_1 + \cdots + \alpha_m = 1$ and $\alpha_1\mathbf{x}_1 + \cdots + \alpha_m\mathbf{x}_m \in D$, we have

$$f(\alpha_1\mathbf{x}_1 + \cdots + \alpha_m\mathbf{x}_m) = \alpha_1 f(\mathbf{x}_1) + \cdots + \alpha_m f(\mathbf{x}_m).$$

Here and in the following let $\Delta_k := \operatorname{conv}\{\mathbf{e}_1, \dots, \mathbf{e}_{k+1}\}$ be the *standard k-dimensional simplex*: This simplex given as the convex hull of the standard basis of $\mathbb{R}^{k+1}$ has the disadvantage of not being full-dimensional in $\mathbb{R}^k$, but it has the extra advantage of being obviously symmetric (with symmetry given by permutation of coordinates). With this, Radon's theorem can be restated as follows (Fig. 1).

Theorem 2.2 (Radon's theorem, affine map version) *Let $f : \Delta_{d+1} \to \mathbb{R}^d$ be an affine map. Then there are disjoint faces σ_1 and σ_2 of the $(d+1)$-simplex Δ_{d+1} with the property that*

$$f(\sigma_1) \cap f(\sigma_2) \neq \emptyset.$$

With this version of Radon's theorem at hand, it is natural to ask: *Would Radon's theorem still hold if instead of an affine map we consider an arbitrary continuous map $f : \Delta_{d+1} \to \mathbb{R}^d$?*

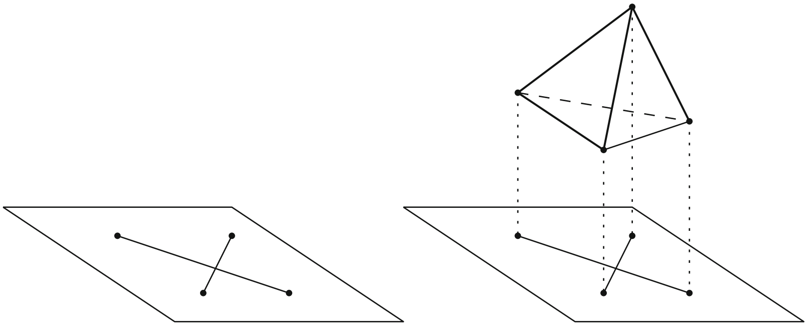


Fig. 1 Illustration of Radon's theorem in the plane for both versions of the theorem

2.2 The Topological Radon Theorem

The question we have just asked was answered in 1979 by Ervin Bajmóczy and Imre Bárány [4], using the Borsuk–Ulam theorem.

Theorem 2.3 (Topological Radon theorem) *Let $f : \Delta_{d+1} \to \mathbb{R}^d$ be any continuous map. Then there are two disjoint faces σ_1 and σ_2 of Δ_{d+1} whose images under f intersect,*

$$f(\sigma_1) \cap f(\sigma_2) \neq \emptyset.$$

Proof Let $\Delta_{d+1} = \operatorname{conv}\{\mathbf{e}_1, \ldots, \mathbf{e}_{d+2}\}$ be the standard simplex. Consider the subcomplex X of the polyhedral complex $\Delta_{d+1} \times \Delta_{d+1}$ given by

$$\begin{aligned} X := \{(x_1, x_2) \in \Delta_{d+1} \times \Delta_{d+1} : &\text{ there are faces } \sigma_1, \sigma_2 \subset \Delta_{d+1} \\ &\text{such that } \sigma_1 \cap \sigma_2 = \emptyset,\ x_1 \in \sigma_1,\ x_2 \in \sigma_2\}. \end{aligned}$$

The group $\mathbb{Z}/2 = \langle \varepsilon \rangle$ acts freely on X by $\varepsilon \cdot (x_1, x_2) = (x_2, x_1)$.

Let us assume that the theorem does not hold. Then there exists a continuous map $f : \Delta_{d+1} \to \mathbb{R}^d$ such that $f(x_1) \neq f(x_2)$ for all $(x_1, x_2) \in X$. Consequently the map $g : X \to S^{d-1}$ given by

$$g(x_1, x_2) := \frac{f(x_1) - f(x_2)}{\|f(x_1) - f(x_2)\|},$$

is continuous and $\mathbb{Z}/2$-equivariant, where the action on $S^{d-1} = S(\mathbb{R}^d)$, the unit sphere in $\mathbb{R}^d$, is the standard antipodal action.

Next we define a continuous $\mathbb{Z}/2$-equivariant map from a d-sphere to X. For this we do not use the standard d-sphere, but the unit sphere $S(W_{d+2})$ in the hyperplane $W_{d+2} := \{(a_1, \ldots, a_{d+2}) \in \mathbb{R}^{d+2} : a_1 + \cdots + a_{d+2} = 0\} \subset \mathbb{R}^{d+2}$, that is,

$$S(W_{d+2}) = \{(a_1, \ldots, a_{d+2}) \in \mathbb{R}^{d+2} : a_1 + \cdots + a_{d+2} = 0,\ a_1^2 + \cdots + a_{d+2}^2 = 1\}.$$

This representation of the d-sphere also has the standard antipodal $\mathbb{Z}/2$-action. The map $h : S(W_{d+2}) \to X$ is defined by

$$h(a_1, \ldots, a_{d+2}) := \Big(\sum_{a_i \geq 0} \frac{a_i}{A} \mathbf{e}_i, \sum_{a_i < 0} \frac{-a_i}{A} \mathbf{e}_i \Big),$$

where $A := \sum_{a_i > 0} a_i = -\sum_{a_i < 0} a_i > 0$. This is easily checked to be well-defined and continuous; the image point lies in the cell $\operatorname{conv}\{\mathbf{e}_i : a_i > 0\} \times \operatorname{conv}\{\mathbf{e}_j : a_j < 0\}$ of the complex $\Delta_{d+1} \times \Delta_{d+1}$.

The composition map $g \circ h : S(W_{d+2}) \to S^{d-1}$ yields a continuous $\mathbb{Z}/2$-equivariant map from a free d-sphere to a free $(d-1)$-sphere that contradicts the Borsuk–Ulam theorem dict. Thus the theorem holds. □

2.3 The Van Kampen–Flores Theorem

The topological Radon theorem guarantees that for every continuous map $\Delta_{d+1} \to \mathbb{R}^d$ there exist two pairwise disjoint faces whose f-images overlap. It is natural to ask: *Is it possible to say something about the dimension of the disjoint faces whose f-images intersect?* In the spirit of Poincaré's classification of mathematical problems [2, Lec. 1] this *binary problem* has a quick answer *no*, but if understood as an *interesting problem* it has an answer: If we are willing to spend an extra vertex/dimension, meaning, put the simplex Δ_{d+2} in place of Δ_{d+1}, we get the following theorem from the 1930s of Egbert R. Van Kampen and Antonio Flores [23, 46].

Theorem 2.4 (Van Kampen–Flores theorem) *Let $d \geq 2$ be an even integer, and let $f : \Delta_{d+2} \to \mathbb{R}^d$ be a continuous map. Then there are disjoint faces σ_1 and σ_2 of Δ_{d+2} of dimension at most $d/2$ whose images under f intersect,*

$$f(\sigma_1) \cap f(\sigma_2) \neq \emptyset.$$

Proof Let $g : \Delta_{d+2} \to \mathbb{R}^{d+1}$ be a continuous map defined by

$$g(x) := \big(f(x), \operatorname{dist}(x, \operatorname{sk}_{d/2}(\Delta_{d+2}))\big)$$

where $\operatorname{sk}_{d/2}(\Delta_{d+2})$ denotes the $d/2$-skeleton of the simplex Δ_{d+2}, and $\operatorname{dist}(x, \operatorname{sk}_{d/2}(\Delta_{d+2}))$ is the distance of the point x from the subcomplex $\operatorname{sk}_{d/2}(\Delta_{d+2})$. Observe that if $x \in \operatorname{relint} \sigma$ and $\operatorname{dist}(x, \operatorname{sk}_{d/2}(\Delta_{d+2})) = 0$, then the simplex σ belongs to the subcomplex $\operatorname{sk}_{d/2}(\Delta_{d+2})$.

Now the topological Radon theorem can be applied to the continuous map $g : \Delta_{d+2} \to \mathbb{R}^{d+1}$. It yields the existence of points $x_1 \in \operatorname{relint} \sigma_1$ and $x_2 \in \operatorname{relint} \sigma_2$, with $\sigma_1 \cap \sigma_2 = \emptyset$, such that $g(x_1) = g(x_2)$, that means,

$$f(x_1) = f(x_2) \quad \text{and} \quad \operatorname{dist}(x_1, \operatorname{sk}_{d/2}(\Delta_{d+2})) = \operatorname{dist}(x_2, \operatorname{sk}_{d/2}(\Delta_{d+2})).$$

If one of the simplices σ_1, or σ_2, would belong to $\operatorname{sk}_{d/2}(\Delta_{d+2})$, then

$$\operatorname{dist}(x_1, \operatorname{sk}_{d/2}(\Delta_{d+2})) = \operatorname{dist}(x_2, \operatorname{sk}_{d/2}(\Delta_{d+2})) = 0$$

implying that both σ_1 and σ_2 belong to $\operatorname{sk}_{d/2}(\Delta_{d+2})$, which would concludes the proof of the theorem.

In order to prove that at least one of the faces σ_1 and σ_2 belongs to $\mathrm{sk}_{d/2}(\Delta_{d+2})$, note that these are two disjoint faces of the simplex Δ_{d+2}, which has $d+3$ vertices, so by the pigeonhole principle one of them has at most $\lfloor(d+3)/2\rfloor = d/2+1$ vertices. □

The proof we have presented is an example of the *constraint method* developed in [12]. An important message of this proof is that the Van Kampen–Flores theorem is a corollary of the topological Radon theorem. It is clear that we could have considered a continuous map f defined only on the $d/2$-skeleton.

All the results we presented so far have always claimed something about intersections of the images of two disjoint faces σ_1 and σ_2, which we refer to as 2-fold overlap, or intersection. *What about r-fold overlaps, for $r > 2$?*

3 The Topological Tverberg Theorem

3.1 The Topological Tverberg Conjecture

In 1964, freezing in a hotel room in Manchester, the Norwegian mathematician Helge Tverberg proved the following r-fold generalization of Radon's theorem [45]. It had been conjectured by Bryan Birch in 1954, who had established the result in the special case of dimension $d = 2$ [9]. The case $d = 1$ is easy, see below. (See [51] for some of the stories surrounding these discoveries.)

Theorem 3.1 (Tverberg's theorem) *Let $d \geq 1$ and $r \geq 2$ be integers, $N = (d+1)(r-1)$, and let $f : \Delta_N \to \mathbb{R}^d$ be an affine map. Then there exist r pairwise disjoint faces $\sigma_1, \dots, \sigma_r$ of the simplex Δ_N whose f-images overlap,*

$$f(\sigma_1) \cap \cdots \cap f(\sigma_r) \neq \emptyset. \tag{1}$$

Any collection of r pairwise disjoint faces $\sigma_1, \dots, \sigma_r$ of the simplex Δ_N having property (1) is called a *Tverberg partition* of the map f.

The dimension of the simplex in the theorem is optimal, it cannot be decreased. To see this consider the affine map $h : \Delta_{N-1} \to \mathbb{R}^d$ given on the vertices of $\Delta_{N-1} = \mathrm{conv}\{\mathbf{e}_1, \dots, \mathbf{e}_N\}$ by

$$\mathbf{e}_i \overset{h}{\longmapsto} u_{\lfloor (i-1)/(r-1) \rfloor} \tag{2}$$

where $\{u_0, \dots, u_d\}$ is an affinely independent set in $\mathbb{R}^d$, e.g., $(u_0, \dots, u_d) = (0, \mathbf{e}_1, \dots, \mathbf{e}_d)$. For each vertex of the simplex $\mathrm{conv}\{u_0, \dots, u_d\}$ the cardinality of its preimage is $r-1$

$$|h^{-1}(\{u_0\})| = \cdots = |h^{-1}(\{u_d\})| = r-1,$$

and so the map h has no Tverberg partition. Even more is true: Any affine map $h : \Delta_{N-1} \to \mathbb{R}^d$ that is in general position cannot have a Tverberg partition.

As in the case of Radon's theorem it is natural to ask: *Would the Tverberg theorem still hold if instead of an affine map $f : \Delta_N \to \mathbb{R}^d$ we would consider an arbitrary continuous map?* This was first asked by Bárány in a 1976 letter to Tverberg. In May of 1978 Tverberg posed the question in Oberwolfach, stating it for a general N-polytope in place of the N-simplex, see [27]. (The problem for a general N-polytope can be reduced to the case of the N-simplex by a theorem of Grünbaum: Every N-polytope as a cell complex is a refinement of the N-simplex [28, p. 200].) Thus, the topological Tverberg conjecture started its life in the late 1970s.

Conjecture 3.2 (Topological Tverberg conjecture) *Let $d \geq 1$ and $r \geq 2$ be integers, $N = (d+1)(r-1)$, and let $f : \Delta_N \to \mathbb{R}^d$ be a continuous map. Then there exist r pairwise disjoint faces $\sigma_1, \ldots, \sigma_r$ of the simplex Δ_N whose f-images overlap,*

$$f(\sigma_1) \cap \cdots \cap f(\sigma_r) \neq \emptyset. \tag{3}$$

The case $r = 2$ of the topological Tverberg conjecture amounts to the topological Radon theorem, so it holds. The topological Tverberg conjecture is also easy to verify for $d = 1$, as follows.

Theorem 3.3 (Topological Tverberg theorem for $d = 1$) *Let $r \geq 2$ be an integer, and let $f : \Delta_{2r-2} \to \mathbb{R}$ be a continuous map. Then there exist r pairwise disjoint faces $\sigma_1, \ldots, \sigma_r$ of the simplex Δ_{2r-2} whose f-images overlap,*

$$f(\sigma_1) \cap \cdots \cap f(\sigma_r) \neq \emptyset.$$

Proof Let $f : \Delta_{2r-2} \to \mathbb{R}$ be continuous. Sort the vertices of the simplex $\Delta_{2r-2} = \operatorname{conv}\{\mathbf{e}_1, \ldots, \mathbf{e}_{2r-1}\}$ such that $f(\mathbf{e}_{\pi(1)}) \leq f(\mathbf{e}_{\pi(2)}) \leq \cdots \leq f(\mathbf{e}_{\pi(2r-2)}) \leq f(\mathbf{e}_{\pi(2r-1)})$. Then the collection of $r-1$ edges and one vertex of Δ_{2r-2}

$$\sigma_1 = [\mathbf{e}_{\pi(1)}, \mathbf{e}_{\pi(2r-1)}],\ \sigma_2 = [\mathbf{e}_{\pi(2)}, \mathbf{e}_{\pi(2r-2)}], \ldots,\ \sigma_{r-1} = [\mathbf{e}_{\pi(r-1)}, \mathbf{e}_{\pi(r+1)}],$$
$$\sigma_r = \{\mathbf{e}_r\}$$

is a Tverberg partition for the map f. □

At first glance the topological Tverberg conjecture is a *binary problem* in the Poincaré classification of mathematical problems. To our surprise it is safe to say, at this point in time, that the topological Tverberg conjecture was one of the most *interesting problems* that shaped interaction between Geometric Combinatorics on one hand and Algebraic and Geometric Topology on the other hand for almost four decades.

After settling the topological Tverberg conjecture for $d = 1$ and $r = 2$ we want to advance. How?

3.2 Equivariant Topology Steps in

Let $d \geq 1$ and $r \geq 2$ be integers, and let $N = (d+1)(r-1)$. Our effort to handle the topological Tverberg conjecture starts with an assumption that there is a counterexample to the conjecture with parameters d and r. Thus there is a continuous map $f : \Delta_N \to \mathbb{R}^d$ such that for every r-tuple $\sigma_1, \dots, \sigma_r$ of pairwise disjoint faces of the simplex Δ_N their f-images do not intersect, that is,

$$f(\sigma_1) \cap \cdots \cap f(\sigma_r) = \emptyset. \tag{4}$$

In order to capture this feature of our counterexample f we parametrize all r-tuples of pairwise disjoint faces of the simplex Δ_N. This can be done in two similar, but different ways.

3.2.1 The *r*-fold 2-wise Deleted Product

The *r-fold* 2*-wise deleted product* dict of a simplicial complex K is the cell complex

$$K^{\times r}_{\Delta(2)} := \{(x_1, \dots, x_r) \in \sigma_1 \times \cdots \times \sigma_r \subset K^{\times r} : \sigma_i \cap \sigma_j = \emptyset \text{ for } i \neq j\},$$

where $\sigma_1, \dots, \sigma_r$ are non-empty faces of K. The symmetric group $\mathfrak{S}_r$ acts (from the left) on $K^{\times r}_{\Delta(2)}$ by

$$\pi \cdot (x_1, \dots, x_r) := (x_{\pi^{-1}(1)}, \dots, x_{\pi^{-1}(r)}),$$

for $\pi \in \mathfrak{S}_r$ and $(x_1, \dots, x_r) \in K^{\times r}_{\Delta(2)}$. This action is free due to the fact that if $(x_1, \dots, x_r) \in K^{\times r}_{\Delta(2)}$ then $x_i \neq x_j$ for all $i \neq j$. We have seen a particular instance before: The complex X that we used in the proof of the topological Radon Theorem 2.3 was $(\Delta_{d+1})^{\times 2}_{\Delta(2)}$. For more details on the deleted product construction see for example [8] or [34, Sec. 6.3].

In the case when K is a simplex the topology of the deleted product $K^{\times r}_{\Delta(2)}$ is known from the following result of Bárány, Shlosman and Szűcs [8, Lem. 1].

Theorem 3.4 *Let N and r be positive integers with $N \geq r-1$. Then $(\Delta_N)^{\times r}_{\Delta(2)}$ is an $(N-r+1)$-dimensional and $(N-r)$-connected CW complex.*

Proof A typical face of the CW complex $(\Delta_N)^{\times r}_{\Delta(2)}$ is of the form $\sigma_1 \times \cdots \times \sigma_r$, where $\sigma_1, \dots, \sigma_r$ are pairwise disjoint simplices. Consequently, the number of vertices of these simplices together cannot exceed $N+1$, or in the language of dimension

$$\dim(\sigma_1) + 1 + \cdots + \dim(\sigma_r) + 1 \leq N + 1.$$

The equality is attained when all the vertices are used, this is when $\sigma_1, \dots, \sigma_r$ is a maximal face of dimension

$$\dim(\sigma_1 \times \cdots \times \sigma_r) = \dim(\sigma_1) + \cdots + \dim(\sigma_r) = N - r + 1.$$

Thus $(\Delta_N)^{\times r}_{\Delta(2)}$ is an $(N - r + 1)$-dimensional CW complex.

For $N = r - 1$ the deleted product $(\Delta_N)^{\times r}_{\Delta(2)}$ is the 0-dimensional simplicial complex $[r]$ and the statement of the theorem holds. Thus, we can assume that $N \geq r$.

For $N \geq r$ we establish the connectivity of the deleted product of a simplex by induction on r making repeated use of the following classical 1957 theorem of Stephen Smale [41, Main Thm.]:

Smale's Theorem *Let X and Y be connected, locally compact, separable metric spaces, and in addition let X be locally contractible. Let $f : X \to Y$ be a continuous surjective proper map, that is, any preimage of a compact set is compact. If for every $y \in Y$ the preimage $f^{-1}(\{y\})$ is locally contractible and n-connected, then the induced homomorphism*

$$f_\# : \pi_i(X) \to \pi_i(Y)$$

is an isomorphism for all $0 \leq i \leq n$, and is an epimorphism for $i = n + 1$.

Recall that Δ_N denotes the standard simplex, whose vertices $\mathbf{e}_1, \dots, \mathbf{e}_{N+1}$ form the standard basis of $\mathbb{R}^{N+1}$. The induction starts with $r = 1$ and the theorem claims that the simplex Δ_N, a contractible space, is $(N - 1)$-connected, which is obviously true.

In the case $r = 2$ consider the surjection $p_1 : (\Delta_N)^{\times 2}_{\Delta(2)} \to \mathrm{sk}_{N-1}(\Delta_N)$ given by the projection on the first factor. Any point x_1 of the $(N - 1)$-skeleton $\mathrm{sk}_{N-1}(\Delta_N)$ of the simplex Δ_N lies in the relative interior of a face,

$$x_1 \in \operatorname{relint}\big(\operatorname{conv}\{\mathbf{e}_i : i \in T \subseteq [N + 1]\}\big)$$

where $1 \leq |T| \leq N$. Let us denote the complementary set of vertices by $S := \{\mathbf{e}_i : i \notin T\} \neq \emptyset$ and its convex hull by $\Delta_S := \operatorname{conv}(S) \cong \Delta_{|S|-1}$. The fiber of the projection map p_1 over x_1 is given by

$$p_1^{-1}(\{x_1\}) = \{(x_1, x_2) \in (\Delta_N)^{\times 2}_{\Delta(2)} : x_2 \in \Delta_S\} \cong \Delta_S,$$

and consequently it is contractible. By Smale's theorem the projection p_1 induces an isomorphism between homotopy groups. Since we are working in the category of CW complexes the Whitehead theorem [19, Thm. 11.2] implies a homotopy equivalence of $(\Delta_N)^{\times 2}_{\Delta(2)}$ and $\mathrm{sk}_{N-1}(\Delta_N)$. The $(N - 1)$-skeleton of a simplex is $(N - 2)$-connected and thus the theorem holds in the case $r = 2$.

For the induction hypothesis assume that $(\Delta_N)^{\times i}_{\Delta(2)}$ is $(N - i)$-connected for all $i \leq k < r$. In the induction step we want to prove that $(\Delta_N)^{\times(k+1)}_{\Delta(2)}$ is $(N - k - 1)$-connected.

Now consider the projection onto the first k factors,

$$p_k : (\Delta_N)^{\times(k+1)}_{\Delta(2)} \to \mathrm{sk}_{N-k}\left((\Delta_N)^{\times k}_{\Delta(2)}\right).$$

Since $(\Delta_N)^{\times k}_{\Delta(2)}$ is $(N-k)$-connected by induction hypothesis, its $(N-k)$-skeleton $\mathrm{sk}_{N-k}\left((\Delta_N)^{\times k}_{\Delta(2)}\right)$ is $(N-k-1)$-connected. For a typical point of the codomain we have that

$$\begin{aligned}(x_1,\dots,x_k) \in \mathrm{relint}\big(&\mathrm{conv}\{\mathbf{e}_i : i \in T_1 \subseteq [N+1]\}\big)\\ &\times\cdots\times \mathrm{relint}\big(\mathrm{conv}\{\mathbf{e}_i : i \in T_k \subseteq [N+1]\}\big),\end{aligned}$$

where $T_i \cap T_j = \emptyset$ for all $1 \le i < j \le k$, and $|T_1| - 1 + \cdots + |T_k| - 1 \le N - k$. As before, consider the complementary set of vertices $S := \{\mathbf{e}_i : i \notin T_1 \cup \cdots \cup T_k\} \neq \emptyset$ and its convex hull $\Delta_S = \mathrm{conv}(S) \cong \Delta_{|S|-1}$. The fiber of the projection map p_k over $(x_1,\dots,x_k)$ is given by

$$p_k^{-1}(\{(x_1,\dots,x_k)\}) = \{(x_1,\dots,x_k,x_{k+1}) \in (\Delta_N)^{\times(k+1)}_{\Delta(2)} : x_{k+1} \in \Delta_S\} \cong \Delta_S,$$

so it is contractible. Again Smale's theorem applied to the projection p_k induces an isomorphism between homotopy groups of $(\Delta_N)^{\times(k+1)}_{\Delta(2)}$ and $\mathrm{sk}_{N-k}\left((\Delta_N)^{\times k}_{\Delta(2)}\right)$. Moreover, the Whitehead theorem implies that these spaces are homotopy equivalent. Since, $\mathrm{sk}_{N-k}\left((\Delta_N)^{\times k}_{\Delta(2)}\right)$ is $(N-k-1)$-connected we have concluded the induction step and the theorem is proved. □

Remark 3.5 Our proof of Theorem 3.4 may be traced back to a proof in the lost preprint version of the paper [8]. Indeed, in the published version the first sentence of [8, Proof of Lem. 1] says:

> For this elementary proof we are indebted to the referee. Our original proof used the Leray spectral sequence.

Here we used Smale's theorem in place of the Leray spectral sequence argument.

3.2.2 The *r*-fold *k*-wise Deleted Join

Let K be a simplicial complex. The *r-fold k-wise deleted join* dict of the simplicial complex K is the simplicial complex

$$\begin{aligned}K^{*r}_{\Delta(k)} := \big\{&\lambda_1 x_1 + \cdots + \lambda_r x_r \in \sigma_1 * \cdots * \sigma_r \subset K^{*r} :\\ &(\forall I \subset [n])\, \mathrm{card}\, I \ge k \Rightarrow \bigcap_{i\in I} \sigma_i = \emptyset\big\},\end{aligned}$$

where $\sigma_1, \dots, \sigma_n$ are faces of K, including the empty face. Thus in the case $k = 2$ we have

$$K^{*r}_{\Delta(2)} := \{\lambda_1 x_1 + \cdots + \lambda_r x_r \in \sigma_1 * \cdots * \sigma_r \subset K^{*r} : \sigma_i \cap \sigma_j = \emptyset \text{ for } i \neq j\}.$$

The symmetric group $\mathfrak{S}_r$ acts (from the left) on $K^{*r}_{\Delta(2)}$ as follows

$$\pi \cdot (\lambda_1 x_1 + \cdots + \lambda_r x_r) := \lambda_{\pi^{-1}(1)} x_{\pi^{-1}(1)} + \cdots + \lambda_{\pi^{-1}(r)} x_{\pi^{-1}(r)},$$

where $\pi \in \mathfrak{S}_n$ and $\lambda_1 x_1 + \cdots + \lambda_r x_r \in K^{*r}_{\Delta(2)}$. This action is free only in the case when $r = 2$.

Examples 3.6 (Compare Fig. 2)

(1) Let $K = \Delta_1$ be the 1-simplex. Then $K^{\times 2}_{\Delta(2)} = S^0$ while $K^{*2}_{\Delta(2)} \cong S^1$.
(2) For $K = S^0$ we have that $K^{\times 2}_{\Delta(2)} = S^0$, and $K^{*2}_{\Delta(2)}$ is a disjoint union of two intervals.
(3) If $K = [3]$ then $K^{*2}_{\Delta(2)} \cong S^1$.
(4) When $K = [k]$, the deleted join $K^{*r}_{\Delta(2)}$ is the $k \times r$ chessboard complex[dict], which is denoted by $\Delta_{k,r}$.

The following lemma establishes the commutativity of the join and the deleted join operations on simplicial complexes. We state it for k-wise deleted joins and prove it here only for 2-wise deleted joins. For more details and insight consult the sections *"Deleted Products Good"* and *"...Deleted Joins Better"* in Matoušek's book, [34, Sections. 5.4 and 5.5].

Lemma 3.7 *Let K and L be simplicial complexes, and let $n \geq 2$ and $k \geq 2$ be integers. There exists an isomorphism of simplicial complexes:*

$$(K * L)^{*n}_{\Delta(k)} \cong K^{*n}_{\Delta(k)} * L^{*n}_{\Delta(k)}.$$

Proof We give a proof only for the case $k = 2$. Let $\sigma_1, \dots, \sigma_n$ and $\tau_1, \dots, \tau_n$ be simplices in K and L, respectively, such that $\sigma_i \cap \sigma_j = \emptyset$ and $\tau_i \cap \tau_j = \emptyset$ for all $i \neq j$.

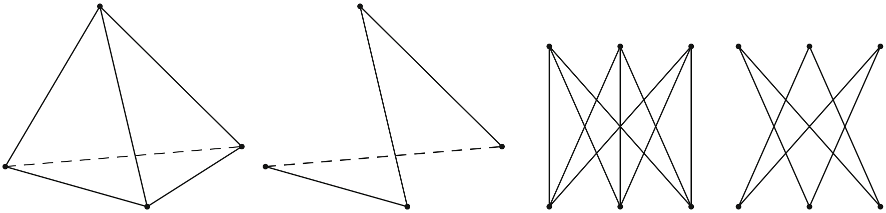


Fig. 2 The complexes K^{*2} and $K^{*2}_{\Delta(2)}$ for $K = \Delta_1$ and $K = [3]$

In addition, since the simplicial complexes K and L have disjoint vertex sets, we get that $\sigma_i \cap \tau_j = \emptyset$ as well for all i and j. Thus, for all $i \neq j$ we obtain an equivalence:

$$(\sigma_i \cup \tau_i) \cap (\sigma_j \cup \tau_j) = \emptyset \qquad \text{if and only if} \qquad \sigma_i \cap \sigma_j = \emptyset \text{ and } \tau_i \cap \tau_j = \emptyset.$$

It induces a bijection between the following simplices of $(K * L)^{*n}_{\Delta(2)}$ and $K^{*n}_{\Delta(2)} * L^{*n}_{\Delta(2)}$ by:

$$(\sigma_1 * \tau_1) *_{\Delta(2)} \cdots *_{\Delta(2)} (\sigma_n * \tau_n) \quad \longleftrightarrow \quad (\sigma_1 *_{\Delta(2)} \cdots *_{\Delta(2)} \sigma_n) * (\tau_1 *_{\Delta(2)} \cdots *_{\Delta(2)} \tau_n).$$

□

A direct consequence of the previous lemma is the following useful fact.

Lemma 3.8 *Let $r \geq 2$ and $2 \leq k \leq r$ be integers. Then*

(1) $(\Delta_N)^{*r}_{\Delta(2)} \cong [r]^{*(N+1)}$,

(2) $(\Delta_N)^{*r}_{\Delta(k)} \cong (\mathrm{sk}_{k-2}(\Delta_{r-1}))^{*(N+1)}$.

Proof $(\Delta_N)^{*r}_{\Delta(k)} \cong ([1]^{*(N+1)})^{*r}_{\Delta(k)} \cong ([1]^{*r}_{\Delta(k)})^{*(N+1)} \cong (\mathrm{sk}_{k-2}(\Delta_{r-1}))^{*(N+1)}$. □

Thus the r-fold 2-wise deleted join of an N-simplex $(\Delta_N)^{*r}_{\Delta(2)}$ is an N-dimensional and $(N-1)$-connected simplicial complex.

3.2.3 Equivariant Maps Induced by *f*

Recall that, at the beginning of Sect. 3.2, we have fixed integers $d \geq 1$ and $r \geq 2$, and in addition we assumed the existence of the continuous map $f : \Delta_N \to \mathbb{R}^d$ that is a counterexample to the topological Tverberg theorem.

Define continuous maps induced by f in the following way:

- the *product map* is

$$P_f : (\Delta_N)^{\times r}_{\Delta(2)} \to (\mathbb{R}^d)^{\times r} \cong (\mathbb{R}^d)^{\oplus r}, \qquad (x_1, \ldots, x_r) \longmapsto (f(x_1), \ldots, f(x_r));$$

- the *join map* is

$$\begin{aligned} J_f : (\Delta_N)^{*r}_{\Delta(2)} \to (\mathbb{R}^{d+1})^{\oplus r}, \qquad & \lambda_1 x_1 + \cdots + \lambda_r x_r \\ & \longmapsto (\lambda_1, \lambda_1 f(x_1)) \oplus \cdots \oplus (\lambda_r, \lambda_r f(x_r)). \end{aligned}$$

The codomains $(\mathbb{R}^d)^{\oplus r}$ and $(\mathbb{R}^{d+1})^{\oplus r}$ of the maps P_f and J_f are equipped with the action of the symmetric group $\mathfrak{S}_r$ given by permutation of the corresponding r factors, that is

$$\pi \cdot (y_1, \ldots, y_r) = (y_{\pi^{-1}(1)}, \ldots, y_{\pi^{-1}(r)}),$$
$$\pi \cdot (z_1, \ldots, z_r) = (z_{\pi^{-1}(1)}, \ldots, z_{\pi^{-1}(r)}),$$

for $(y_1, \dots, y_r) \in (\mathbb{R}^d)^{\oplus r}$ and $(z_1, \dots, z_r) \in (\mathbb{R}^{d+1})^{\oplus r}$. Then both maps P_f and J_f are $\mathfrak{S}_r$-equivariant. Indeed, the following diagrams commute:

$$\begin{tikzcd}
(x_1, \dots, x_r) \arrow[r, "P_f"] \arrow[d, "\pi\cdot"'] & (f(x_1), \dots, f(x_r)) \arrow[d, "\pi\cdot"'] \\
(x_{\pi^{-1}(1)}, \dots, x_{\pi^{-1}(r)}) \arrow[r, "P_f"] & (f(x_{\pi^{-1}(1)}), \dots, f(x_{\pi^{-1}(r)})),
\end{tikzcd}$$

and

$$\begin{tikzcd}
\lambda_1 x_1 + \cdots + \lambda_r x_r \arrow[r, "J_f"] \arrow[d, "\pi\cdot"'] & (\lambda_1, \lambda_1 f(x_1)) \oplus \cdots \oplus (\lambda_r, \lambda_r f(x_r)) \arrow[d, "\pi\cdot"'] \\
\lambda_{\pi^{-1}(1)} x_{\pi^{-1}(1)} + \cdots + \lambda_{\pi^{-1}(r)} x_{\pi^{-1}(r)} \arrow[r, "J_f"] & (\lambda_{\pi^{-1}(1)}, \lambda_{\pi^{-1}(1)} f(x_{\pi^{-1}(1)})) \oplus \cdots \oplus (\lambda_{\pi^{-1}(r)}, \lambda_{\pi^{-1}(r)} f(x_{\pi^{-1}(r)})).
\end{tikzcd}$$

The $\mathfrak{S}_r$-invariant subspaces

$$D_P := \{(y_1, \dots, y_r) \in (\mathbb{R}^d)^{\oplus r} : y_1 = \cdots = y_r\},$$
$$D_J := \{(z_1, \dots, z_r) \in (\mathbb{R}^{d+1})^{\oplus r} : z_1 = \cdots = z_r\},$$

of the codomains $(\mathbb{R}^d)^{\oplus r}$ and $(\mathbb{R}^{d+1})^{\oplus r}$, respectively, are called the *thin diagonals*. The crucial property of the maps P_f and J_f, for a counterexample continuous map $f : \Delta_N \to \mathbb{R}^d$, is that

$$\operatorname{im}(P_f) \cap D_P = \emptyset \qquad \text{and} \qquad \operatorname{im}(J_f) \cap D_J = \emptyset. \tag{5}$$

Indeed, the property (4) of the map f immediately implies that $\operatorname{im}(P_f)$ and D_P are disjoint. For the second relation of (5) assume that

$$(\lambda_1, \lambda_1 f(x_1)) \oplus \cdots \oplus (\lambda_r, \lambda_r f(x_r)) \in \operatorname{im}(J_f) \cap D_J \neq \emptyset$$

for some $\lambda_1 x_1 + \cdots + \lambda_r x_r \in (\Delta_N)^{*r}_{\Delta(2)}$. Then $\lambda_1 = \cdots = \lambda_r = \frac{1}{r}$ and consequently $f(x_1) = \cdots = f(x_r)$.

Therefore, the maps P_f and J_f induce $\mathfrak{S}_r$-equivariant maps

$$(\Delta_N)^{\times r}_{\Delta(2)} \to (\mathbb{R}^d)^{\oplus r} \backslash D_P \qquad \text{and} \qquad (\Delta_N)^{*r}_{\Delta(2)} \to (\mathbb{R}^{d+1})^{\oplus r} \backslash D_J \tag{6}$$

that, with an obvious abuse of notation, are again denoted by P_f and J_f, respectively. Let us denote by

$$R_P : (\mathbb{R}^d)^{\oplus r} \backslash D_P \to D_P^{\perp} \backslash \{0\} \to S(D_P^{\perp}),$$
$$R_J : (\mathbb{R}^{d+1})^{\oplus r} \backslash D_J \to D_J^{\perp} \backslash \{0\} \to S(D_J^{\perp}), \tag{7}$$

the compositions of projections and deformation retractions. Here $U^{\perp}$ denotes the orthogonal complement of the subspace U in the relevant ambient real vector space, while $S(V)$ denotes the unit sphere in the real vector space V. Both maps R_P and R_J are $\mathfrak{S}_r$-equivariant maps with respect to the introduced actions.

Furthermore, let $\mathbb{R}^r$ be a vectors space with the (left) action of the symmetric group $\mathfrak{S}_r$ given by the permutation of coordinates. Then the subspace $W_r = \big\{(t_1,\dots,t_r) \in \mathbb{R}^r : \sum_{i=1}^r t_i = 0\big\}$ is an $\mathfrak{S}_r$-invariant subspace of dimension $r-1$. There is an isomorphism of real $\mathfrak{S}_r$-representations

$$D_P^{\perp} \cong W_r^{\oplus d} \qquad \text{and} \qquad D_J^{\perp} \cong W_r^{\oplus (d+1)}.$$

Using this identification of $\mathfrak{S}_r$-representations the $\mathfrak{S}_r$-equivariant maps R_P and R_J, defined in (7), can be presented by

$$R_P : (\mathbb{R}^d)^{\oplus r}\backslash D_P \to S(W_r^{\oplus d}), \quad R_J : (\mathbb{R}^{d+1})^{\oplus r}\backslash D_J \to S(W_r^{\oplus (d+1)}). \tag{8}$$

Finally we have the theorem we were looking for. It will give us a chance to employ methods of algebraic topology to attack the topological Tverberg conjecture.

Theorem 3.9 *Let $d \geq 1$ and $r \geq 2$ be integers, and let $N = (d+1)(r-1)$. If there exists a counterexample to the topological Tverberg conjecture, then there exist $\mathfrak{S}_r$-equivariant maps*

$$(\Delta_N)^{\times r}_{\Delta(2)} \to S(W_r^{\oplus d}) \qquad \textit{and} \qquad (\Delta_N)^{*r}_{\Delta(2)} \to S(W_r^{\oplus (d+1)}).$$

Proof If $f : \Delta_N \to \mathbb{R}^d$ is a counterexample to the topological Tverberg conjecture, then by composing maps from (6) and (8) we get $\mathfrak{S}_r$-equivariant maps

$$R_P \circ P_f : (\Delta_N)^{\times r}_{\Delta(2)} \to S(W_r^{\oplus d}) \qquad \text{and} \qquad R_J \circ J_f : (\Delta_N)^{*r}_{\Delta(2)} \to S(W_r^{\oplus (d+1)}).$$

□

Now we have constructed our equivariant maps. The aim is to *find as many r's as possible such that an $\mathfrak{S}_r$-equivariant map*

$$(\Delta_N)^{\times r}_{\Delta(2)} \to S(W_r^{\oplus d}), \qquad \textit{or} \qquad (\Delta_N)^{*r}_{\Delta(2)} \to S(W_r^{\oplus (d+1)}), \tag{9}$$

cannot exist. For this we keep in mind that the $\mathfrak{S}_r$-action on $(\Delta_N)^{\times r}_{\Delta(2)}$ is free, while for $r \geq 3$ the $\mathfrak{S}_r$-action on $(\Delta_N)^{*r}_{\Delta(2)}$ is not free.

3.3 The Topological Tverberg Theorem

The story of the topological Tverberg conjecture continues with a 1981 breakthrough of Bárány, Shlosman and Szűcs [8]. They proved that in the case when r is a prime, there is no $\mathbb{Z}/r$-equivariant map $(\Delta_N)^{\times r}_{\Delta(2)} \to S(W_r^{\oplus d})$, and consequently no

$\mathfrak{S}_r$-equivariant map can exist. Hence, Theorem 3.9 settles the topological Tverberg conjecture in the case when r is a prime. We give a proof of this result relying on the following theorem of Dold [21] [34, Thm. 6.2.6]:

> **Dold's theorem** *Let G be a non-trivial finite group. For an n-connected G-space X and at most n-dimensional free G-CW complex Y there cannot be any continuous G-equivariant map $X \to Y$.*

Theorem 3.10 (Topological Tverberg theorem for primes *r*) *Let $d \geq 1$ be an integer, let $r \geq 2$ be a prime, $N = (d+1)(r-1)$, and let $f : \Delta_N \to \mathbb{R}^d$ be a continuous map. Then there exist r pairwise disjoint faces $\sigma_1, \dots, \sigma_r$ of the simplex Δ_N whose f-images overlap, that is*

$$f(\sigma_1) \cap \cdots \cap f(\sigma_r) \neq \emptyset.$$

Proof According to Theorem 3.9 it suffices to prove that there cannot be any $\mathfrak{S}_r$-equivariant map $(\Delta_N)^{\times r}_{\Delta(2)} \to S(W_r^{\oplus d})$. Let $\mathbb{Z}/r$ be the subgroup of $\mathfrak{S}_r$ generated by the cyclic permutation $(123 \dots r)$. Then it is enough to prove that there is no $\mathbb{Z}/r$-equivariant map $(\Delta_N)^{\times r}_{\Delta(2)} \to S(W_r^{\oplus d})$. For that we are going to use Dold's theorem.

The assumption that r is a prime implies that the action of $\mathbb{Z}/r$ on the sphere $S(W_r^{\oplus d})$ is free. Now, since

- $(\Delta_N)^{\times r}_{\Delta(2)}$ is an $(N-r)$-connected $\mathbb{Z}/r$-space, and
- $S(W_r^{\oplus d})$ is a free $(N-r)$-dimensional $\mathbb{Z}/r$-CW complex,

the theorem of Dold implies that a $\mathbb{Z}/r$-equivariant map $(\Delta_N)^{\times r}_{\Delta(2)} \to S(W_r^{\oplus d})$ cannot exist. □

The same argument yields that there cannot be any $\mathfrak{S}_r$-equivariant map $(\Delta_N)^{*r}_{\Delta(2)} \to S(W_r^{\oplus(d+1)})$, when r is a prime. Observe that for an application of the theorem of Dold the nature of the group action on the domain is of no importance.

The next remarkable step followed a few years later. In 1987 in his landmark unpublished manuscript Özaydin [36] extended the result of Bárány, Shlosman and Szűcs and proved that the topological Tverberg conjecture holds for r a prime power. He proved even more and left the topological Tverberg conjecture as a teaser for generations of mathematicians to come. But this story will come a bit later.

The first published proof of the topological Tverberg theorem for r a prime power appeared in a paper of Aleksei Yu. Volovikov [47]; see Remark 3.12. Here we give a proof of the topological Tverberg theorem for prime powers based on a comparison of Serre spectral sequences which uses a consequence of the localization theorem dict for equivariant cohomology dict [29, Cor. 1, p. 45]. For background on spectral sequences we refer to the textbooks by John McCleary [35] and by Anatoly Fomenko and Dmitry Fuchs [24].

Theorem 3.11 (Topological Tverberg theorem for prime powers *r*) *Let $d \geq 1$ be an integer, let $r \geq 2$ be a prime power, $N = (d+1)(r-1)$, and let $f : \Delta_N \to \mathbb{R}^d$*

be a continuous map. Then there exist r pairwise disjoint faces $\sigma_1, \ldots, \sigma_r$ *of the simplex* Δ_N *whose f-images overlap, that is*

$$f(\sigma_1) \cap \cdots \cap f(\sigma_r) \neq \emptyset.$$

Proof Let $d \geq 1$ be an integer, and let $r = p^n$ for p a prime. By Theorem 3.9 it suffices to prove that there cannot be any $\mathfrak{S}_r$-equivariant map $(\Delta_N)^{\times r}_{\Delta(2)} \to S(W_r^{\oplus d})$.

Consider the elementary abelian group $(\mathbb{Z}/p)^n$ and the regular embedding reg : $(\mathbb{Z}/p)^n \to \mathfrak{S}_r$, as explained in [1, Ex. 2.7, p. 100]. It is given by the left translation action of $(\mathbb{Z}/p)^n$ on itself: To each element $g \in (\mathbb{Z}/p)^n$ we associate the permutation $L_g : (\mathbb{Z}/p)^n \to (\mathbb{Z}/p)^n$ from $\mathrm{Sym}((\mathbb{Z}/p)^n) \cong \mathfrak{S}_r$ given by $L_g(x) = g + x$. We identify the elementary abelian group $(\mathbb{Z}/p)^n$ with the subgroup im(reg) of the symmetric group $\mathfrak{S}_r$. Thus, in order to prove the non-existence of an $\mathfrak{S}_r$-equivariant map it suffices to prove the non-existence of a $(\mathbb{Z}/p)^n$-equivariant map $(\Delta_N)^{\times r}_{\Delta(2)} \to S(W_r^{\oplus d})$.

Our proof takes several steps; the crucial ingredient is a comparison of Serre spectral sequences. As it will be by contradiction, let us now assume that a $(\mathbb{Z}/p)^n$-equivariant map $\varphi : (\Delta_N)^{\times r}_{\Delta(2)} \to S(W_r^{\oplus d})$ exists.

(1) Let λ denote the Borel construction $^{\text{dict}}$ fiber bundle

$$\lambda \quad : \quad (\Delta_N)^{\times r}_{\Delta(2)} \to \mathrm{E}(\mathbb{Z}/p)^n \times_{(\mathbb{Z}/p)^n} (\Delta_N)^{\times r}_{\Delta(2)} \to \mathrm{B}(\mathbb{Z}/p)^n,$$

while ρ denotes the Borel construction fiber bundle

$$\rho \quad : \quad S(W_r^{\oplus d}) \to \mathrm{E}(\mathbb{Z}/p)^n \times_{(\mathbb{Z}/p)^n} S(W_r^{\oplus d}) \to \mathrm{B}(\mathbb{Z}/p)^n.$$

Then the map φ would induce the following morphism between fiber bundles λ and ρ:

$$\begin{tikzcd}
\mathrm{E}(\mathbb{Z}/p)^n \times_{(\mathbb{Z}/p)^n} (\Delta_N)^{\times r}_{\Delta(2)} \arrow[r, "\mathrm{id} \times_{(\mathbb{Z}/p)^n} \varphi"] \arrow[d] & \mathrm{E}(\mathbb{Z}/p)^n \times_{(\mathbb{Z}/p)^n} S(W_r^{\oplus d}) \arrow[d] \\
\mathrm{B}(\mathbb{Z}/p)^n \arrow[r, "="] & \mathrm{B}(\mathbb{Z}/p)^n.
\end{tikzcd}$$

This bundle morphism induces a morphism of associated cohomology Serre spectral sequences:

$$\begin{aligned} E_s^{i,j}(\lambda) &:= E_s^{i,j}(\mathrm{E}(\mathbb{Z}/p)^n \times_{(\mathbb{Z}/p)^n} (\Delta_N)^{\times r}_{\Delta(2)}) \\ &\xleftarrow{\Phi_s^{i,j}} E_s^{i,j}(\mathrm{E}(\mathbb{Z}/p)^n \times_{(\mathbb{Z}/p)^n} S(W_r^{\oplus d})) =: E_s^{i,j}(\rho) \end{aligned}$$

such that on the zero row of the second term

$$\begin{aligned} E_2^{i,0}(\lambda) &:= E_2^{i,0}(\mathrm{E}(\mathbb{Z}/p)^n \times_{(\mathbb{Z}/p)^n} (\Delta_N)^{\times r}_{\Delta(2)}) \\ &\xleftarrow{\Phi_2^{i,0}} E_2^{i,0}(\mathrm{E}(\mathbb{Z}/p)^n \times_{(\mathbb{Z}/p)^n} S(W_r^{\oplus d})) =: E_2^{i,0}(\rho) \end{aligned}$$

is the identity. Here for the morphisms we use the simplified notation $\Phi_s^{i,j} := E_s^{i,j}(\mathrm{id} \times_{(\mathbb{Z}/p)^n} \varphi)$.

Before calculating both spectral sequences we recall the cohomology of $\mathrm{B}(\mathbb{Z}/p)^n$ with coefficients in the field $\mathbb{F}_p$. For $p = 2$ we have:

$$H^*(\mathrm{B}\,((\mathbb{Z}/2)^n)\,;\mathbb{F}_2) = H^*((\mathbb{Z}/2)^n;\mathbb{F}_2) \cong \mathbb{F}_2[t_1,\dots,t_n],$$

where $\deg t_i = 1$, and for $p \geq 3$ we set:

$$H^*(\mathrm{B}\,((\mathbb{Z}/p)^n)\,;\mathbb{F}_p) = H^*((\mathbb{Z}/p)^n;\mathbb{F}_p) \cong \mathbb{F}_p[t_1,\dots,t_n] \otimes \Lambda[e_1,\dots,e_n],$$

where $\deg t_i = 2, \deg e_i = 1$, and $\Lambda[\,\cdot\,]$ denotes the exterior algebra.

(2) First, we consider the Serre spectral sequence, with coefficients in the field $\mathbb{F}_p$, associated to the fiber bundle λ. The E_2-term of this spectral sequence can be computed as follows:

$$\begin{aligned} E_2^{i,j}(\lambda) &= H^i(\mathrm{B}\,((\mathbb{Z}/p)^n)\,;\mathcal{H}^j((\Delta_N)^{\times r}_{\Delta(2)};\mathbb{F}_p)) = H^i((\mathbb{Z}/p)^n;H^j((\Delta_N)^{\times r}_{\Delta(2)};\mathbb{F}_p)) \\ &= \begin{cases} H^i((\mathbb{Z}/p)^n;\mathbb{F}_p), & \text{for } j = 0, \\ H^i((\mathbb{Z}/p)^n;H^{N-r+1}((\Delta_N)^{\times r}_{\Delta(2)};\mathbb{F}_p)), & \text{for } j = N-r+1, \\ 0, & \text{otherwise,} \end{cases} \end{aligned}$$

since by Theorem 3.4 the deleted product $(\Delta_N)^{\times r}_{\Delta(2)}$ is an $(N-r+1)$-dimensional, $(N-r)$-connected simplicial complex and consequently $H^j((\Delta_N)^{\times r}_{\Delta(2)};\mathbb{F}_p) \neq 0$ only for $j = 0$ or $j = N-r+1$. Thus, the only possibly non-zero differential of the spectral sequence is ∂_{N-r+2} and therefore $E_2^{i,0}(\lambda) \cong E_\infty^{i,0}(\lambda)$ for $i \leq N-r+1$.

(3) The second Serre spectral sequence, with coefficients in the field $\mathbb{F}_p$, we consider is associated to the fiber bundle ρ. In this case the fundamental group of the base space $\pi_1(\mathrm{B}(\mathbb{Z}/p)^n) \cong (\mathbb{Z}/p)^n$ acts trivially on the cohomology $H^*(S(W_r^{\oplus d});\mathbb{F}_p)$. Indeed, when $p = 2$ the group $(\mathbb{Z}/2)^n$ can only act trivially on $H^0(S(W_r^{\oplus d});\mathbb{F}_2) \cong \mathbb{F}_2$ and on $H^{N-r}(S(W_r^{\oplus d});\mathbb{F}_2) \cong \mathbb{F}_2$. For p an odd prime all elements of the group $(\mathbb{Z}/p)^n$ have odd order and therefore the action is trivial on the $\mathbb{F}_p$ vector spaces $H^0(S(W_r^{\oplus d}); \mathbb{F}_p) \cong \mathbb{F}_p$ and $H^{N-r}(S(W_r^{\oplus d}); \mathbb{F}_p) \cong \mathbb{F}_p$. Thus the E_2-term of this spectral sequence is of the form

$$\begin{aligned} E_2^{i,j}(\rho) &= H^i(\mathrm{B}\,((\mathbb{Z}/p)^n)\,; \mathcal{H}^j(S(W_r^{\oplus d}); \mathbb{F}_p)) = H^i((\mathbb{Z}/p)^n; H^j(S(W_r^{\oplus d}); \mathbb{F}_p)) \\ &\cong H^i((\mathbb{Z}/p)^n; \mathbb{F}_p) \otimes_{\mathbb{F}_p} H^j(S(W_r^{\oplus d}); \mathbb{F}_p) \\ &\cong \begin{cases} H^i((\mathbb{Z}/p)^n; \mathbb{F}_p), & \text{for } j = 0 \text{ or } N-r, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Moreover, if $\ell \in H^{N-r}(S(W_r^{\oplus d}); \mathbb{F}_p) \cong \mathbb{F}_p$ denotes a generator then the $(N-r)$-row of the E_2-term is a free $H^*((\mathbb{Z}/p)^n; \mathbb{F}_p)$-module generated by the element $1 \otimes_{\mathbb{F}_p} \ell \in E_2^{0,N-r}(\rho) \cong H^{N-r}(S(W_r^{\oplus d}); \mathbb{F}_p)$. The only possible non-zero differential is

$$\partial_{N-r+1} : E_{N-r+1}^{0,N-r}(\rho) \to E_{N-r+1}^{N-r+1,0}(\rho).$$

Consequently we have that

- E_2 and E_{N-r+1} terms coincide, that is $E_2^{*,*}(\rho) \cong E_{N-r+1}^{*,*}(\rho)$,
- $(N-r)$-row of the E_{N-r+1}-term is a free $H^*((\mathbb{Z}/p)^n; \mathbb{F}_p)$-module generated by $1 \otimes_{\mathbb{F}_p} \ell \in E_{N-r+1}^{0,N-r}(\rho)$,
- ∂_{N-r+1}, as all differentials, is an $H^*((\mathbb{Z}/p)^n; \mathbb{F}_p)$-module morphism.

Therefore, the differential ∂_{N-r+1} is zero if and only if

$$\partial_{N-r+1}(1 \otimes_{\mathbb{F}_p} \ell) = 0 \in E_{N-r+1}^{N-r+1,0}(\rho) \cong E_2^{N-r+1,0}(\rho).$$

Furthermore, if $\partial_{N-r+1}(1 \otimes_{\mathbb{F}_p} \ell) = 0$, then $E_2^{*,*}(\rho) \cong E_\infty^{*,*}(\rho)$. Hence, the projection map

$$\mathrm{E}(\mathbb{Z}/p)^n \times_{(\mathbb{Z}/p)^n} S(W_r^{\oplus d}) \to \mathrm{B}(\mathbb{Z}/p)^n$$

induces a monomorphism in cohomology

$$H^*(\mathrm{B}(\mathbb{Z}/p)^n; \mathbb{F}_p) \to H^*(\mathrm{E}(\mathbb{Z}/p)^n \times_{(\mathbb{Z}/p)^n} S(W_r^{\oplus d}); \mathbb{F}_p).$$

Now the following consequence of the localization theorem [29, Cor. 1, p. 45], which in the case of finite groups holds only for elementary abelian groups, comes into play:

Theorem *Let p be a prime, $G = (\mathbb{Z}/p)^n$ with $n \geq 1$, and let X be a finite G-CW complex. The fixed point set X^G of the space X is non-empty if and only if the map in cohomology $H^*(\mathrm{B}G; \mathbb{F}_p) \to H^*(\mathrm{E}G \times_G X; \mathbb{F}_p)$, induced by the projection $\mathrm{E}G \times_G X \to \mathrm{B}G$, is a monomorphism.*

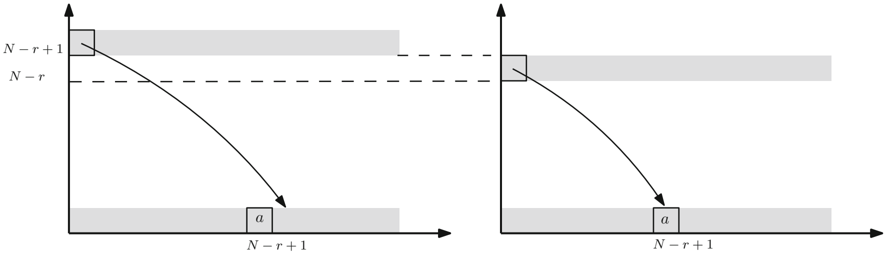


Fig. 3 Illustration of the spectral sequences $E_*^{*,*}(\lambda)$ and $E_*^{*,*}(\rho)$ and the morphism between them $\Phi_*^{*,*} : E_*^{*,*}(\lambda) \leftarrow E_*^{*,*}(\rho)$ that is the identity between the 0-rows up to the E_{N-r+1}-term

Since the fixed point set $S(W_r^{\oplus d})^{(\mathbb{Z}/p)^n} = \emptyset$ of the sphere is empty, the theorem we just quoted implies that the map in cohomology

$$H^*(\mathrm{B}(\mathbb{Z}/p)^n; \mathbb{F}_p) \to H^*(\mathrm{E}(\mathbb{Z}/p)^n \times_{(\mathbb{Z}/p)^n} S(W_r^{\oplus d}); \mathbb{F}_p).$$

is *not* a monomorphism. Consequently, the element

$$a := \partial_{N-r+1}(1 \otimes_{\mathbb{F}_p} \ell) \neq 0 \in E_{N-r+1}^{N-r+1,0}(\rho) \cong E^{N-r+1,0}(\rho)$$

is *not* zero.

(4) Finally, to reach a contradiction with the assumption that the $(\mathbb{Z}/p)^n$-equivariant map φ exists we track the element $a := \partial_{N-r+1}(1 \otimes_{\mathbb{F}_p} \ell) \neq 0 \in E_{N-r+1}^{N-r+1,0}(\rho) \cong E_2^{N-r+1,0}(\rho)$ along the morphism of spectral sequences (Fig. 3)

$$\Phi_s^{N-r+1,0} : E_s^{N-r+1,0}(\rho) \to E_s^{N-r+1,0}(\lambda).$$

Since the differentials in both spectral sequences are zero in all terms $E_s(\rho)$ and $E_s(\lambda)$ for $2 \le s \le N-r$ we have that $\Phi_{s'}^{*,0}$ is the identity for $2 \le s' \le N-r+1$. In particular, the morphism

$$\Phi_{N-r+1}^{N-r+1,0} : E_{N-r+1}^{N-r+1,0}(\rho) \to E_{N-r+1}^{N-r+1,0}(\lambda)$$

is still identity as it was in the second term, and so $\Phi_{N-r+1}^{N-r+1,0}(a) = a$. Passing to the $(N-r+1)$-term, with a slight abuse of notation, we have that

$$\Phi_{N-r+2}^{N-r+1,0}([a]) = [a],$$

where $[a]$ denotes the class induced by a in the appropriate $(N-r+2)$-term of the spectral sequences. Since $a := \partial_{N-r+1}(1 \otimes_{\mathbb{F}_p} \ell) \in E_{N-r+1}^{N-r+1,0}(\rho)$ and

$0 \neq a \in E_2^{N-r+1,0}(\lambda) \cong E_\infty^{N-r+1,0}(\lambda)$ passing to the next E_{N-r+2}-term we reach a contradiction:

$$\Phi_{N-r+2}^{N-r+1,0}(0) = [a] = a \neq 0,$$

because the class of the element a in $E_{N-r+2}^{N-r+1,0}(\rho)$ vanishes (domain of $\Phi_{N-r+2}^{N-r+1,0}$) while in $E_{N-r+2}^{N-r+1,0}(\lambda)$ it does not vanish (codomain of $\Phi_{N-r+2}^{N-r+1,0}$). Hence, there cannot be any $(\mathbb{Z}/p)^n$-equivariant map $(\Delta_N)^{\times r}_{\Delta(2)} \to S(W_r^{\oplus d})$, and the proof of the theorem is complete.

□

In the language of the Fadell–Husseini index dict, as introduced in [22], we have computed that

$$\mathrm{index}_{(\mathbb{Z}/p)^n}((\Delta_N)^{\times r}_{\Delta(2)}; \mathbb{F}_p) \subseteq H^{\geq N-r+2}(\mathrm{B}(\mathbb{Z}/p)^n; \mathbb{F}_p).$$

Furthermore, we showed the existence of an element $a \in H^{\geq N-r+1}(\mathrm{B}(\mathbb{Z}/p)^n; \mathbb{F}_p)$ that has the property

$$0 \neq a \in \mathrm{index}_{(\mathbb{Z}/p)^n}(S(W_r^{\oplus d}); \mathbb{F}_p) \cap H^{N-r+1}(\mathrm{B}(\mathbb{Z}/p)^n; \mathbb{F}_p). \tag{10}$$

Consequently $\mathrm{index}_{(\mathbb{Z}/p)^n}(S(W_r^{\oplus d}); \mathbb{F}_p) \not\subseteq \mathrm{index}_{(\mathbb{Z}/p)^n}((\Delta_N)^{\times r}_{\Delta(2)}; \mathbb{F}_p)$ and so the monotonicity property of the Fadell–Husseini index implies the non-existence of a $(\mathbb{Z}/p)^n$-equivariant map $(\Delta_N)^{\times r}_{\Delta(2)} \to S(W_r^{\oplus d})$.

The element a with the property (10) can be specified explicitly. It is the Euler class of the vector bundle

$$W_r^{\oplus d} \to \mathrm{E}(\mathbb{Z}/p)^n \times_{(\mathbb{Z}/p)^n} W_r^{\oplus d} \to \mathrm{B}(\mathbb{Z}/p)^n.$$

From the work of Mann and Milgram [33] we get that for an odd prime p

$$a = \omega \cdot \Big(\prod_{(\alpha_1,\dots,\alpha_n)\in\mathbb{F}_p^n\setminus\{0\}} (\alpha_1 t_1 + \cdots + \alpha_n t_n) \Big)^{d/2},$$

where $\omega \in \mathbb{F}_p\setminus\{0\}$, while for $p = 2$ we have that

$$a = \Big(\prod_{(\alpha_1,\dots,\alpha_n)\in\mathbb{F}_2^n\setminus\{0\}} (\alpha_1 t_1 + \cdots + \alpha_n t_n) \Big)^{d}.$$

The square root in $\mathbb{F}_p[t_1, \dots, t_n]$ is not uniquely determined for an odd prime p and d odd: The factor ω accounts for an arbitrary square root being taken.

Remark 3.12 Volovikov, in his 1996 paper [47], proved the following extension of the topological Tverberg theorem for continuous maps to manifolds:

> **Theorem** *Let $d \geq 1$ be an integer, let $r \geq 2$ be a prime power, and $N = (d+1)(r-1)$. For any topological d-manifold M and any continuous map $f : \Delta_N \to M$, there exist r pairwise disjoint faces $\sigma_1, \ldots, \sigma_r$ of the simplex Δ_N whose f-images overlap, that is*
>
> $$f(\sigma_1) \cap \cdots \cap f(\sigma_r) \neq \emptyset.$$

4 Corollaries of the Topological Tverberg Theorem

Over time many results were discovered that were believed to be substantial extensions or analogs of the topological Tverberg theorem, such as the generalized Van Kampen–Flores theorem of Karanbir Sarkaria [38] and Aleksei Volovikov [48], the colored Tverberg theorems of Rade Živaljević and Siniša Vrećica [49, 53] and Pablo Soberón's result on Tverberg points with equal barycentric coordinates [42]. It turned out only recently that the elementary idea of constraint functions together with the concept of "unavoidable complexes" introduced in [12] transforms all these results into simple corollaries of the topological Tverberg theorem.

Well, if all these results are corollaries, is there any genuine extension of the topological Tverberg theorem? The answer to this question will bring us to the fundamental work of Bárány and Larman [7], and the optimal colored Tverberg theorem [17] from 2009. But this will be the story of the final section of this paper.

4.1 The Generalized Van Kampen–Flores Theorem

The first corollary we prove is the following generalized Van Kampen–Flores Theorem that was originally proved by Sarkaria [38] for primes and then by Volovikov [48] for prime powers. The fact that this result can be derived easily from the topological Tverberg theorem by adding an extra component to the map was first sketched by Gromov in [26, Sec. 2.9c]; this can be seen as a first instance of the constraint method [12, Thm. 6.3] "at work."

Theorem 4.1 (The generalized Van Kampen–Flores Theorem) *Let $d \geq 1$ be an integer, let r be a prime power, let $k \geq \lceil \frac{r-1}{r} d \rceil$ and $N = (d+2)(r-1)$, and let $f : \Delta_N \to \mathbb{R}^d$ be a continuous map. Then there exist r pairwise disjoint faces $\sigma_1, \ldots, \sigma_r$ in the k-skeleton $\mathrm{sk}_k(\Delta_N)$ of the simplex Δ_N whose f-images overlap,*

$$f(\sigma_1) \cap \cdots \cap f(\sigma_r) \neq \emptyset.$$

Proof For the proof we use two ingredients, the topological Tverberg theorem and the pigeonhole principle. First, consider the continuous map $g : \Delta_N \to \mathbb{R}^{d+1}$ defined by

$$g(x) = (f(x), \mathrm{dist}(x, \mathrm{sk}_k(\Delta_N)).$$

Since $N = (d+2)(r-1) = ((d+1)+1)(r-1)$ and r is a prime power we can apply the topological Tverberg theorem to the map g. Consequently, there exist r pairwise disjoint faces $\sigma_1, \dots, \sigma_r$ with points $x_1 \in \mathrm{relint}\,\sigma_1, \dots, x_r \in \mathrm{relint}\,\sigma_r$ such that $g(x_1) = \cdots = g(x_r)$, that is,

$$f(x_1) = \cdots = f(x_r) \qquad \text{and} \qquad \mathrm{dist}(x_1, \mathrm{sk}_k(\Delta_N)) = \cdots = \mathrm{dist}(x_r, \mathrm{sk}_k(\Delta_N)).$$

One of the faces $\sigma_1, \dots, \sigma_r$ has to belong to $\mathrm{sk}_k(\Delta_N)$. Indeed, if all the faces $\sigma_1, \dots, \sigma_r$, which are disjoint, would not belong to $\mathrm{sk}_k(\Delta_N)$, then the simplex Δ_N should have at least

$$|\sigma_1| + \cdots + |\sigma_r| \geq r(k+2) \geq r\big(\lceil \tfrac{r-1}{r} d \rceil + 2\big) \geq (r-1)(d+2) + 2 = N + 2$$

vertices. Thus, since one of the faces is in the k-skeleton $\mathrm{dist}(x_1, \mathrm{sk}_k(\Delta_N)) = \cdots = \mathrm{dist}(x_r, \mathrm{sk}_k(\Delta_N)) = 0$, and consequently $\sigma_1 \in \mathrm{sk}_k(\Delta_N), \dots, \sigma_r \in \mathrm{sk}_k(\Delta_N)$, completing the proof of the theorem. □

4.2 The Colored Tverberg Problem of Bárány and Larman

In their 1990 study on halving lines and halving planes, Bárány, Zoltan Füredi and László Lovász [6] realized a need for a colored version of the Tverberg theorem. The sentence from this paper

> For this we need a colored version of Tverberg's theorem.

opened a new chapter in the study of extensions of the Tverberg theorem, both affine and topological. Soon after, in 1992, Bárány and David Larman in [7] formulated the colored Tverberg problem and brought to light a conjecture that motivated the progress in the area for decades to come.

Let $N \geq 1$ be an integer and let $\mathcal{C}$ be the set of vertices of the simplex Δ_N. A *coloring* of the set of vertices $\mathcal{C}$ by ℓ colors is a partition $(C_1, \dots, C_\ell)$ of $\mathcal{C}$, that is $\mathcal{C} = C_1 \cup \cdots \cup C_\ell$ and $C_i \cap C_j = \emptyset$ for $1 \leq i < j \leq \ell$. The elements of the partition $(C_1, \dots, C_\ell)$ are called *color classes*. A face σ of the simplex Δ_N is a *rainbow* face if $|\sigma \cap C_i| \leq 1$ for all $1 \leq i \leq \ell$. The subcomplex of all rainbow faces of the simplex Δ_N induced by the coloring $(C_1, \dots, C_\ell)$ will be denoted by $R_{(C_1,\dots,C_\ell)}$ and will be called the *rainbow subcomplex*. There is an isomorphism of simplicial complexes $R_{(C_1,\dots,C_\ell)} \cong C_1 * \cdots * C_\ell$.

Problem 4.2 (Bárány–Larman colored Tverberg problem) Let $d \geq 1$ and $r \geq 2$ be integers. Determine the smallest number $n = n(d, r)$ such that for every affine map $f : \Delta_{n-1} \to \mathbb{R}^d$, and every coloring $(C_1, \dots, C_{d+1})$ of the vertex set $\mathcal{C}$ of the simplex Δ_{n-1} by $d + 1$ colors with each color of size at least r, there exist r pairwise disjoint rainbow faces $\sigma_1, \dots, \sigma_r$ of Δ_{n-1} whose f-images overlap,

$$f(\sigma_1) \cap \cdots \cap f(\sigma_r) \neq \emptyset.$$

A trivial lower bound for the function $n(d, r)$ is $(d + 1)r$. Bárány and Larman proved that the trivial lower bound is tight in the cases $n(r, 1) = 2r$ and $n(r, 2) = 3r$, and presented a proof by Lovász for $n(2, d) = 2(d + 1)$. Furthermore, they conjectured the following equality.

Conjecture 4.3 (Bárány–Larman conjecture) *Let $r \geq 2$ and $d \geq 1$ be integers. Then $n(d, r) = (d + 1)r$.*

Now we present the proof of Lovász for the Bárány–Larman conjecture in the case $r = 2$ from the paper of Bárány and Larman [7, Thm. (iii)].

Theorem 4.4 *Let $d \geq 1$ be an integer. Then $n(2, d) = 2(d + 1)$.*

Proof Let $n = 2(d + 1)$, and let $f : \Delta_{n-1} \to \mathbb{R}^d$ be an affine map. Furthermore, consider a coloring $(C_1, \dots, C_{d+1})$ of the vertex set $\mathcal{C}$ of the simplex Δ_{n-1} by $d + 1$ colors where $|C_1| = \cdots = |C_{d+1}| = 2$. Denote $C_i = \{v_i, -v_i\}$ for $1 \leq i \leq d + 1$. The subcomplex of all rainbow faces of the simplex Δ_{n-1} is the join $R := R_{(C_1,\dots,C_{d+1})} = C_1 * \cdots * C_{d+1}$. In this case, the rainbow subcomplex R can be identified with the boundary of the cross-polytope $[2]^{*(d+1)} \cong S^d$. Here $[2]$, as before, denotes the 0-dimensional simplicial complex with two vertices.

The restriction map $f|_R : [2]^{*(d+1)} \to \mathbb{R}^d$ is a piecewise affine map, and therefore continuous. The Borsuk–Ulam theorem yields the existence of a point $x \in [2]^{*(d+1)} \cong S^d$ on the sphere with the property that $f|_R(x) = f|_R(-x)$. The point $x \in [2]^{*(d+1)}$ belongs to the relative interior of a unique simplex in the boundary of the cross-polytope $[2]^{*(d+1)}$,

$$x \in \operatorname{relint}\big(\operatorname{conv}\{\varepsilon_{i_1} v_{i_1}, \dots, \varepsilon_{i_k} v_{i_k}\}\big),$$

where $\varepsilon_{i_a} \in \{-1, +1\}$ and $1 \leq k \leq d + 1$. Thus,

$$-x \in \operatorname{relint}\big(\operatorname{conv}\{-\varepsilon_{i_1} v_{i_1}, \dots, -\varepsilon_{i_k} v_{i_k}\}\big).$$

Since the rainbow faces

$$\operatorname{conv}\{\varepsilon_{i_1} v_{i_1}, \dots, \varepsilon_{i_k} v_{i_k}\} \text{ and } \operatorname{conv}\{-\varepsilon_{i_1} v_{i_1}, \dots, -\varepsilon_{i_k} v_{i_k}\}$$

are disjoint, and

$$\begin{aligned} f|_R(x) &= f|_R(-x) \in \\ &f|_R\big(\operatorname{relint}(\operatorname{conv}\{\varepsilon_{i_1} v_{i_1}, \dots, \varepsilon_{i_k} v_{i_k}\})\big) \cap f|_R\big(\operatorname{relint}(\operatorname{conv}\{-\varepsilon_{i_1} v_{i_1}, \dots, -\varepsilon_{i_k} v_{i_k}\})\big), \end{aligned}$$

we have proved the theorem. □

4.3 The Colored Tverberg Problem of Živaljević and Vrećica

In response to the work of Bárány and Larman a modified colored Tverberg problem was presented by Živaljević and Vrećica in their influential paper [53] from 1992.

Problem 4.5 (The Živaljević–Vrećica colored Tverberg problem) Let $d \geq 1$ and $r \geq 2$ be integers. Determine the smallest number $t = t(d, r)$ (or $t = tt(d, r)$) such that for every affine (or continuous) map $f : \Delta \to \mathbb{R}^d$, and every coloring $(C_1, \ldots, C_{d+1})$ of the vertex set $\mathcal{C}$ of the simplex Δ by $d + 1$ colors with each color of size at least t, there exist r pairwise disjoint rainbow faces $\sigma_1, \ldots, \sigma_r$ of Δ whose f-images overlap, that is

$$f(\sigma_1) \cap \cdots \cap f(\sigma_r) \neq \emptyset.$$

Observe that in the language of the function $t(d, r)$ the Bárány–Larman conjecture says that $t(d, r) = r$ for all $r \geq 2$ and $d \geq 1$. Furthermore, proving that $t(d, r) < +\infty$ does *not* imply $n(d, r) < +\infty$, while proving $t(d, r) = r$ would imply that $n(d, r) = r(d + 1)$.

In order to address the modified problem Živaljević and Vrećica needed to know the connectivity of chessboard complexes. For that they recalled the following result of Anders Björner, Lovász, Vrećica and Živaljević [11, Thm. 1.1]. Its connectivity lower bound is best possible according to [39].

Theorem 4.6 *Let $m \geq 1$ and $n \geq 1$ be integers. The chessboard $\Delta_{m,n}$ is ν-connected, where*

$$\nu = \min\left\{m, n, \left\lfloor \tfrac{m+n+1}{3} \right\rfloor\right\} - 2.$$

Proof Without loss of generality we can assume that $1 \leq m \leq n$. The proof proceeds by induction on $\min\{m, n\} = m$. In the case $m = 1$ the statement of the theorem is obviously true. For $m = 2$ we distinguish between two cases:

- If $n = 2$, then $\nu = \min\{2, 2, \left\lfloor \frac{5}{3} \right\rfloor\} - 2 = -1$ and $\Delta_{2,2}$ is just a disjoint union of two edges, and
- If $n \geq 3$, then $\nu = \min\{2, 2, \left\lfloor \frac{n+3}{3} \right\rfloor\} - 2 = 0$ and $\Delta_{2,n}$ is path connected.

Let $m \geq 3$, and let us assume that the statement of the theorem holds for every chessboard $\Delta_{m',n'}$ where $1 \leq \min\{m', n'\} < m$. Now we prove the statement of the theorem for the chessboard $\Delta_{m,n}$.

Let K_ℓ for $1 \leq \ell \leq n$ be a subcomplex of $\Delta_{m,n}$ defined by

$$\{(i_0,j_0), \ldots, (i_k,j_k)\} \in K_\ell \iff \{(i_0,j_0), \ldots, (i_k,j_k), (1, \ell)\} \in \Delta_{m,n}.$$

The family of subcomplexes $\mathcal{K} := \{K_\ell : 1 \leq \ell \leq n\}$ covers the chessboard $\Delta_{m,n}$. Moreover, each subcomplex K_ℓ is a cone over the chessboard $\Delta_{m-1,n-1}$, and therefore contractible. Since, for $\sigma \subseteq [n]$ we have that

$$\bigcap\{K_\ell : \ell \in \sigma\} = \emptyset \iff \sigma = [n],$$

the nerve dict $N_{\mathcal{K}}$ of the family $\mathcal{K}$ is homeomorphic to the boundary of an $(n-1)$-simplex $\partial\Delta_{n-1}$. Thus, $N_{\mathcal{K}} \cong S^{n-2}$ is $(n-3)$-connected. Furthermore, for $\sigma \subseteq [n]$ with the property that $2 \leq |\sigma| \leq n-1$ the intersection $\bigcap\{K_\ell : \ell \in \sigma\}$ is homeomorphism with the chessboard $\Delta_{m-1,n-|\sigma|}$. The induction hypothesis can be applied to each of these intersections. Therefore,

$$\begin{aligned}\operatorname{conn}\left(\bigcap\{K_\ell : \ell \in \sigma\}\right) &= \operatorname{conn}(\Delta_{m-1,n-|\sigma|})\\ &\geq \min\left\{m-1, n-|\sigma|, \left\lfloor \tfrac{m+n-|\sigma|}{3} \right\rfloor\right\} - 2.\end{aligned}$$

Now we will apply the following connectivity version of the Nerve theorem dict due to Björner, see [10, Thm. 10.6].

Theorem *Let K be a finite simplicial complex, or a regular CW-complex, and let $\mathcal{K} := \{K_i : i \in I\}$ be a cover of K by a family of subcomplexes, $K = \bigcup\{K_i : i \in I\}$.*

(1) *If for every face σ of the nerve $N_{\mathcal{K}}$ the intersection $\bigcap\{K_i : i \in \sigma\}$ is contractible, then K and $N_{\mathcal{K}}$ are homotopy equivalent, $K \simeq N_{\mathcal{K}}$.*
(2) *If for every face σ of the nerve $N_{\mathcal{K}}$ the intersection $\bigcap\{K_i : i \in \sigma\}$ is $(k-|\sigma|+1)$-connected, then the complex K is k-connected if and only if the nerve $N_{\mathcal{K}}$ is k-connected.*

In the case of the covering $\mathcal{K}$ of the chessboard $\Delta_{m,n}$, where $2 < m \leq n$, we have that

- for every face σ of the nerve $N_{\mathcal{K}}$ the intersection $\bigcap\{K_i : i \in \sigma\}$ is contractible when $|\sigma| = 1$, and

$$\begin{aligned}\operatorname{conn}\left(\bigcap\{K_\ell : \ell \in \sigma\}\right) &\geq \min\left\{m-1, n-|\sigma|, \left\lfloor \tfrac{m+n-|\sigma|}{3} \right\rfloor\right\} - 2\\ &\geq \min\left\{m, n, \left\lfloor \tfrac{m+n+1}{3} \right\rfloor\right\} - 2 - |\sigma| + 1\\ &\geq \nu - |\sigma| + 1,\end{aligned}$$

  when $2 \leq |\sigma| \leq n-1$, while
- the nerve $N_{\mathcal{K}}$ of the family $\mathcal{K}$ is $(n-3)$-connected with

$$n - 3 \geq \min\left\{m, n, \left\lfloor \tfrac{m+n+1}{3} \right\rfloor\right\} - 2 = \nu.$$

Therefore, according to the Nerve theorem applied for the cover $\mathcal{K}$ the chessboard $\Delta_{m,n}$ is ν-connected. This concludes the induction step. □

The knowledge on the connectivity of the chessboard complexes was the decisive information both for the original proof of the Živaljević and Vrećica colored Tverberg theorem [53, Thm. 1], which worked only for primes, and for the following version of the proof for prime powers; see also the proof of Živaljević [52, Thm. 3.2 (2)].

Theorem 4.7 (Colored Tverberg theorem of Živaljević and Vrećica) *Let $d \geq 1$ be an integer, and let $r \geq 2$ be a prime power. For every continuous map $f : \Delta \to \mathbb{R}^d$, and every coloring $(C_1, \dots, C_{d+1})$ of the vertex set $\mathcal{C}$ of the simplex Δ by $d+1$ colors with each color of size at least $2r-1$, there exist r pairwise disjoint rainbow faces $\sigma_1, \dots, \sigma_r$ of Δ whose f-images overlap, that is*

$$f(\sigma_1) \cap \dots \cap f(\sigma_r) \neq \emptyset.$$

In the language of the function $tt(d,r)$ the previous theorem yields the upper bound $tt(d,r) \leq 2r-1$ when r is a prime power. This bound implies the bound $t(d,r) \leq tt(d,r) \leq 4r-3$ for arbitrary r via Bertrand's postulate.

Proof Let $r = p^n$ for p a prime and $n \geq 1$. Let $f : \Delta \to \mathbb{R}^d$ be a continuous map from a simplex Δ whose set of vertices $\mathcal{C}$ is colored by $d+1$ colors $(C_1, \dots, C_{d+1})$. Without loss of generality assume that $|C_1| = \dots = |C_{d+1}| = 2r-1$. In addition assume that the map f is a counterexample for the statement of the theorem. Set $M := (d+1)(2r-1)-1$ and $N := (d+1)(r-1)$, so Δ is an M-dimensional simplex. Now, the proof of the theorem will be presented in several steps.

(1) The rainbow subcomplex of the simplex Δ induced by the coloring $(C_1, \dots, C_{d+1})$ in this case is

$$R_{(C_1,\dots,C_{d+1})} \cong C_1 * \dots * C_{d+1} \cong [2r-1]^{*(d+1)}.$$

The r-fold 2-wise deleted join of the rainbow subcomplex $R_{(C_1,\dots,C_{d+1})}$ can be identified, with the help of Lemma 3.7 and Example 3.6, as follows

$$\begin{aligned}(R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)} &\cong \big([2r-1]^{*(d+1)}\big)^{*r}_{\Delta(2)} \\ &\cong \big([2r-1]^{*r}_{\Delta(2)}\big)^{*(d+1)} \cong (\Delta_{2r-1,r})^{*(d+1)}.\end{aligned}$$

The action of the symmetric group $\mathfrak{S}_r$ on the chessboard $\Delta_{2r-1,r}$ is assumed to be given by permutation of columns of the chessboard, that is

$$\pi \cdot \{(i_0,j_0), \dots, (i_k,j_k)\} = \{(i_0,\pi(j_0)), \dots, (i_k,\pi(j_k))\},$$

for $\pi \in \mathfrak{S}_r$ and $\{(i_0,j_0),\dots,(i_k,j_k)\}$ a simplex in $\Delta_{2r-1,r}$. Furthermore, the chessboard $\Delta_{2r-1,r}$ is an $(r-1)$-dimensional and according to Theorem 4.6 an $(r-2)$-connected simplicial complex. Therefore

$$\begin{aligned} \dim\left((\Delta_{2r-1,r})^{*(d+1)}\right) &= (d+1)r-1 = N+d, \\ \operatorname{conn}\left((\Delta_{2r-1,r})^{*(d+1)}\right) &= (d+1)r-2 = N+d-1. \end{aligned} \tag{11}$$

(2) Now, along the lines of Sect. 3.2.3, the continuous map $f : \Delta \to \mathbb{R}^d$ induces the join map

$$J_f : (\Delta)^{*r}_{\Delta(2)} \to (\mathbb{R}^{d+1})^{\oplus r},$$
$$\lambda_1 x_1 + \cdots + \lambda_r x_r \longmapsto (\lambda_1, \lambda_1 f(x_1)) \oplus \cdots \oplus (\lambda_r, \lambda_r f(x_r)).$$

Both domain and codomain of the join map J_f are equipped with the action of the symmetric group $\mathfrak{S}_r$ in such a way that J_f is an $\mathfrak{S}_r$-equivariant map. The deleted join of the rainbow complex $(R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)}$ is an $\mathfrak{S}_r$-invariant subcomplex of $(\Delta)^{*r}_{\Delta(2)}$. Thus, the restriction map

$$J'_f := J_f|_{(R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)}} : (R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)} \to (\mathbb{R}^{d+1})^{\oplus r}$$

is also an $\mathfrak{S}_r$-equivariant map. Next consider the thin diagonal

$$D_J = \{(z_1,\dots,z_r) \in (\mathbb{R}^{d+1})^{\oplus r} : z_1 = \cdots = z_r\}.$$

This is an $\mathfrak{S}_r$-invariant subspace of $(\mathbb{R}^{d+1})^{\oplus r}$. The key property of the map J'_f we have constructed for any counterexample continuous map $f : \Delta \to \mathbb{R}^d$, is that $\operatorname{im}(J'_f) \cap D_J = \emptyset$. Thus J'_f induces an $\mathfrak{S}_r$-equivariant map

$$(R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)} \to (\mathbb{R}^{d+1})^{\oplus r} \backslash D_J \tag{12}$$

which we, with an obvious abuse of notation, again denote by J'_f. Furthermore, let

$$R_J : (\mathbb{R}^{d+1})^{\oplus r} \backslash D_J \to D_J^{\perp} \backslash \{0\} \to S(D_J^{\perp}) \tag{13}$$

be the composition of the appropriate projection and deformation retraction. The map R_J is $\mathfrak{S}_r$-equivariant. Recall from Sect. 3.2.3 that there is an isomorphism of real $\mathfrak{S}_r$-representations $D_J^{\perp} \cong W_r^{\oplus(d+1)}$. Here $W_r = \left\{(t_1,\dots,t_r) \in \mathbb{R}^r : \sum_{i=1}^r t_i = 0\right\}$ and it is equipped with the (left) action of the symmetric group $\mathfrak{S}_r$ given by permutation of coordinates. After the identification of the $\mathfrak{S}_r$-representations the $\mathfrak{S}_r$-equivariant map R_J defined in (13) has the form

$$R_J : (\mathbb{R}^{d+1})^{\oplus r}\backslash D_J \to S(W_r^{\oplus(d+1)}). \tag{14}$$

Thus we have proved that if there exists a counterexample map f for the theorem, then there exists an $\mathfrak{S}_r$-equivariant map

$$(R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)} \to S(W_r^{\oplus(d+1)}). \tag{15}$$

In the final step we reach a contradiction by proving that an $\mathfrak{S}_r$-equivariant map (15) cannot exist, concluding that a counterexample f could not exist in the first place. The proof of the non-existence of an equivariant map is following the footsteps of the proof of Theorem 3.11.

(3) Consider the elementary abelian group $(\mathbb{Z}/p)^n$ and its regular embedding $\mathrm{reg} : (\mathbb{Z}/p)^n \to \mathfrak{S}_r$. Now it suffices to prove the non-existence of a $(\mathbb{Z}/p)^n$-equivariant map $(R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)} \to S(W_r^{\oplus(d+1)})$. To prove the non-existence of such a map assume the opposite: let $\varphi : (R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)} \to S(W_r^{\oplus(d+1)})$ be a $(\mathbb{Z}/p)^n$-equivariant map.

Denote by λ the Borel construction fiber bundle

$$\lambda \quad : \quad (R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)} \to \mathrm{E}(\mathbb{Z}/p)^n \times_{(\mathbb{Z}/p)^n} (R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)} \to \mathrm{B}(\mathbb{Z}/p)^n,$$

and by ρ the following Borel construction fiber bundle

$$\rho \quad : \quad S(W_r^{\oplus(d+1)}) \to \mathrm{E}(\mathbb{Z}/p)^n \times_{(\mathbb{Z}/p)^n} S(W_r^{\oplus(d+1)}) \to \mathrm{B}(\mathbb{Z}/p)^n.$$

Then the map φ induces the following morphism of fiber bundles

$$\begin{tikzcd}
\mathrm{E}(\mathbb{Z}/p)^n \times_{(\mathbb{Z}/p)^n} (R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)} \arrow[r, "\mathrm{id}\times_{(\mathbb{Z}/p)^n}\varphi"] \arrow[d] & \mathrm{E}(\mathbb{Z}/p)^n \times_{(\mathbb{Z}/p)^n} S(W_r^{\oplus(d+1)}) \arrow[d] \\
\mathrm{B}(\mathbb{Z}/p)^n \arrow[r, "="] & \mathrm{B}(\mathbb{Z}/p)^n.
\end{tikzcd}$$

In turn, this morphism induces a morphism of corresponding Serre spectral sequences

$$E_s^{i,j}(\lambda) := E_s^{i,j}(\mathrm{E}(\mathbb{Z}/p)^n \times_{(\mathbb{Z}/p)^n} (R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)})$$

$$\xleftarrow{\Phi_s^{i,j}} E_s^{i,j}(\mathrm{E}(\mathbb{Z}/p)^n \times_{(\mathbb{Z}/p)^n} S(W_r^{\oplus(d+1)})) =: E_s^{i,j}(\rho)$$

with the property that on the zero row of the second term the induced map

$$\begin{aligned} E_2^{i,0}(\lambda) &:= E_2^{i,0}(\mathrm{E}(\mathbb{Z}/p)^n \times_{(\mathbb{Z}/p)^n} (R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)}) \\ &\xleftarrow{\Phi_2^{i,0}} E_2^{i,0}(\mathrm{E}(\mathbb{Z}/p)^n \times_{(\mathbb{Z}/p)^n} S(W_r^{\oplus(d+1)})) =: E_2^{i,0}(\rho) \end{aligned}$$

is the identity. Again we use simplified notation by setting $\Phi_s^{i,j} := E_s^{i,j}(\mathrm{id} \times_{(\mathbb{Z}/p)^n} \varphi)$.

In the case when the prime p is 2 then

$$H^*(\mathrm{B}\,((\mathbb{Z}/2)^n)\,;\mathbb{F}_2) = H^*((\mathbb{Z}/2)^n;\mathbb{F}_2) \cong \mathbb{F}_2[t_1,\dots,t_n],$$

where $\deg t_i = 1$, while in the case $p \geq 3$ we have

$$H^*(\mathrm{B}\,((\mathbb{Z}/p)^n)\,;\mathbb{F}_p) = H^*((\mathbb{Z}/p)^n;\mathbb{F}_p) \cong \mathbb{F}_p[t_1,\dots,t_n] \otimes \Lambda[e_1,\dots,e_n],$$

where $\deg t_i{=}2$, $\deg e_i = 1$, and $\Lambda[\,\cdot\,]$ denotes the exterior algebra.

(4) First we consider the Serre spectral sequence, with coefficients in the field $\mathbb{F}_p$, associated to the fiber bundle λ. Using the connectivity of $(R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)}$ derived in (11), we get that the E_2-term of this spectral sequence is

$$\begin{aligned} E_2^{i,j}(\lambda) &= H^i(\mathrm{B}\,((\mathbb{Z}/p)^n)\,;\mathcal{H}^j((R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)};\mathbb{F}_p)) \\ &= H^i((\mathbb{Z}/p)^n;H^j((R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)};\mathbb{F}_p)) \\ &\cong \begin{cases} H^i((\mathbb{Z}/p)^n;\mathbb{F}_p), & \text{for } j = 0, \\ H^i((\mathbb{Z}/p)^n;H^{N+d}((R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)};\mathbb{F}_p)), & \text{for } j = N+d, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Consequently, the only possibly non-zero differential of this spectral sequence is ∂_{N+d+1} and therefore $E_2^{i,0}(\lambda) \cong E_\infty^{i,0}(\lambda)$ for $i \leq N+d$.

(5) The Serre spectral sequence, with coefficients in the field $\mathbb{F}_p$, associated to the fiber bundle ρ is similar to the one appearing in the proof of Theorem 3.11. Briefly, the E_2-term of this spectral sequence is

$$\begin{aligned} E_2^{i,j}(\rho) &= H^i((\mathbb{Z}/p)^n;H^j(S(W_r^{\oplus(d+1)});\mathbb{F}_p)) \\ &\cong \begin{cases} H^i((\mathbb{Z}/p)^n;\mathbb{F}_p), & \text{for } j = 0 \text{ or } N-1, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Letting $\ell \in H^{N-1}(S(W_r^{\oplus(d+1)});\mathbb{F}_p) \cong \mathbb{F}_p$ denote a generator, then the $(N-1)$-row of the E_2-term can be seen as a free $H^*((\mathbb{Z}/p)^n;\mathbb{F}_p)$-module generated by $1 \otimes_{\mathbb{F}_p} \ell \in E_2^{0,N-1}(\rho) \cong H^{N-1}(S(W_r^{\oplus(d+1)});\mathbb{F}_p)$. Thus the only possible non-zero differential is $\partial_N : E_N^{i,N-1}(\rho) \to E_N^{N+i,0}(\rho)$ and it is completely determined by

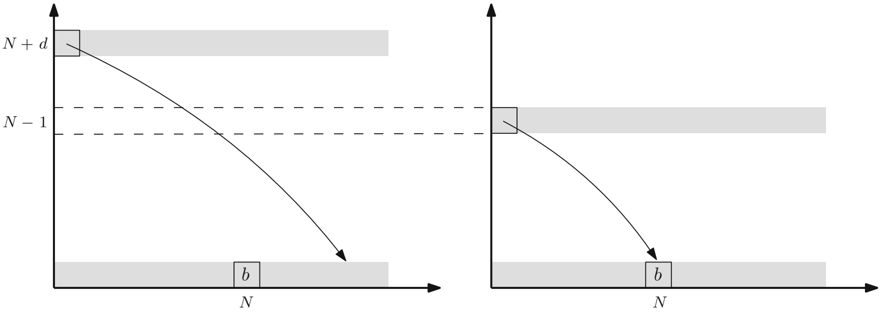


Fig. 4 Illustration of the spectral sequences $E_*^{*,*}(\lambda)$ and $E_*^{*,*}(\rho)$ and the morphism between them $\Phi_*^{*,*} : E_*^{*,*}(\lambda) \leftarrow E_*^{*,*}(\rho)$ that is the identity between the 0-rows up to the E_N-term

the image $\partial_N(1 \otimes_{\mathbb{F}_p} \ell)$. As in the proof of Theorem 3.11, a consequence of the localization theorem implies that $b := \partial_N(1 \otimes_{\mathbb{F}_p} \ell) \neq 0 \in E_N^{N,0}(\rho) \cong E_2^{N,0}(\rho)$ is *not* zero.

(6) To reach the desired contradiction we track the element $b \in E_N^{N,0}(\rho) \cong E_2^{N,0}(\rho)$ along the morphism of spectral sequences (Fig. 4)

$$\Phi_s^{N,0} : E_s^{N,0}(\rho) \to E_s^{N,0}(\lambda).$$

The differentials in both spectral sequences are zero in all terms $E_s(\rho)$ and $E_s(\lambda)$ for $2 \leq s \leq N-1$. Thus, $\Phi_{s'}^{*,0}$ is an isomorphism for all $2 \leq s' \leq N$. In particular, the morphism $\Phi_N^{N,0} : E_N^{N,0}(\rho) \to E_N^{N,0}(\lambda)$ is the identity, as it was in the second term, and so $\Phi_N^{N,0}(b) = b$. When passing to the $(N+1)$-term, with a slight abuse of notation, we get

$$\Phi_{N+1}^{N,0}([b]) = [b],$$

where $[b]$ denotes the class induces by b in the appropriate $(N+1)$-term of the spectral sequences. Since $b := \partial_N(1 \otimes_{\mathbb{F}_p} \ell) \in E_N^{N,0}(\rho)$ and $0 \neq b \in E_2^{N,0}(\lambda) \cong E_\infty^{N,0}(\lambda)$ we have reached a contradiction:

$$\Phi_{N+1}^{N,0}(0) = [b] = b \neq 0.$$

Therefore, there cannot be any $(\mathbb{Z}/p)^n$-equivariant map $(R_{(C_1,\dots,C_{d+1})})_{\Delta(2)}^{*r} \to S(W_r^{\oplus(d+1)})$, and the proof of the theorem is complete.

□

As part of the proof of the previous theorem the following general criterion was derived.

Corollary 4.8 *Let $(C_1,\dots,C_m)$ be a coloring of the simplex Δ by m colors. If there is no $\mathfrak{S}_r$-equivariant map*

$$\Delta_{|C_1|,r} * \cdots * \Delta_{|C_m|,r} \cong (R_{(C_1,\dots,C_m)})^{*r}_{\Delta(2)} \to S(W_r^{\oplus(d+1)}),$$

then for every continuous map $f : \Delta \to \mathbb{R}^d$ there exist r pairwise disjoint rainbow faces $\sigma_1,\dots,\sigma_r$ of Δ whose f-images overlap,

$$f(\sigma_1) \cap \cdots \cap f(\sigma_r) \neq \emptyset.$$

The proof of Theorem 4.7 could have been written in the language of the Fadell–Husseini index [22]. The non-existence of an $(\mathbb{Z}/p)^n$-equivariant map $(R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)} \to S(W_r^{\oplus(d+1)})$ would then follow from the observation that

$$\operatorname{index}_{(\mathbb{Z}/p)^n}((R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)};\mathbb{F}_p) \not\supseteq \operatorname{index}_{(\mathbb{Z}/p)^n}(S(W_r^{\oplus(d+1)});\mathbb{F}_p).$$

More precisely, we have computed that

$$\begin{aligned}\operatorname{index}_{(\mathbb{Z}/p)^n}((R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)};\mathbb{F}_p) &= \operatorname{index}_{(\mathbb{Z}/p)^n}((\Delta_{2r-1,r})^{*(d+1)};\mathbb{F}_p)\\ &\subseteq H^{\geq N+d+1}(\mathrm{B}(\mathbb{Z}/p)^n;\mathbb{F}_p),\end{aligned}$$

when $|C_1| = \cdots = |C_{d+1}| = 2r-1$. Actually we proved more:

$$\operatorname{index}_{(\mathbb{Z}/p)^n}((\Delta_{2r-1,r})^{*k};\mathbb{F}_p) \subseteq H^{\geq kr}(\mathrm{B}(\mathbb{Z}/p)^n;\mathbb{F}_p). \tag{16}$$

Furthermore we have found an element $b \in H^N(\mathrm{B}(\mathbb{Z}/p)^n;\mathbb{F}_p)$ with the property that

$$0 \neq b \in \operatorname{index}_{(\mathbb{Z}/p)^n}(S(W_r^{\oplus(d+1)});\mathbb{F}_p) \cap H^N(\mathrm{B}(\mathbb{Z}/p)^n;\mathbb{F}_p),$$

and moreover

$$\operatorname{index}_{(\mathbb{Z}/p)^n}(S(W_r^{\oplus(d+1)});\mathbb{F}_p) = \langle b\rangle. \tag{17}$$

The element b with this property is the Euler class of the vector bundle

$$W_r^{\oplus(d+1)} \to \mathrm{E}(\mathbb{Z}/p)^n \times_{(\mathbb{Z}/p)^n} W_r^{\oplus(d+1)} \to \mathrm{B}(\mathbb{Z}/p)^n.$$

The work of Mann and Milgram [33] allows us to specify the element b completely: For p an odd prime it is

$$b = \omega \cdot \Big(\prod_{(\alpha_1,\dots,\alpha_n)\in\mathbb{F}_p^n\setminus\{0\}} (\alpha_1 t_1 + \cdots + \alpha_n t_n)\Big)^{(d+1)/2},$$

where $\omega \in \mathbb{F}_p\backslash\{0\}$, while for $p = 2$ it is

$$b = \Big(\prod_{(\alpha_1,\dots,\alpha_n)\in\mathbb{F}_2^n\backslash\{0\}} (\alpha_1 t_1 + \cdots + \alpha_k t_n)\Big)^{d+1}.$$

The square root in $\mathbb{F}_p[t_1,\dots,t_n]$ is not uniquely determined for an odd prime p and d odd. Thus we consider an arbitrary square root.

Combining these index computations we have that

$$\begin{aligned} 0 \neq b &\in \operatorname{index}_{(\mathbb{Z}/p)^n}(S(W_r^{\oplus(d+1)});\mathbb{F}_p) \cap H^N(\mathrm{B}(\mathbb{Z}/p)^n;\mathbb{F}_p) \\ &\not\subseteq \operatorname{index}_{(\mathbb{Z}/p)^n}((\Delta_{2r-1,r})^{*(d+1)};\mathbb{F}_p) \\ &\subseteq H^{\geq N+d+1}(\mathrm{B}(\mathbb{Z}/p)^n;\mathbb{F}_p). \end{aligned} \tag{18}$$

If a $(\mathbb{Z}/p)^n$-equivariant map $(R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)} \to S(W_r^{\oplus(d+1)})$ exists, then the monotonicity property of the Fadell–Husseini index yields the inclusion

$$\operatorname{index}_{(\mathbb{Z}/p)^n}((R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)};\mathbb{F}_p) \supseteq \operatorname{index}_{(\mathbb{Z}/p)^n}(S(W_r^{\oplus(d+1)});\mathbb{F}_p),$$

which does not hold, as we just proved. Thus the $(\mathbb{Z}/p)^n$-equivariant map in question does not exist.

Now observe the difference of dimensions in (18) and compare the dimension of the element b and the dimension of the group cohomology where the index of the join $(\Delta_{2r-1,r})^{*(d+1)}$ lives. *We could have proved more.* Indeed, using the index computation (16) we have that

$$\begin{aligned} 0 \neq b &\in \operatorname{index}_{(\mathbb{Z}/p)^n}(S(W_r^{\oplus(d+1)});\mathbb{F}_p) \cap H^N(\mathrm{B}(\mathbb{Z}/p)^n;\mathbb{F}_p) \\ &\not\subseteq \operatorname{index}_{(\mathbb{Z}/p)^n}((\Delta_{2r-1,r})^{*k};\mathbb{F}_p) \\ &\subseteq H^{\geq kr}(\mathrm{B}(\mathbb{Z}/p)^n;\mathbb{F}_p) \end{aligned}$$

as long as $kr \geq N + 1$. We have just concluded that, if $kr \geq N + 1$, then there is no $(\mathbb{Z}/p)^n$-equivariant map

$$(\Delta_{2r-1,r})^{*k} \cong (R_{(C_1,\dots,C_k)})^{*r}_{\Delta(2)} \to S(W_r^{\oplus(d+1)}).$$

Thus with Corollary 4.8 we have proved the following "colored Tverberg theorem of type B" [49, Thm. 4].

Theorem 4.9 (The Colored Tverberg theorem of type B of Vrećica and Živaljević) *Let $d \geq 1$ and $k \geq 1$ be integers, $N = (d+1)(r-1)$, and let $r \geq 2$ be a prime power. For every continuous map $f : \Delta \to \mathbb{R}^d$, and every coloring $(C_1,\dots,C_k)$ of the vertex set $\mathcal{C}$ of the simplex Δ by k colors, with each color of* *size at least* $2r-1$ *and* $kr \geq N+1$, *there exist* r *pairwise disjoint rainbow faces* $\sigma_1, \dots, \sigma_r$ *of* Δ *whose* f*-images overlap, that is*

$$f(\sigma_1) \cap \cdots \cap f(\sigma_r) \neq \emptyset.$$

As we have just seen, the proof of the colored Tverberg theorem of Živaljević and Vrećica is in fact also a proof of a type B colored Tverberg theorem. Is it possible that this proof hides a way to prove, for example the Bárány–Larman conjecture? For this we would need to prove that for some or all r and $|C_1| = \cdots = |C_{d+1}| = r$ there is no $\mathfrak{S}_r$-equivariant map

$$\Delta_{r,r}^{*(d+1)} \cong (R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)} \to S(W_r^{\oplus(d+1)}). \tag{19}$$

The connectivity of the chessboard $\Delta_{r,r}$ is only $\left(\lfloor \frac{2r+1}{3} \rfloor - 2\right)$ and therefore the scheme of the proof of Theorem 4.7 cannot be used. Even worse, the complete approach fails, as the following theorem of Blagojević, Matschke and Ziegler [17, Prop. 4.1] shows that an $\mathfrak{S}_r$-equivariant map (19) does exist.

Theorem 4.10 *Let* $r \geq 2$ *and* $d \geq 1$ *be integers. There exists an* $\mathfrak{S}_r$*-equivariant map*

$$\Delta_{r,r}^{*(d+1)} \to S(W_r^{\oplus(d+1)}).$$

Proof For this we use equivariant obstruction theory, as presented by Tammo tom Dieck [44, Sec. II.3].

Let $N := (d+1)(r-1)$, $M := r(d+1)-1$, and let $(C_1, \dots, C_{d+1})$ be a coloring of the vertex set of the simplex Δ_M by $d+1$ colors of the same size r, that is $|C_1| = \cdots = |C_{d+1}| = r$. As we know the deleted join $(R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)}$ of the rainbow complex is isomorphic to the join of chessboards $\Delta_{r,r}^{*(d+1)}$. The action of the symmetric group $\mathfrak{S}_r$ on the complex $\Delta_{r,r}^{*(d+1)}$ is not free. The subcomplex of $\Delta_{r,r}^{*(d+1)}$ whose points have non-trivial stabilizers with respect to the action of $\mathfrak{S}_r$ can be described as follows:

$$\begin{aligned}(\Delta_{r,r}^{*(d+1)})^{>1} &= ((R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)})^{>1} \\ &= \{\lambda_1 x_1 + \cdots + \lambda_r x_r \in (R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)} : \lambda_i = \lambda_j = 0 \text{ for some } i \neq j\}.\end{aligned}$$

Here for a G-space (CW complex) X we use notation $X^{>1}$ for the subspace (subcomplex) of all points (cells) with non-trivial stabilizer, meaning that $X \backslash X^{>1}$ is a free G-space.

Let $f : \Delta_M \to \mathbb{R}^d$ be any continuous map. As explained in Sect. 3.2.3 the map f induces the join map given by

$$J_f : (\Delta_M)^{*r}_{\Delta(2)} \to (\mathbb{R}^{d+1})^{\oplus r}, \qquad \lambda_1 x_1 + \cdots + \lambda_r x_r \longmapsto (\lambda_1, \lambda_1 f(x_1)) \oplus \cdots \oplus (\lambda_r, \lambda_r f(x_r)).$$

Since the rainbow complex $(R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)}$ is an $\mathfrak{S}_r$-invariant subcomplex of $(\Delta)^{*r}_{\Delta(2)}$, the restriction

$$J'_f := J_f|_{(R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)}} : (R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)} \to (\mathbb{R}^{d+1})^{\oplus r}$$

is also an $\mathfrak{S}_r$-equivariant map. Moreover $\operatorname{im}(J'_f|_{((R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)})^{>1}}) \cap D_J = \emptyset$ where, as before, $D_J = \{(z_1,\dots,z_r) \in (\mathbb{R}^{d+1})^{\oplus r} : z_1 = \cdots = z_r\}$. Thus the map J'_f induces an $\mathfrak{S}_r$-equivariant map

$$(\Delta^{*(d+1)}_{r,r})^{>1} = ((R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)})^{>1} \to (\mathbb{R}^{d+1})^{\oplus r}\backslash D_J.$$

Composing this map with the $\mathfrak{S}_r$-equivariant retraction $R_j : (\mathbb{R}^{d+1})^{\oplus r}\backslash D_J \to S(D_J^{\perp}) \cong S(W_r^{\oplus(d+1)})$ introduced in (7), we get a continuous $\mathfrak{S}_r$-equivariant map

$$(\Delta^{*(d+1)}_{r,r})^{>1} = ((R_{(C_1,\dots,C_{d+1})})^{*r}_{\Delta(2)})^{>1} \to S(W_r^{\oplus(d+1)}). \tag{20}$$

The $(r-1)$-dimensional chessboard complex $\Delta_{r,r}$ equivariantly retracts to a subcomplex of dimension $r-2$. Indeed, for each facet of $\Delta_{r,r}$ there is an elementary collapse obtained by deleting all of its subfacets (faces of dimension $r-2$) that contain the vertex in the r-th column. Performing these collapses to all facets of $\Delta_{r,r}$, we get that $\Delta_{r,r}$ collapses $\mathfrak{S}_r$-equivariantly to an $(r-2)$-dimensional subcomplex of $\Delta_{r,r}$. Consequently, the join $(\Delta_{r,r})^{*(d+1)}$ equivariantly retracts to a subcomplex K of dimension $(d+1)(r-1)-1$. Thus in order to prove the existence of an $\mathfrak{S}_r$-equivariant map $\Delta^{*(d+1)}_{r,r} \to S(W_r^{\oplus(d+1)})$ it suffices to construct an $\mathfrak{S}_r$-equivariant map $K \to S(W_r^{\oplus(d+1)})$. Since

- $\dim K = \dim S(W_r^{\oplus(d+1)}) = N-1$,
- $S(W_r^{\oplus(d+1)})$ is $(N-1)$-simple dict and $(N-2)$-connected,

and the groups where the obstructions would live are zero, the equivariant obstruction theory yields the existence of an $\mathfrak{S}_r$-equivariant map $K \to S(W_r^{\oplus(d+1)})$, provided that an $\mathfrak{S}_r$-equivariant map $K^{>1} \to S(W_r^{\oplus(d+1)})$ exists. The subcomplex of all points with non-trivial stabilizer $K^{>1} = K \cap (\Delta^{*(d+1)}_{r,r})^{>1}$ is a subcomplex of $(\Delta^{*(d+1)}_{r,r})^{>1}$ and therefore the map (20) restricted to $K^{>1}$ completes the argument.

□

After this theorem an urgent question emerges: *How are we going to handle the Bárány–Larman conjecture?* An answer to this question will bring us to our last section and the optimal colored Tverberg theorem.

4.4 The Weak Colored Tverberg Theorem

How many colored Tverberg theorems can we get directly from the topological Tverberg theorem without major topological machinery? Here is an answer given by [12, Thm. 5.3].

Theorem 4.11 (The weak colored Tverberg theorem) *Let $d \geq 1$ be an integer, let r be a prime power, $N = (2d + 2)(r - 1)$, and let $f : \Delta_N \to \mathbb{R}^d$ be a continuous map. If the vertices of the simplex Δ_N are colored by $d + 1$ colors, where each color class has cardinality at most $2r - 1$, then there are r pairwise disjoint rainbow faces $\sigma_1, \dots, \sigma_r$ of Δ_N whose f-images overlap, that is*

$$f(\sigma_1) \cap \cdots \cap f(\sigma_r) \neq \emptyset.$$

Proof Let $\mathcal{C}$ be the set of vertices of the simplex Δ_N and let $(C_1, \dots, C_{d+1})$ be a coloring of $\mathcal{C}$ where $|C_i| \leq 2r - 1$ for all $1 \leq i \leq d + 1$. To each color class C_i we associate the subcomplex Σ_i of Δ_N defined by

$$\Sigma_i := \{\sigma \in \Delta_N : |\sigma \cap C_i| \leq 1\}.$$

Observe that the intersection $\Sigma_1 \cap \cdots \cap \Sigma_{d+1}$ is the subcomplex of all rainbow faces of Δ_N with respect to the given coloring. Next consider the continuous map $g : \Delta_N \to \mathbb{R}^{2d+1}$ defined by

$$g(x) = (f(x), \operatorname{dist}(x, \Sigma_1), \operatorname{dist}(x, \Sigma_2), \dots, \operatorname{dist}(x, \Sigma_{d+1})).$$

Since $N = (2d + 2)(r - 1) = ((2d + 1) + 1)(r - 1)$ and r is a prime power, we can apply the topological Tverberg theorem to g. Consequently there are r pairwise disjoint faces $\sigma_1, \dots, \sigma_r$ with points $x_1 \in \operatorname{relint} \sigma_1, \dots, x_r \in \operatorname{relint} \sigma_r$ such that $g(x_1) = \cdots = g(x_r)$, that is,

$$\begin{aligned} f(x_1) &= \cdots = f(x_r), \\ \operatorname{dist}(x_1, \Sigma_1) &= \cdots = \operatorname{dist}(x_r, \Sigma_1), \\ &\cdots \\ \operatorname{dist}(x_1, \Sigma_{d+1}) &= \cdots = \operatorname{dist}(x_r, \Sigma_{d+1}). \end{aligned}$$

Now observe that for every subcomplex Σ_i one of the faces $\sigma_1, \dots, \sigma_r$ is contained in it. Indeed, if this would not hold then we would have $|\sigma_1 \cap C_i| \geq 2, \dots, |\sigma_r \cap C_i| \geq 2$, and consequently we would obtain the following contradiction:

$$2r - 1 \geq |C_i| \geq |\sigma_1 \cap C_i| + \cdots + |\sigma_r \cap C_i| \geq 2r.$$

Hence the distances, which were previously know to be equal, have to vanish,

$$\operatorname{dist}(x_1, \Sigma_1) = \cdots = \operatorname{dist}(x_r, \Sigma_1) = 0,$$

$$\cdots$$

$$\operatorname{dist}(x_1, \Sigma_{d+1}) = \cdots = \operatorname{dist}(x_r, \Sigma_{d+1}) = 0,$$

implying that $x_i \in \Sigma_1 \cap \cdots \cap \Sigma_{d+1}$ for every $1 \le i \le r$. Since $\Sigma_1, \dots, \Sigma_{d+1}$ are subcomplexes of Δ_N and $x_1 \in \operatorname{relint} \sigma_1, \dots, x_r \in \operatorname{relint} \sigma_r$ it follows that the faces $\sigma_1, \dots, \sigma_r$ belong to the subcomplex $\Sigma_1 \cap \cdots \cap \Sigma_{d+1}$, that is, $\sigma_1, \dots, \sigma_r$ are rainbow faces. □

A special case of the weak colored Tverberg theorem we just proved, namely $|C_1| = \cdots = |C_{d+1}| = 2r - 1$, yields $t(d, r) \le tt(d, r) \le 2r - 1$ for r a prime power. This is the colored Tverberg theorem of Živaljević and Vrećica presented in Theorem 4.7.

Along the lines of the previous theorem we can prove the following colored Van Kampen–Flores theorem, where the number of color classes is at most $d + 1$.

Theorem 4.12 (The colored Van Kampen–Flores theorem) *Let $d \ge 1$ be an integer, let r be a prime power, let $k \ge \lceil d\frac{r-1}{r} \rceil + 1$ be an integer, and $N = (d + k + 1)(r - 1)$. Let $f : \Delta_N \to \mathbb{R}^d$ be a continuous map. If the vertices of the simplex Δ_N are colored by k colors, where each color class has cardinality at most $2r-1$, then there are r pairwise disjoint rainbow faces $\sigma_1, \dots, \sigma_r$ of Δ_N whose f-images overlap,*

$$f(\sigma_1) \cap \cdots \cap f(\sigma_r) \neq \emptyset.$$

Proof Let $\mathcal{C}$ be the set of vertices of the simplex Δ_N and let $(C_1, \dots, C_k)$ be a coloring where $|C_i| \le 2r - 1$ for all $1 \le i \le k$. Such a coloring exists because $k(2r - 1) \ge (d + k + 1)(r - 1)$ is equivalent to our assumption $k \ge \lceil d\frac{r-1}{r} \rceil + 1$. To each color class C_i we associate the subcomplex Σ_i of Δ_N defined as before by

$$\Sigma_i := \{\sigma \in \Delta_N : |\sigma \cap C_i| \le 1\}.$$

The subcomplex $\Sigma_1 \cap \cdots \cap \Sigma_k$ is a subcomplex of all rainbow faces of Δ_N with respect to the given coloring. Consider the continuous map $g : \Delta_N \to \mathbb{R}^{d+k}$ defined by

$$g(x) = (f(x), \operatorname{dist}(x, \Sigma_1), \operatorname{dist}(x, \Sigma_2), \dots, \operatorname{dist}(x, \Sigma_k)).$$

Since $N = (d+k+1)(r-1)$ and r is a prime power the topological Tverberg theorem can be applied to the map g. Therefore, there are r pairwise disjoint faces $\sigma_1, \dots, \sigma_r$

with points $x_1 \in \operatorname{relint}\sigma_1, \dots, x_r \in \operatorname{relint}\sigma_r$ such that $g(x_1) = \cdots = g(x_r)$, that is,

$$\begin{aligned} f(x_1) &= \cdots = f(x_r), \\ \operatorname{dist}(x_1, \Sigma_1) &= \cdots = \operatorname{dist}(x_r, \Sigma_1), \\ &\cdots \\ \operatorname{dist}(x_1, \Sigma_k) &= \cdots = \operatorname{dist}(x_r, \Sigma_k). \end{aligned}$$

Now observe that every subcomplex Σ_i contains one of the faces $\sigma_1, \dots, \sigma_r$. Indeed, if this would not hold then $|\sigma_1 \cap C_i| \geq 2, \dots, |\sigma_r \cap C_i| \geq 2$, and we would get the contradiction

$$2r - 1 \geq |C_i| \geq |\sigma_1 \cap C_i| + \cdots + |\sigma_r \cap C_i| \geq 2r.$$

Consequently the distances, which were previously known to be equal, have to vanish

$$\operatorname{dist}(x_1, \Sigma_1) = \cdots = \operatorname{dist}(x_r, \Sigma_1) = 0, \ \cdots, \ \operatorname{dist}(x_1, \Sigma_k) = \cdots = \operatorname{dist}(x_r, \Sigma_k) = 0,$$

implying that $x_i \in \Sigma_1 \cap \cdots \cap \Sigma_k$ for every $1 \leq i \leq r$. Since $\Sigma_1, \dots, \Sigma_k$ are subcomplexes and $x_1 \in \operatorname{relint}\sigma_1, \dots, x_r \in \operatorname{relint}\sigma_r$, it follows that

$$\sigma_1 \in \Sigma_1 \cap \cdots \cap \Sigma_k, \ \dots \ , \sigma_r \in \Sigma_1 \cap \cdots \cap \Sigma_k,$$

that is, $\sigma_1, \dots, \sigma_r$ are rainbow faces. □

The "colored Tverberg theorem of type B" of Vrećica and Živaljević, Theorem 4.9, is a particular case of this theorem, when the color classes have the same size.

4.5 *Tverberg Points with Equal Barycentric Coordinates*

The last corollary of the Topological Tverberg theorem that we present here is the topological version [12, Thm. 8.1] of a recent result by Soberón [42, Thm. 1.1] [43, Thm. 1].

Let $N \geq 1$ be an integer, let $\mathcal{C}$ be the set of vertices of the simplex Δ_N, and let $(C_1, \dots, C_\ell)$ be a coloring of $\mathcal{C}$. Every point x in the rainbow subcomplex $R_{(C_1,\dots,C_\ell)}$ has a unique presentation in barycentric coordinates as $x = \sum_{i=1}^{\ell} \lambda_i^x v_i^x$ where $0 \leq \lambda_i^x \leq 1$ and $v_i^x \in C_i$ for all $0 \leq i \leq \ell - 1$. Two points $x = \sum_{i=1}^{\ell} \lambda_i^x v_i^x$ and $y = \sum_{i=1}^{\ell} \lambda_i^y v_i^y$ in the rainbow subcomplex $R_{(C_1,\dots,C_\ell)}$ have *equal barycentric coordinates* if $\lambda_i^x = \lambda_i^y$ for all $1 \leq i \leq \ell$.

Theorem 4.13 *Let $d \geq 1$ be an integer, let r be a prime power, $N = r((r-1)d+1)-1 = (r-1)(rd+1)$, and let $f : \Delta_N \to \mathbb{R}^d$ be a continuous map. If the vertices of the simplex Δ_N are colored by $(r-1)d+1$ colors where each colored class is of size r, then there are points $x_1, \dots, x_r$ with equal barycentric coordinates that belong to r pairwise disjoint rainbow faces $\sigma_1, \dots, \sigma_r$ of Δ_N whose f-images coincide, that is*

$$f(x_1) = \cdots = f(x_r).$$

Proof Let $\ell = (r-1)d+1$, and let $(C_1, \dots, C_\ell)$ be a coloring of the vertex set $\mathcal{C} = \{v_0, \dots, v_N\}$ of the simplex Δ_N. Each point x of the simplex Δ_N can be uniquely presented in the barycentric coordinates as $x = \sum_{j=0}^{N} \lambda_j^x v_j$. For every color class C_i, $1 \leq i \leq \ell$, we define the function $h_i : \Delta_N \to \mathbb{R}$ by $h_i\big(\sum_{j=0}^{N} \lambda_j^x v_j\big) = \sum_{v_j \in C_i} \lambda_j^x$. All functions h_j are affine functions and $\sum_{i=1}^{\ell} h_i(x) = \sum_{j=0}^{N} \lambda_j^x = 1$ for every $x \in \Delta_N$.

Now consider the function $g : \Delta_N \to \mathbb{R}^{rd}$ given by

$$g(x) = (f(x), h_1(x), \dots, h_{\ell-1}(x)).$$

Since $N = (r-1)(rd+1)$, the topological Tverberg theorem applied to the function g implies that there exist r pairwise disjoint faces $\sigma_1, \dots, \sigma_r$ of Δ_N and r points $x_1 \in \operatorname{relint} \sigma_1, \dots, x_r \in \operatorname{relint} \sigma_r$ such that $f(x_1) = \cdots = f(x_r)$ and $h_i(x_1) = \cdots = h_i(x_r)$ for $1 \leq i \leq \ell - 1$. In addition, the equality $\sum_{i=1}^{\ell} h_i(x) = 1$ implies that also $h_\ell(x_1) = \cdots = h_\ell(x_r)$.

Assume now that $|\sigma_j \cap C_i| \geq 1$ for some $1 \leq j \leq r$ and some $1 \leq i \leq \ell$. Then $h_i(x_j) > 0$ since $x_j \in \operatorname{relint} \sigma_j$. Consequently, $h_i(x_1) = \cdots = h_i(x_r) > 0$ implying that $|\sigma_j \cap C_i| \geq 1$ for all $1 \leq j \leq r$. Since $|C_i| = r$ and $\sigma_1, \dots, \sigma_r$ are pairwise disjoint it follows that each σ_j has precisely one vertex in the color class C_i. Thus, repeating the argument for each color class we conclude that all faces $\sigma_1, \dots, \sigma_r$ are rainbow faces. The immediate consequence of this fact is that $h_i(x_j)$, $1 \leq i \leq \ell$, are the barycentric coordinates of the point x_i and so all the points $x_1, \dots, x_r$ have equal barycentric coordinates. □

5 Counterexamples to the Topological Tverberg Conjecture

Now we are going to get to a very recent piece of the topological Tverberg puzzle: We show how counterexamples to the topological Tverberg conjecture for any number of parts that his not a prime power were derived by Frick [13, 25] from the remarkable works of Özaydin [36] and of Mabillard and Wagner [31, 32], via a lemma of Gromov [26, p. 445] that is an instance of the constraint method of Blagojević, Frick, Ziegler [12, Lemmas 4.1(iii) and 4.2].

5.1 Existence of Equivariant Maps if r *is not a prime power*

First we present the second main result of Özaydin's landmark manuscript [36, Thm. 4.2].

Theorem 5.1 *Let $d \geq 1$ and $r \geq 6$ be integers, and let $N = (d+1)(r-1)$. If r is not a prime power, then there exists an $\mathfrak{S}_r$-equivariant map*

$$(\Delta_N)^{\times r}_{\Delta(2)} \to S(W_r^{\oplus d}). \tag{21}$$

Proof In order to prove the existence of a continuous $\mathfrak{S}_r$-equivariant map $(\Delta_N)^{\times r}_{\Delta(2)} \to S(W_r^{\oplus d})$ we again use the equivariant obstruction theory. Since

- $(\Delta_N)^{\times r}_{\Delta(2)}$ is an $(N-r+1)$-dimensional, $(N-r)$-connected free $\mathfrak{S}_r$-CW complex, and
- $S(W_r^{\oplus d})$ is a path-connected $(N-r-1)$-connected, $(N-r)$-simple $\mathfrak{S}_r$-space,

we have that an $\mathfrak{S}_r$-equivariant map $(\Delta_N)^{\times r}_{\Delta(2)} \to S(W_r^{\oplus d})$ exists if and only if the primary obstruction $^{\text{dict}}$

$$[\mathfrak{o}_{\mathfrak{S}_r}^{N-r+1}(\mathrm{pt})] \in \mathcal{H}_{\mathfrak{S}_r}^{N-r+1}((\Delta_N)^{\times r}_{\Delta(2)}, \pi_{N-r}S(W_r^{\oplus d}))$$

vanishes. The obstruction element $[\mathfrak{o}_{\mathfrak{S}_r}^{N-r+1}(\mathrm{pt})] = [\mathfrak{o}_{\mathfrak{S}_r}^{N-r+1}(f)]$ does not depend on the particular $\mathfrak{S}_r$-equivariant map $f : \mathrm{sk}_{N-r}\big((\Delta_N)^{\times r}_{\Delta(2)}\big) \to S(W_r^{\oplus d})$ used to define the obstruction cocycle $\mathfrak{o}_{\mathfrak{S}_r}^{N-r+1}(f)$. Thus, in order to prove the existence of an $\mathfrak{S}_r$-equivariant map (21) it suffices to prove that the obstruction element $[\mathfrak{o}_{\mathfrak{S}_r}^{N-r+1}(f)]$ vanishes for some particular choice of f.

Let p be a prime such that $p \mid |\mathfrak{S}_r| = r!$, and let $\mathfrak{S}_r^{(p)}$ denotes a p-Sylow subgroup of $\mathfrak{S}_r$. Since r is not a prime power each p-Sylow subgroup of $\mathfrak{S}_r$ does not act transitively on the set $[r]$, and hence the fixed point set $S(W_r^{\oplus d})^{\mathfrak{S}_r^{(p)}} \neq \emptyset$ is non-empty. Thus there exists a (constant) $\mathfrak{S}_r^{(p)}$-equivariant map $(\Delta_N)^{\times r}_{\Delta(2)} \to S(W_r^{\oplus d})$, or equivalently the primary obstruction element with respect to $\mathfrak{S}_r^{(p)}$ vanishes, that is, $[\mathfrak{o}_{\mathfrak{S}_r^{(p)}}^{N-r+1}(\mathrm{pt})] = [\mathfrak{o}_{\mathfrak{S}_r^{(p)}}^{N-r+1}(f)] = 0$. Here, the $\mathfrak{S}_r$-equivariant map f is considered only as an $\mathfrak{S}_r^{(p)}$-equivariant map.

The restriction $^{\text{dict}}$ homomorphism

$$\mathrm{res} : \mathcal{H}_{\mathfrak{S}_r}^{N-r+1}((\Delta_N)^{\times r}_{\Delta(2)}, \pi_{N-r}S(W_r^{\oplus d})) \to \mathcal{H}_{\mathfrak{S}_r^{(p)}}^{N-r+1}((\Delta_N)^{\times r}_{\Delta(2)}, \pi_{N-r}S(W_r^{\oplus d})),$$

is defined on the cochain level in [14, Lem. 5.4]. According to the definition of the obstruction cochain (already on the cochain level) the restriction homomorphism sends the obstruction cochain $\mathfrak{o}^{N-r+1}_{\mathfrak{S}_r}(f)$ to the obstruction cochain $\mathfrak{o}^{N-r+1}_{\mathfrak{S}_r^{(p)}}(f)$. Consequently the same hold for obstruction elements

$$\mathrm{res}([\mathfrak{o}^{N-r+1}_{\mathfrak{S}_r}(f)]) = [\mathfrak{o}^{N-r+1}_{\mathfrak{S}_r^{(p)}}(f)].$$

Now, composing the restriction homomorphism with the transfer dict homomorphism

$$\mathrm{tr} : \mathcal{H}^{N-r+1}_{\mathfrak{S}_r^{(p)}}((\Delta_N)^{\times r}_{\Delta(2)}, \pi_{N-r}S(W_r^{\oplus d})) \to \mathcal{H}^{N-r+1}_{\mathfrak{S}_r}((\Delta_N)^{\times r}_{\Delta(2)}, \pi_{N-r}S(W_r^{\oplus d})),$$

also defined on the cochain level in [14, Lem. 5.4], we get

$$[\mathfrak{S}_r : \mathfrak{S}_r^{(p)}] \cdot [\mathfrak{o}^{N-r+1}_{\mathfrak{S}_r}(f)] = \mathrm{tr} \circ \mathrm{res}([\mathfrak{o}^{N-r+1}_{\mathfrak{S}_r}(f)]) = \mathrm{tr}([\mathfrak{o}^{N-r+1}_{\mathfrak{S}_r^{(p)}}(f)]) = \mathrm{tr}(0) = 0.$$

Finally, since $[\mathfrak{S}_r : \mathfrak{S}_r^{(p)}] \cdot [\mathfrak{o}^{N-r+1}_{\mathfrak{S}_r}(f)] = 0$ for every prime p that divides the order of the group $\mathfrak{S}_r$, it follows that the obstruction element $[\mathfrak{o}^{N-r+1}_{\mathfrak{S}_r}(f)]$ must vanish, and the existence of an $\mathfrak{S}_r$-equivariant map (21) is established. □

Corollary 5.2 *Let $d \geq 1$ be an integer, let $r \geq 6$ be an integer that is not a prime power and let $N = (d+1)(r-1)$. For any free $\mathfrak{S}_r$-CW complex X of dimension at most $N - r + 1$ there exists an $\mathfrak{S}_r$-equivariant map*

$$X \to S(W_r^{\oplus d}).$$

Proof The free $\mathfrak{S}_r$-CW complex X has dimension at most $N - r + 1$, and the deleted product $(\Delta_N)^{\times r}_{\Delta(2)}$ is $(N-r)$-connected, therefore there are no obstructions for the existence of an $\mathfrak{S}_r$-equivariant map $h : X \to (\Delta_N)^{\times r}_{\Delta(2)}$. Next, let $f : (\Delta_N)^{\times r}_{\Delta(2)} \to S(W_r^{\oplus d})$ be an $\mathfrak{S}_r$-equivariant map whose existence was guaranteed by Theorem 5.1. The composition $f \circ h : X \to S(W_r^{\oplus d})$ yields the required $\mathfrak{S}_r$-equivariant map. □

With this theorem Özaydin *only* proved that the "deleted product approach" towards solving the topological Tverberg conjecture fails in the case that r is not a prime power. *What about the "deleted join approach"?* This question was discussed in [16, Sec. 3.4].

Theorem 5.3 *Let $d \geq 1$ and r be integers, and let $N = (d+1)(r-1)$. If r is not a prime power, then there exists an $\mathfrak{S}_r$-equivariant map*

$$(\Delta_N)^{*r}_{\Delta(2)} \to S(W_r^{\oplus(d+1)}). \tag{22}$$

Proof Since $r \geq 6$ is not a prime power Theorem 5.1 implies the existence of an $\mathfrak{S}_r$-equivariant map

$$f : (\Delta_N)^{\times r}_{\Delta(2)} \to S(W_r^{\oplus d}).$$

Now an $\mathfrak{S}_r$-equivariant map

$$g : (\Delta_N)^{*r}_{\Delta(2)} \to S(W_r^{\oplus(d+1)}) \cong S(W_r \oplus W_r^{\oplus d})$$

can be defined by

$$g(\lambda_1 x_1 + \cdots + \lambda_r x_r) = \frac{1}{\nu}\big((\lambda_1 - \tfrac{1}{r}, \ldots, \lambda_r - \tfrac{1}{r}) \oplus \prod_{i=1}^{r} \lambda_i \cdot f(x_1, \ldots, x_r)\big),$$

where $\nu := \|\big((\lambda_1 - \frac{1}{r}, \ldots, \lambda_r - \frac{1}{r}) \oplus \prod_{i=1}^{r} \lambda_i \cdot f(x_1, \ldots, x_r)\big)\|$. The function g is well defined, continuous and $\mathfrak{S}_r$-equivariant. Thus an $\mathfrak{S}_r$-equivariant map (22) exists.

□

Now we see that not only the "deleted product approach" fails if r is not a prime power, but the "deleted join approach" fails as well. *Is this an indication that the topological Tverberg theorem fails if the number of parts is not a prime power?*

5.2 *The Topological Tverberg Conjecture does not hold if* r *is not a prime power*

It is time to show that the topological Tverberg conjecture fails in the case that r is not a prime power. This will be done following the presentation given in [13].

Based on the work of Mabillard and Wagner [31, 32] we will prove that the generalized Van Kampen–Flores theorem for any r that is not a prime power fails, as demonstrated by Frick [13, 25]. Since, by the constraint method, the generalized Van Kampen–Flores theorem for fixed number of overlaps r is a consequence of the topological Tverberg theorem for the same number of overlaps r, failure of the generalized Van Kampen–Flores theorem implies the failure of the topological Tverberg theorem.

Theorem 5.4 (The generalized Van Kampen–Flores theorem fails when *r* is not a prime power) *Let $k \geq 3$ be an integer, and let $r \geq 6$ be an integer that is not a prime power. For any integer $N > 0$ there exists a continuous map $f : \Delta_N \to \mathbb{R}^{rk}$ such that for any r pairwise disjoint faces $\sigma_1, \ldots, \sigma_r$ from the $((r-1)k)$-skeleton $\mathrm{sk}_{(r-1)k}(\Delta_N)$ of the simplex Δ_N the corresponding f-images do not overlap,*

$$f(\sigma_1) \cap \cdots \cap f(\sigma_r) = \emptyset.$$

Proof The deleted product $(\mathrm{sk}_{(r-1)k}(\Delta_N))^{\times r}_{\Delta(2)}$ is a free $\mathfrak{S}_r$-space of dimension at most $d := (r-1)rk$. Since r is not a power of a prime, according to Corollary 5.2, there exists an $\mathfrak{S}_r$-equivariant map

$$h : (\mathrm{sk}_{(r-1)k}(\Delta_N))^{\times r}_{\Delta(2)} \to S(W_r^{\oplus d}). \tag{23}$$

Now we use the following result of Mabillard and Wagner [31, Thm. 3], [32, Thm. 7], for which an alternative proof is given in [3]. Skopenkov [40] gives a user's guide.

Theorem *Let $r \geq 2$ and $k \geq 3$ be integers, and let K be an $((r-1)k)$-dimensional simplicial complex. Then the following statements are equivalent:*

(i) *There exists a continuous $\mathfrak{S}_r$-equivariant map $K^{\times r}_{\Delta(2)} \to S(W_r^{\oplus rk})$.*

(ii) *There exists a continuous map $f : K \to \mathbb{R}^{rk}$ such that for any r pairwise disjoint faces $\sigma_1, \dots, \sigma_r$ of K we have that $f(\sigma_1) \cap \cdots \cap f(\sigma_r) = \emptyset$.*

If we apply this result to the $\mathfrak{S}_r$-equivariant map h in (23) we get a continuous map $f : \mathrm{sk}_{(r-1)k}(\Delta_N) \to \mathbb{R}^{rk}$ with the property that for any collection of r pairwise disjoint faces $\sigma_1, \dots, \sigma_r$ in $\mathrm{sk}_{(r-1)k}(\Delta_N)$ the corresponding f-images do not overlap,

$$f(\sigma_1) \cap \cdots \cap f(\sigma_r) = \emptyset.$$

□

Thus we have proved that in the case when r is not a prime power the generalized Van Kampen–Flores theorem fails. As we have pointed out this means that the corresponding topological Tverberg theorem also fails [13, Thm. 4.3].

Theorem 5.5 (The topological Tverberg theorem fails for any r that is not a prime power) *Let $k \geq 3$ and $r \geq 6$ be integers, and let $N = (r-1)(rk+2)$. If r is not a prime power, then there exists a continuous map $g : \Delta_N \to \mathbb{R}^{rk+1}$ such that for any r pairwise disjoint faces $\sigma_1, \dots, \sigma_r$ of Δ_N the corresponding g images do not overlap,*

$$g(\sigma_1) \cap \cdots \cap g(\sigma_r) = \emptyset.$$

Proof Since r is not a power of a prime, Theorem 5.4 yields a continuous map $f : \Delta_N \to \mathbb{R}^{rk}$ such that for any r pairwise disjoint faces $\sigma_1, \dots, \sigma_r$ in $\mathrm{sk}_{(r-1)k}\,\Delta_N$

$$f(\sigma_1) \cap \cdots \cap f(\sigma_r) = \emptyset.$$

Motivated by the proof of Theorem 4.1 we consider the function $g : \Delta_N \to \mathbb{R}^{rk+1}$ defined by

$$g(x) = (f(x), \mathrm{dist}(x, \mathrm{sk}_{(r-1)k}(\Delta_N))).$$

We prove that the map g fails the topological Tverberg conjecture.

Assume, to the contrary, that there are r pairwise disjoint faces $\sigma_1, \dots, \sigma_r$ in Δ_N and r points

$$x_1 \in \mathrm{relint}\,\sigma_1, \ \dots, \ x_r \in \mathrm{relint}\,\sigma_r$$

such that $g(x_1) = \cdots = g(x_r)$. Consequently,

$$\mathrm{dist}(x_1, \mathrm{sk}_{(r-1)k}(\Delta_N)) = \cdots = \mathrm{dist}(x_r, \mathrm{sk}_{(r-1)k}(\Delta_N)).$$

Next, at least one of the faces $\sigma_1, \ldots, \sigma_r$ is in $\mathrm{sk}_{(r-1)k}(\Delta_N)$. Indeed, if all the faces σ_i would have dimension at least $(r-1)k+1$, then we would get the following contradiction:

$$\begin{aligned} N+1 = (r-1)(rk+2)+1 = |\Delta_N| \geq |\sigma_1| + \cdots + |\sigma_r| \geq ((r-1)rk+2) \\ = (r-1)(rk+2)+2 > N+1. \end{aligned}$$

Since one of the faces $\sigma_1, \ldots, \sigma_r$ is in $\mathrm{sk}_{(r-1)k}(\Delta_N)$, all the distances vanish, meaning that

$$\mathrm{dist}(x_1, \mathrm{sk}_{(r-1)k}(\Delta_N)) = \cdots = \mathrm{dist}(x_r, \mathrm{sk}_{(r-1)k}(\Delta_N)) = 0.$$

Therefore, all the faces $\sigma_1, \ldots, \sigma_r$ belong to $\mathrm{sk}_{(r-1)k}(\Delta_N)$ contradicting the choice of the map f. Thus the map g is a counterexample to the topological Tverberg theorem. □

Remark 5.6 The smallest counterexample to the topological Tverberg theorem that can be obtained from Theorem 5.5 is a continuous map $\Delta_{100} \rightarrow \mathbb{R}^{19}$ with the property that no six pairwise disjoint faces in Δ_{100} have f-images that overlap. Recently, using additional ideas, Sergey Avvakumov, Isaac Mabillard, Arkadiy Skopenkov and Uli Wagner [3] have improved this to get counterexamples $\Delta_{65} \rightarrow \mathbb{R}^{12}$.

6 The Bárány–Larman Conjecture and the Optimal Colored Tverberg Theorem

Let us briefly recall the original colored Tverberg problem posed by Bárány and Larman in their 1992 paper [7], see Sect. 4.2.

Problem 6.1 (Bárány–Larman colored Tverberg problem) Let $d \geq 1$ and $r \geq 2$ be integers. Determine the smallest number $n = n(d, r)$ such that for every affine (continuous) map $f : \Delta_{n-1} \rightarrow \mathbb{R}^d$, and every coloring $(C_1, \ldots, C_{d+1})$ of the vertex set $\mathcal{C}$ of the simplex Δ_{n-1} by $d+1$ colors with each color of size at least r, there exist r pairwise disjoint rainbow faces $\sigma_1, \ldots, \sigma_r$ of Δ_{n-1} whose f-images overlap, that is

$$f(\sigma_1) \cap \cdots \cap f(\sigma_r) \neq \emptyset.$$

A trivial lower bound for the function $n(d, r)$ is $(d+1)r$ and it is natural to conjecture the following.

Conjecture 6.2 (Bárány–Larman Conjecture) *Let $r \geq 2$ and $d \geq 1$ be integers. Then $n(d, r) = (d + 1)r$.*

In Sect. 4.3 we tried to use the approach of Živaljević and Vrećica to solve the Bárány–Larman conjecture and we failed dramatically. We hoped to prove that an $\mathfrak{S}_r$-equivariant map

$$\Delta_{r,r}^{*(d+1)} \to S(W_r^{\oplus(d+1)})$$

does not exist, but Theorem 4.10 gave us exactly the opposite, the existence of this map. *What can we do now?* We change the question, and prove the non-existence of an $\mathfrak{S}_{r+1}$-equivariant map

$$(R_{(C_1,\ldots,C_{d+2})})^{*r+1}_{\Delta(2)} \cong \Delta_{r,r+1}^{*(d+1)} * [r + 1] \to S(W_{r+1}^{\oplus(d+1)})$$

instead; here $|C_1| = \cdots = |C_{d+1}| = r$, and $|C_{d+2}| = 1$. *Still, why should we be interested in such a result?*

Theorem 6.3 *Let $r \geq 2$ and $d \geq 1$ be integers. If there is no $\mathfrak{S}_{r+1}$-equivariant map*

$$\Delta_{r,r+1}^{*(d+1)} * [r + 1] \to S(W_{r+1}^{\oplus(d+1)}),$$

then $n(d, r) = (d + 1)r$ and $tt(d, r) = r$.

Proof Let $(C_1, \ldots , C_{d+1})$ be a coloring of the vertices of the simplex Δ with $|C_1| = \cdots = |C_{d+1}| = r$, and let $f : \Delta \to \mathbb{R}^d$ be a continuous map. Construct a simplex Δ' as a pyramid over Δ, and let C_{d+2} be the additional color class containing only the apex of the pyramid. Thus, $(C_1, \ldots , C_{d+2})$ is a coloring of the vertices of the simplex Δ'.

Let us assume that an $\mathfrak{S}_{r+1}$-equivariant map $\Delta_{r,r+1}^{*(d+1)} * [r + 1] \to S(W_{r+1}^{\oplus(d+1)})$ does not exist. The non-existence of this map in combination with Corollary 4.8 implies that there exist $r + 1$ pairwise disjoint rainbow faces $\sigma_1, \ldots , \sigma_{r+1}$ of the simplex Δ' whose f-images overlap, $f(\sigma_1) \cap \cdots \cap f(\sigma_{r+1}) \neq \emptyset$.

Without loss of generality we can assume that $\sigma_{r+1} \cap C_{d+2} \neq \emptyset$. Then the faces $\sigma_1, \ldots , \sigma_r$ are rainbow faces of the simplex Δ with respect to the coloring $(C_1, \ldots , C_{d+1})$ and

$$f(\sigma_1) \cap \cdots \cap f(\sigma_r) \neq \emptyset.$$

Hence, $tt(d, r) = r$ and consequently $n(d, r) = (d + 1)r$. □

The theorem we just proved tells us that in order to make an advance on the Bárány–Larman conjecture we should try to prove the non-existence of a continuous $\mathfrak{S}_{r+1}$-equivariant map

$$\Delta_{r,r+1}^{*(d+1)} * [r+1] \to S(W_{r+1}^{\oplus(d+1)}) \tag{24}$$

at least for some values of r.

Now in order to prove the non-existence of a continuous $\mathfrak{S}_{r+1}$-equivariant map (24), for $r+1 =: p$ an odd prime, we will compute the Fadell–Husseini index of the join $\Delta_{r,r+1}^{*(d+1)} * [r+1] = \Delta_{p-1,p}^{*(d+1)} * [p]$ with respect to the cyclic group and compare the result with the index of the sphere $S(W_{r+1}^{\oplus(d+1)})$.

6.1 The Fadell–Husseini Index of Chessboards

Let $p := r+1$ be an odd prime. In this section we compute the Fadell–Husseini index of chessboards

$$\mathrm{index}_{\mathbb{Z}/p}(\Delta_{k,p}; \mathbb{F}_p) \subseteq H^*(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p) = H^*(\mathbb{Z}/p; \mathbb{F}_p), \qquad k \geq 1,$$

and their joins with respect to the cyclic subgroup $\mathbb{Z}/p$ of the symmetric group $\mathfrak{S}_p$. Recall that

$$H^*(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p) = H^*(\mathbb{Z}/p; \mathbb{F}_p) = \mathbb{F}_p[t] \otimes \Lambda[e],$$

where $\deg t = 2$, $\deg e = 1$, and $\Lambda[\,\cdot\,]$ denotes the exterior algebra. First, we collect some simple facts about the Fadell–Husseini index of chessboards.

Lemma 6.4 *Let $k \geq 1$ be an integer, and let p be an odd prime. Then*

(i) $\mathrm{index}_{\mathbb{Z}/p}\, \Delta_{1,p} = H^{\geq 1}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p)$,
(ii) $\mathrm{index}_{\mathbb{Z}/p}\, \Delta_{2p-1,p} = H^{\geq p}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p)$,
(iii) $\mathrm{index}_{\mathbb{Z}/p}\, \Delta_{1,p} \subseteq \mathrm{index}_{\mathbb{Z}/p}\, \Delta_{2,p} \subseteq \cdots \subseteq \mathrm{index}_{\mathbb{Z}/p}\, \Delta_{2p-1,p} = \mathrm{index}_{\mathbb{Z}/p}\, \Delta_{2p,p} = \cdots = H^{\geq p}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p)$.

Proof For the statement (i) observe that $\Delta_{1,p} = [p]$ and therefore $\mathrm{E}(\mathbb{Z}/p) \times_{\mathbb{Z}/p} \Delta_{1,p} \cong \mathrm{E}(\mathbb{Z}/p)$. Since $\mathrm{E}(\mathbb{Z}/p)$ is a contractible space,

$$\mathrm{index}_{\mathbb{Z}/p}\, \Delta_{1,p} = \ker\big(H^*(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p) \to H^*(\mathrm{E}(\mathbb{Z}/p); \mathbb{F}_p)\big) = H^{\geq 1}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p).$$

In order to prove (ii) recall that $\Delta_{2p-1,p}$ is a $(p-1)$-dimensional $(p-2)$-connected free $\mathbb{Z}/p$ simplicial complex, see Theorem 4.6. The Serre spectral sequence associated to the Borel construction fiber bundle

$$\Delta_{2p-1,p} \to \mathrm{E}(\mathbb{Z}/p) \times_{\mathbb{Z}/p} \Delta_{2p-1,p} \to \mathrm{B}(\mathbb{Z}/p)$$

has the E_2-term given by

$$E_2^{i,j} = H^i(\mathrm{B}(\mathbb{Z}/p); \mathcal{H}^j(\Delta_{2p-1,p}; \mathbb{F}_p)) = H^i(\mathbb{Z}/p; H^j(\Delta_{2p-1,p}; \mathbb{F}_p))$$
$$= \begin{cases} H^i(\mathbb{Z}/p; \mathbb{F}_p), & \text{for } j = 0, \\ H^i(\mathbb{Z}/p; H^{p-1}(\Delta_{2p-1,p}; \mathbb{F}_p)), & \text{for } j = p-1, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, $E_\infty^{i,0} \cong E_2^{i,0} \cong H^i(\mathbb{Z}/p; \mathbb{F}_p)$ for $0 \leq i \leq p-1$. Consequently $\mathrm{index}_{\mathbb{Z}/p}\,\Delta_{2p-1,p} \subseteq H^{\geq p}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p)$. Since $\Delta_{2p-1,p}$ is a free $\mathbb{Z}/p$ simplicial complex, we get $\mathrm{E}(\mathbb{Z}/p) \times_{\mathbb{Z}/p} \Delta_{2p-1,p} \simeq \Delta_{2p-1,p}/\mathbb{Z}/p$, implying that $H^i(\mathrm{E}(\mathbb{Z}/p) \times_{\mathbb{Z}/p} \Delta_{2p-1,p}; \mathbb{F}_p) = 0$ for $i \geq p$. Since the spectral sequence $E_*^{*,*}$ converges to the cohomology of the Borel construction $H^*(\mathrm{E}(\mathbb{Z}/p) \times_{\mathbb{Z}/p} \Delta_{2p-1,p}; \mathbb{F}_p)$, we have that $E_\infty^{i,j} = 0$ for $i + j \geq p$. In particular, $E_\infty^{i,0} = 0$ for $i \geq p$, implying that

$$\mathrm{index}_{\mathbb{Z}/p}\,\Delta_{2p-1,p} = H^{\geq p}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p).$$

For (iii) observe that there is a sequence of $\mathbb{Z}/p$-equivariant inclusions

$$\Delta_{1,p} \hookrightarrow \Delta_{2,p} \hookrightarrow \cdots \hookrightarrow \Delta_{k,p} \hookrightarrow \Delta_{k+1,p} \hookrightarrow \cdots$$

given by the inclusions of the corresponding vertex sets

$$[1] \times [p] \hookrightarrow [2] \times [p] \hookrightarrow \cdots \hookrightarrow [k] \times [p] \hookrightarrow [k+1] \times [p] \hookrightarrow \cdots .$$

Consequently the monotonicity property of the Fadell–Husseini index, combined with the fact that for $k \geq 2p-1$ all chessboards $\Delta_{k,p}$ are $(p-1)$-dimensional and $(p-2)$-connected free $\mathbb{Z}/p$ simplicial complexes, implies that

$$\mathrm{index}_{\mathbb{Z}/p}\,\Delta_{1,p} \subseteq \mathrm{index}_{\mathbb{Z}/p}\,\Delta_{2,p} \subseteq \cdots \subseteq \mathrm{index}_{\mathbb{Z}/p}\,\Delta_{2p-1,p}$$
$$= \mathrm{index}_{\mathbb{Z}/p}\,\Delta_{2p,p} = \cdots = H^{\geq p}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p).$$

□

In the next step we compute the index of the chessboard $\Delta_{p-1,p}$. For that we need to establish the following fact.

Lemma 6.5 *Let p be an odd prime. There exists a $\mathbb{Z}/p$-equivariant map*

$$f : \Delta_{p-1,p} \to S(W_p)$$

such that the induced map in cohomology

$$f^* : H^{p-2}(S(W_p); \mathbb{F}_p) \to H^{p-2}(\Delta_{p-1,p}; \mathbb{F}_p)$$

is an isomorphism.

Proof Let $\mathbf{e}_1, \ldots, \mathbf{e}_p$ be a standard basis of $\mathbb{R}^p$, let $\mathbf{e} := \frac{1}{p}(\mathbf{e}_1 + \cdots + \mathbf{e}_p)$, and let $v_i := \mathbf{e}_i - \mathbf{e}$ for $1 \leq i \leq p$. Denote now by $\Delta_{p-1} \subseteq W_p$ the simplex $\operatorname{conv}\{v_1, \ldots, v_p\}$, which is invariant with respect to the action of the cyclic group $\mathbb{Z}/p$. Moreover, its boundary $\partial\Delta_{p-1}$ is equivariantly homeomorphic to the representation sphere $S(W_p)$.

We define a continuous map $f : \Delta_{p-1,p} \to \partial\Delta_{p-1} \cong S(W_p)$ to be the $\mathbb{Z}/p$-equivariant simplicial map given on the vertex set of $\Delta_{p-1,p}$ by $(i,j) \longmapsto v_j$, where $(i,j) \in [p-1] \times [p]$. It remains to be verified that $f^* : H^{p-2}(S(W_p); \mathbb{F}_p) \to H^{p-2}(\Delta_{p-1,p}; \mathbb{F}_p)$ is an isomorphism.

Since $p \geq 3$, the chessboard complex $\Delta_{p-1,p}$ is a connected, orientable pseudomanifold of dimension $p-2$, for this see [30, p. 145]. Thus $H_{p-2}(\Delta_{p-1,p}; \mathbb{Z}) = \mathbb{Z}$ and an orientation class is given by the chain

$$z_{p-1,p} = \sum_{\pi \in \mathfrak{S}_p} (\operatorname{sgn} \pi)\langle (1, \pi(1)), \ldots, (p-1, \pi(p-1))\rangle.$$

Then on the chain level we have that

$$\begin{aligned}
f_\#(z_{p-1,p}) &= \sum_{\pi \in \mathfrak{S}_p} (\operatorname{sgn} \pi)\langle v_{\pi(1)}, \ldots, v_{\pi(p-1)}\rangle \\
&= \sum_{\pi \in \mathfrak{S}_p} (\operatorname{sgn} \pi)\langle v_{\pi(1)}, \ldots, v_{\pi(p-1)}, \widehat{v_{\pi(p)}}\rangle \\
&= \sum_{k=1}^{p} \sum_{\pi \in \mathfrak{S}_p : \pi(p)=k} (-1)^{p+k} (\operatorname{sgn} \pi)^2 \langle v_1, \ldots, \widehat{v_k}, \ldots, v_p\rangle \\
&= \sum_{k=1}^{p} (-1)^{k-1} \sum_{\pi \in \mathfrak{S}_p : \pi(p)=k} \langle v_1, \ldots, \widehat{v_k}, \ldots, v_p\rangle \\
&= \sum_{k=1}^{p} (-1)^{k-1} (p-1)! \langle v_1, \ldots, \widehat{v_k}, \ldots, v_p\rangle \\
&= (p-1)! \sum_{k=1}^{p} (-1)^{k-1} \langle v_1, \ldots, \widehat{v_k}, \ldots, v_p\rangle.
\end{aligned}$$

For this calculation keep in mind that p is an odd prime. The chain $\sum_{k=1}^{p} (-1)^{k-1} \langle v_1, \ldots, \widehat{v_k}, \ldots, v_p\rangle$ is a generator of the top homology of the sphere $\partial\Delta_{p-1} \cong S(W_p)$. Therefore, the induced map in homology

$$f_* : H_{p-2}(\Delta_{p-1,p}; \mathbb{Z}) \to H_{p-2}(S(W_p); \mathbb{Z})$$

is just a multiplication by $(p-1)! \equiv -1 \pmod{p}$. Using the naturality of the universal coefficient isomorphism [19, Cor. 7.5] we have that the induced map in homology with $\mathbb{F}_p$ field coefficients

$$f_* : H_{p-2}(\Delta_{p-1,p}; \mathbb{F}_p) \to H_{p-2}(S(W_p); \mathbb{F}_p)$$

is again multiplication by $(p-1)!$. Since $(p-1)!$ and p are relatively prime the multiplication by $(p-1)!$ is an isomorphism. Now using yet another universal coefficient isomorphism [19, Cor. 7.2] for the coefficients in a field we get that the induced map in cohomology with $\mathbb{F}_p$ coefficients

$$f^* : H^{p-2}(S(W_p);\mathbb{F}_p) \to H^{p-2}(\Delta_{p-1,p};\mathbb{F}_p)$$

is an isomorphism. □

Now we have all ingredients needed to compute the index of the chessboard $\Delta_{p-1,p}$.

Theorem 6.6 $\mathrm{index}_{\mathbb{Z}/p}\,\Delta_{p-1,p} = \mathrm{index}_{\mathbb{Z}/p}\,S(W_p) = H^{\geq p-1}(\mathrm{B}(\mathbb{Z}/p);\mathbb{F}_p)$.

Proof Let us denote by λ the Borel construction fiber bundle

$$\lambda \quad : \quad \Delta_{p-1,p} \to \mathrm{E}(\mathbb{Z}/p)\times_{\mathbb{Z}/p}\Delta_{p-1,p} \to \mathrm{B}(\mathbb{Z}/p),$$

and by ρ the Borel construction fiber bundle

$$\rho \quad : \quad S(W_p) \to \mathrm{E}(\mathbb{Z}/p)^n \times_{\mathbb{Z}/p} S(W_p) \to \mathrm{B}(\mathbb{Z}/p).$$

The $\mathbb{Z}/p$-equivariant map $f : \Delta_{p-1,p} \to S(W_p)$ constructed in Lemma 6.5 induces a morphism of the Borel construction fiber bundles λ and ρ:

$$\begin{tikzcd}
\mathrm{E}(\mathbb{Z}/p)\times_{\mathbb{Z}/p}\Delta_{p-1,p} \arrow[r, "\mathrm{id}\times_{\mathbb{Z}/p}f"] \arrow[d] & \mathrm{E}(\mathbb{Z}/p)\times_{\mathbb{Z}/p}S(W_p) \arrow[d] \\
\mathrm{B}(\mathbb{Z}/p) \arrow[r, "="] & \mathrm{B}(\mathbb{Z}/p).
\end{tikzcd}$$

This morphism induces a morphism of the corresponding Serre spectral sequences

$$E_s^{i,j}(\lambda) := E_s^{i,j}(\mathrm{E}(\mathbb{Z}/p)\times_{\mathbb{Z}/p}\Delta_{p-1,p}) \xleftarrow{f_s^{i,j}} E_s^{i,j}(\mathrm{E}(\mathbb{Z}/p)\times_{\mathbb{Z}/p}S(W_p)) =: E_s^{i,j}(\rho)$$

with the property that on the zero row of the second term the induced map

$$E_2^{i,0}(\lambda) = E_2^{i,0}(\mathrm{E}(\mathbb{Z}/p)\times_{\mathbb{Z}/p}\Delta_{p-1,p}) \xleftarrow{f_2^{i,0}} E_2^{i,0}(\mathrm{E}(\mathbb{Z}/p)\times_{\mathbb{Z}/p}S(W_p)) = E_2^{i,0}(\rho)$$

is the identity. Here we use simplified notation $f_s^{i,j} := E_s^{i,j}(\mathrm{id}\times_{\mathbb{Z}/p}f)$. In the E_2-term, since the homomorphism $f^* : H^{p-2}(S(W_p);\mathbb{F}_p) \to H^{p-2}(\Delta_{p-1,p};\mathbb{F}_p)$ induces an isomorphism on the $(p-2)$-cohomology, and $\mathbb{Z}/p$ acts trivially on both cohomologies $H^{p-2}(S(W_p);\mathbb{F}_p) \cong H^{p-2}(\Delta_{p-1,p};\mathbb{F}_p) \cong \mathbb{F}_p$, the morphism of spectral sequences

$$f_2^{i,p-2} : E_2^{i,p-2}(\rho) \to E_2^{i,p-2}(\lambda) \tag{25}$$

is an isomorphism.

The E_2-term of the Serre spectral sequence associated to the fiber bundle ρ is given by

$$\begin{aligned} E_2^{i,j}(\rho) &= H^i(\mathrm{B}(\mathbb{Z}/p);\mathcal{H}^j(S(W_p);\mathbb{F}_p)) = H^i(\mathbb{Z}/p;H^j(S(W_p);\mathbb{F}_p)) \\ &\cong H^i(\mathbb{Z}/p;\mathbb{F}_p)\otimes_{\mathbb{F}_p} H^j(S(W_p);\mathbb{F}_p), \end{aligned}$$

because $\mathbb{Z}/p$ acts trivially on the cohomology $H^*(S(W_p);\mathbb{F}_p)$. Thus the only possible non-trivial differential is

$$\partial_{p-1} : E_2^{i,p-2}(\rho) \cong E_{p-1}^{i,p-2}(\rho) \to E_2^{i+p-1,0}(\rho) \cong E_{p-1}^{i+p-1,0}(\rho).$$

Let $\ell \in H^{p-2}(S(W_p);\mathbb{F}_p)$ denote a generator. Then the $(p-2)$-row of the E_2-term, as an $H^*(\mathbb{Z}/p;\mathbb{F}_p)$-module, is generated by $1 \otimes_{\mathbb{F}_p} \ell \in E_2^{0,p-2}(\rho)$. Since the differentials are $H^*(\mathbb{Z}/p;\mathbb{F}_p)$-module maps it follows that the differential ∂_{p-1} is completely determined by its image $\partial_{p-1}(1 \otimes_{\mathbb{F}_p} \ell) \in E_{p-1}^{p-1,0}(\rho) \cong E_2^{p-1,0}(\rho)$. In order to find the image of the differential notice that $\mathbb{Z}/p$ acts freely on the sphere $S(W_p)$ and consequently $\mathrm{E}(\mathbb{Z}/p) \times_{\mathbb{Z}/p} S(W_p) \simeq S(W_p)/\mathbb{Z}/p$. Since the spectral sequence $E_s^{i,j}$ converges to the cohomology $H^*(\mathrm{E}(\mathbb{Z}/p) \times_{\mathbb{Z}/p} S(W_p);\mathbb{F}_p)$ we have that $E_\infty^{i,j}(\rho) \cong E_p^{i,j}(\rho) = 0$ for $i + j \geq p - 1$. Thus,

$$\partial_{p-1}(1 \otimes_{\mathbb{F}_p} \ell) = \omega \cdot t^{(p-1)/2} \neq 0$$

for some $\omega \in \mathbb{F}_p\backslash\{0\}$. Moreover,

$$\mathrm{index}_{\mathbb{Z}/p}\, S(W_p) = \left\langle t^{(p-1)/2} \right\rangle = H^{\geq p-1}(\mathrm{B}(\mathbb{Z}/p);\mathbb{F}_p).$$

The E_2-term of the Serre spectral sequence associated to the fiber bundle λ is given by

$$E_2^{i,j}(\lambda) = H^i(\mathrm{B}(\mathbb{Z}/p);\mathcal{H}^j(\Delta_{p-1,p};\mathbb{F}_p)) = H^i(\mathbb{Z}/p;H^j(\Delta_{p-1,p};\mathbb{F}_p)).$$

In particular, $E_2^{i,0}(\lambda) \cong H^i(\mathbb{Z}/p;\mathbb{F}_p)$ and $E_2^{i,p-2}(\lambda) \cong H^i(\mathbb{Z}/p;\mathbb{F}_p)$, because $\mathbb{Z}/p$ acts trivially on the cohomology $H^{p-2}(\Delta_{p-1,p};\mathbb{F}_p) \cong \mathbb{F}_p$. Let $z := f_2^{0,p-2}(1 \otimes_{\mathbb{F}_p} \ell)$. As we have seen in (25) the map $f_2^{0,p-2}$ is an isomorphism. Thus z is a generator of $E_2^{0,p-2}(\lambda) \cong \mathbb{F}_p$, and moreover z is a generator of the $(p-2)$-row of the E_2-term as an $H^*(\mathbb{Z}/p;\mathbb{F}_p)$-module. As in the case of the spectral sequence $E_s^{i,j}(\rho)$ the fact that $\mathbb{Z}/p$-acts freely on the chessboard $\Delta_{p-1,p}$ implies that $E_\infty^{i,j}(\lambda) \cong E_p^{i,j}(\lambda) = 0$ for $i + j \geq p - 1$.

Since $f_s^{i,j}$ is a morphism of spectral sequences it has to commute with the differentials. In particular, for $2 \le s \le p-2$ we have

$$\partial_s(z) = \partial_s(f_s^{0,p-2}(1 \otimes_{\mathbb{F}_p} \ell)) = f_s^{s,p-s-1}(\partial_s(1 \otimes_{\mathbb{F}_p} \ell)) = 0.$$

Now the fact that z is a generator of the $(p-2)$-row of the E_2-term as an $H^*(\mathbb{Z}/p;\mathbb{F}_p)$-module yields

$$E_{p-1}^{i,p-2}(\lambda) \cong E_2^{i,p-2}(\lambda) \cong H^i(\mathbb{Z}/p;\mathbb{F}_p).$$

If in addition $\partial_{p-1}(z) = 0$, then for every $i \ge 0$

$$E_p^{i,p-2}(\lambda) \cong E_{p-1}^{i,p-2}(\lambda) \cong E_2^{i,p-2}(\lambda) \cong H^i(\mathbb{Z}/p;\mathbb{F}_p) \neq 0,$$

which contradicts the fact that $E_\infty^{i,j}(\lambda) \cong E_p^{i,j}(\lambda) = 0$ for $i + j \ge p-1$. In summary we have that

$$\begin{aligned}\partial_{p-1}(z) &= \partial_{p-1}(f_s^{0,p-2}(1 \otimes_{\mathbb{F}_p} \ell)) = f_{p-1}^{p-1,0}(\partial_{p-1}(1 \otimes_{\mathbb{F}_p} \ell)) = f_{p-1}^{p-1,o}(\omega \cdot t^{(p-1)/2}) \\ &= \omega \cdot f_{p-1}^{p-1,0}(t^{(p-1)/2}) \neq 0.\end{aligned}$$

Moreover, we have that

$$\partial_{p-1} : E_{p-1}^{i,p-2}(\lambda) \to E_{p-1}^{i+p-1,0}(\lambda)$$

must be an isomorphism for every $i \ge 0$. Hence, for $i \ge 0$ we have that

$$E_{p-1}^{i+p-1,0}(\lambda) \cong E_2^{i+p-1,0}(\lambda) \cong H^{i+p-1}(\mathbb{Z}/p;\mathbb{F}_p) \cong \mathbb{F}_p. \tag{26}$$

Since, $f_{p-1}^{p-1,0}(t^{(p-1)/2}) \neq 0$ and $f_2^{p-1,0}$ is the identity map we conclude that $f_{p-1}^{p-1,0}(t^{(p-1)/2}) = t^{(p-1)/2}$ and consequently,

$$\text{index}_{\mathbb{Z}/p}\, \Delta_{p-1,p} \subseteq \left\langle t^{(p-1)/2} \right\rangle = H^{\ge p-1}(\mathrm{B}(\mathbb{Z}/p);\mathbb{F}_p).$$

Finally we claim that no non-zero differential can arrive to the 0-row on E_s-term for $2 \le s \le p-2$, implying that

$$\text{index}_{\mathbb{Z}/p}\, \Delta_{p-1,p} = \left\langle t^{(p-1)/2} \right\rangle = H^{\ge p-1}(\mathrm{B}(\mathbb{Z}/p);\mathbb{F}_p),$$

and concluding the proof of the theorem. Indeed, if this is not true, then there exists a minimal s such that $2 \le s \le p-2$ and $0 \neq \partial_s(y) = t^a e^b \in E_s^{i,0}(\lambda)$ for some y and $0 \le i \le p-2$. Since differentials are $H^*(\mathbb{Z}/p;\mathbb{F}_p)$-module maps we have that $\partial_s(t^c \cdot y) = t^c \cdot \partial_s(y) = t^{a+c}e^b \in E_s^{i+2c,0}(\lambda)$ for every $c \ge 0$. Consequently,

$E_{s+1}^{i+2c,0}(\lambda) = 0$ for every $c \geq 0$ contradicting the existence of the isomorphisms (26). Thus, no non-zero differential can arrive to the 0-row before the E_{p-1}-term. □

The proof of the previous theorem, combined together with the fact that a join of pseudomanifolds is a pseudomanifold, yields the following corollary [15, Cor. 2.6].

Corollary 6.7 *Let* $m \geq 1$ *be an integer. Then*

(i) $\operatorname{index}_{\mathbb{Z}/p} \Delta^{*m}_{p-1,p} = \operatorname{index}_{\mathbb{Z}/p} S(W_p^{\oplus m}) = H^{\geq m(p-1)}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p)$,

(ii) $\operatorname{index}_{\mathbb{Z}/p}(\Delta^{*m}_{p-1,p} * [p]) = \operatorname{index}_{\mathbb{Z}/p}(S(W_p^{\oplus m}) * [p]) = H^{\geq m(p-1)+1}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p)$,

(iii) $\operatorname{index}_{\mathbb{Z}/p}(\Delta^{*m}_{p-1,p} * \Delta_{2p-1,p}) = \operatorname{index}_{\mathbb{Z}/p}(S(W_p^{\oplus m}) * [p]^{*p-1}) = H^{\geq m(p-1)+p}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p)$.

In the next step we compute the index of the chessboard $\Delta_{k,p}$ for $1 \leq k \leq p-2$.

Theorem 6.8 $\operatorname{index}_{\mathbb{Z}/p} \Delta_{k,p} = H^{\geq k}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p)$, *for* $1 \leq k \leq p-1$.

Proof Let $1 \leq k \leq p-2$ be an integer. The chessboard $\Delta_{k,p}$ is a $(k-1)$-dimensional free $\mathbb{Z}/p$ simplicial complex. Thus $\mathrm{E}(\mathbb{Z}/p) \times_{\mathbb{Z}/p} \Delta_{k,p} \simeq \Delta_{k,p}/\mathbb{Z}/p$ and consequently $H^i(\mathrm{E}(\mathbb{Z}/p) \times_{\mathbb{Z}/p} \Delta_{k,p}; \mathbb{F}_p) = 0$ for all $i \geq k$. Therefore, $\operatorname{index}_{\mathbb{Z}/p} \Delta_{k,p} \supseteq H^{\geq k}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p)$. For $k = 1$ the theorem follows from Lemma 6.4(i). Furthermore, for $k = p-1$ the statement is the content of Theorem 6.6.

Now let us assume that $2 \leq k \leq p-3$ is even. Then $p-1-k$ is also even. Now consider the $\mathbb{Z}/p$-equivariant inclusion map

$$\Delta_{p-1,p} \to \Delta_{k,p} * \Delta_{p-1-k,p}.$$

From the monotonicity and join properties of the Fadell–Husseini index we have that

$$\operatorname{index}_{\mathbb{Z}/p} \Delta_{k,p} \cdot \operatorname{index}_{\mathbb{Z}/p} \Delta_{p-1-k,p} \subseteq \operatorname{index}_{\mathbb{Z}/p}(\Delta_{k,p} * \Delta_{p-1-k,p}) \subseteq \operatorname{index}_{\mathbb{Z}/p} \Delta_{p-1,p}.$$

Since $p-1-k$ is even and, as we have seen,

$$\operatorname{index}_{\mathbb{Z}/p} \Delta_{p-1-k,p} \supseteq H^{\geq p-1-k}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p) = \langle t^{(p-1-k)/2} \rangle$$

we have that $t^{(p-1-k)/2} \in \operatorname{index}_{\mathbb{Z}/p} \Delta_{p-1-k,p}$. On the other hand, assume that there is an element $u \in \operatorname{index}_{\mathbb{Z}/p} \Delta_{k,p}$ such that $\deg(u) \leq k-1$. Then we have reached a contradiction

$$\begin{aligned} 0 \neq u \cdot t^{(p-1-k)/2} \in \operatorname{index}_{\mathbb{Z}/p} \Delta_{k,p} \cdot \operatorname{index}_{\mathbb{Z}/p} \Delta_{p-1-k,p} &\subseteq \operatorname{index}_{\mathbb{Z}/p} \Delta_{p-1,p} \\ &= H^{\geq p-1}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p), \end{aligned}$$

because $\deg(u \cdot t^{(p-1-k)/2}) = \deg(u) + \deg(t^{(p-1-k)/2}) = \deg(u) + p - 1 - k \leq p-2$. Thus we have proved that for even k

$$\operatorname{index}_{\mathbb{Z}/p} \Delta_{k,p} = H^{\geq k}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p).$$

Next let us assume that $3 \le k \le p-2$ is odd. As we observed at the start of the proof

$$\operatorname{index}_{\mathbb{Z}/p} \Delta_{k,p} \supseteq H^{\ge k}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p) = \langle t^{(k-1)/2}e, t^{(k+1)/2} \rangle.$$

The $\mathbb{Z}/p$-equivariant inclusion map $\Delta_{k-1,p} \subseteq \Delta_{k,p}$ together with the computation of the index for even integers implies that

$$\begin{aligned}\langle t^{(k-1)/2} \rangle &= H^{\ge k-1}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p) \\ &= \operatorname{index}_{\mathbb{Z}/p} \Delta_{k-1,p} \supseteq \operatorname{index}_{\mathbb{Z}/p} \Delta_{k,p} \supseteq H^{\ge k}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p).\end{aligned}$$

In order to conclude the proof of the theorem it remains to prove that $t^{(k-1)/2} \notin \operatorname{index}_{\mathbb{Z}/p} \Delta_{k,p}$. This would yield the equality

$$\operatorname{index}_{\mathbb{Z}/p} \Delta_{k,p} = H^{\ge k}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p)$$

for all odd k. Indeed, assume the opposite, that is, $t^{(k-1)/2} \in \operatorname{index}_{\mathbb{Z}/p} \Delta_{k,p}$. The $\mathbb{Z}/p$-equivariant inclusion $\Delta_{k+1,p} \subseteq \Delta_{1,p} * \Delta_{k,p}$ combined with the monotonicity and join properties of the Fadell–Husseini index imply that

$$\operatorname{index}_{\mathbb{Z}/p} \Delta_{1,p} \cdot \operatorname{index}_{\mathbb{Z}/p} \Delta_{k,p} \subseteq \operatorname{index}_{\mathbb{Z}/p}(\Delta_{1,p} * \Delta_{k,p}) \subseteq \operatorname{index}_{\mathbb{Z}/p} \Delta_{k+1,p}.$$

Since $e \in \operatorname{index}_{\mathbb{Z}/p} \Delta_{1,p}$, and we have assumed that $t^{(k-1)/2} \in \operatorname{index}_{\mathbb{Z}/p} \Delta_{k,p}$, the previous relation implies that

$$t^{(k-1)/2}e \in \operatorname{index}_{\mathbb{Z}/p} \Delta_{k+1,p} = H^{\ge k+1}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p) = \langle t^{(k+1)/2} \rangle,$$

a contradiction. Hence $t^{(k-1)/2} \notin \operatorname{index}_{\mathbb{Z}/p} \Delta_{k,p}$, and the proof of the theorem is complete. □

Let us review the results on the Fadell–Husseini index of chessboards we have obtained so far:

$$\begin{tikzcd}
\operatorname{index}_{\mathbb{Z}/p} \Delta_{1,p} \arrow[d, leftrightarrow, "="'] & \operatorname{index}_{\mathbb{Z}/p} \Delta_{2,p} \arrow[l, "\supseteq"'] \arrow[d, leftrightarrow, "="'] & \dots \arrow[l, "\supseteq"'] & \operatorname{index}_{\mathbb{Z}/p} \Delta_{p-1,p} \arrow[l, "\supseteq"'] \arrow[d, leftrightarrow, "="'] & \dots \arrow[l, "\supseteq"'] & \operatorname{index}_{\mathbb{Z}/p} \Delta_{2p-1,p} \arrow[l, "\supseteq"'] \arrow[d, leftrightarrow, "="'] & {} \arrow[l, "="'] \\
H^{\ge 1}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p) & H^{\ge 2}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p) \arrow[l, "\supseteq"'] & \dots \arrow[l, "\supseteq"'] & H^{\ge p-1}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p) \arrow[l, "\supseteq"'] & \dots \arrow[l, "\supseteq"'] & H^{\ge p}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p) \arrow[l, "\supseteq"'] & {} \arrow[l, "="']
\end{tikzcd}$$

The remaining question indicated by this diagram is: *For which chessboard* $\Delta_{k,p}$ *with* $p-1 \le k \le 2p-1$ *does the first jump in the index* $H^{\ge p-1}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p)$ *to* $H^{\ge p}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p)$ *happen?*

Theorem 6.9 $\operatorname{index}_{\mathbb{Z}/p} \Delta_{k,p} = H^{\ge p-1}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p)$, *for* $p-1 \le k \le 2p-2$.

Proof It suffices to show that $\operatorname{index}_{\mathbb{Z}/p} \Delta_{2p-2,p} = H^{\ge p-1}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p)$. For this we are going to prove that $t^{(p-1)/2} \in \operatorname{index}_{\mathbb{Z}/p} \Delta_{2p-2,p}$.

Consider the following composition of maps

$$\begin{aligned}\Delta_{2p-2,p} \to \Delta_{p-1,p} *_{\Delta(2)} \Delta_{p-1,p} \overset{f*f}{\to} \partial\Delta_{p-1} *_{\Delta(2)} \partial\Delta_{p-1} \to \\ \{\lambda x + (1-\lambda)y \in S^{p-2} * S^{p-2} : \lambda \neq \tfrac{1}{2} \text{ or } x \neq y\} \to S^{p-2} \cong S(W_p),\end{aligned}$$

where the first map is an inclusion, the second map is the two-fold join of the map $f : \Delta_{p-1,p} \to \partial\Delta_{p-1}$, $\Delta_{p-1} \cong S(W_p)$ introduced in Lemma 6.5, the third map is again an inclusion, while the last map is a deformation retraction. All the maps in this composition are $\mathbb{Z}/p$-equivariant. The monotonicity property of the Fadell–Husseini index implies that

$$\operatorname{index}_{\mathbb{Z}/p} \Delta_{2p-2,p} \supseteq \operatorname{index}_{\mathbb{Z}/p} S(W_p) = \langle t^{(p-1)/2} \rangle = H^{\geq p-1}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p),$$

according to (17). Thus $t^{(p-1)/2} \in \operatorname{index}_{\mathbb{Z}/p} \Delta_{2p-2,p}$, and we have concluded the proof of the theorem. □

Now we have the answer to our question. The jump happens in the last possible moment, that is for the index of $\Delta_{2p-1,p}$. The proof of this is due to Carsten Schultz.

We conclude the section with a very useful corollary [15, Cor. 2.6], which also hides a proof for the upcoming optimal colored Tverberg theorem 6.14.

Corollary 6.10 *Let* $1 \leq k_1, \ldots, k_n \leq p-1$. *Then*

$$\operatorname{index}_{\mathbb{Z}/p}(\Delta_{k_1,p} * \cdots * \Delta_{k_n,p}) = H^{\geq k_1+\cdots+k_n}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p).$$

Proof Let $K := \Delta_{k_1,p} * \cdots * \Delta_{k_n,p}$, $K' := \Delta_{p-1-k_1,p} * \cdots * \Delta_{p-1-k_n,p}$, and $L := \Delta^{*n}_{p-1,p}$. Then there is a $\mathbb{Z}/p$-equivariant inclusion $L \to K * K'$. Again the monotonicity and join properties of the Fadell–Husseini index imply that

$$\operatorname{index}_{\mathbb{Z}/p} L \supseteq \operatorname{index}_{\mathbb{Z}/p}(K * K') \supseteq \operatorname{index}_{\mathbb{Z}/p} K \cdot \operatorname{index}_{\mathbb{Z}/p} K'.$$

Furthermore $\dim L = \dim K + \dim K' + 1$. The complexes K and K' are free $\mathbb{Z}/p$-spaces and therefore, as previously observed, it follows that

$$\begin{aligned}&\operatorname{index}_{\mathbb{Z}/p} K \supseteq H^{\geq \dim K+1}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p) \qquad \text{and} \\ &\operatorname{index}_{\mathbb{Z}/p} K' \supseteq H^{\geq \dim K'+1}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p).\end{aligned}$$

Since, by Corollary 6.7, $\operatorname{index}_{\mathbb{Z}/p} L = H^{\geq \dim L+1}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p)$ and $\dim L + 1$ is an even integer, the relation between the indexes

$$\operatorname{index}_{\mathbb{Z}/p} L \supseteq \operatorname{index}_{\mathbb{Z}/p} K \cdot \operatorname{index}_{\mathbb{Z}/p} K'$$

implies that

$$\mathrm{index}_{\mathbb{Z}/p} K = H^{\geq \dim K+1}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p),$$

as claimed. We have also proved that $\mathrm{index}_{\mathbb{Z}/p} K' = H^{\geq \dim K'+1}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p)$. □

6.2 *The Bárány–Larman Conjecture and the Optimal Colored Tverberg Theorem*

Finally we will, motivated by Theorem 6.3, utilize the computation of the Fadell–Husseini index for the chessboards to prove the following result [17, Prop. 4.2].

Theorem 6.11 *Let $d \geq 1$ be an integer, and let p be an odd prime. There is no $\mathfrak{S}_p$-equivariant map*

$$\Delta_{p-1,p}^{*(d+1)} * [p] \to S(W_p^{\oplus(d+1)}).$$

Proof It suffices to prove that there is no $\mathbb{Z}/p$-equivariant map $\Delta_{p-1,p}^{*(d+1)} * [p] \to S(W_p^{\oplus(d+1)})$, where $\mathbb{Z}/p$ is a subgroup of the symmetric group $\mathfrak{S}_p$ generated by the cycle $(12\ldots p)$. The proof uses the monotonicity property of the Fadell–Husseini index.

According to (17) and the join property for the spheres, the index of the sphere $S(W_p^{\oplus(d+1)})$ is

$$\mathrm{index}_{\mathbb{Z}/p} S(W_p^{\oplus(d+1)}) = \langle t^{(d+1)(p-1)/2} \rangle = H^{\geq (d+1)(p-1)}(\mathrm{B}(\mathbb{Z}/p); \mathbb{F}_p).$$

Using Corollary 6.7 we get that

$$\mathrm{index}_{\mathbb{Z}/p}(\Delta_{p-1,p}^{*(d+1)} * [p]) = H^{\geq (d+1)(p-1)+1}(\mathrm{E}(\mathbb{Z}/p); \mathbb{F}_p),$$

and consequently $t^{(d+1)(p-1)/2} \notin \mathrm{index}_{\mathbb{Z}/p}(\Delta_{p-1,p}^{*(d+1)} * [p])$. Thus,

$$\mathrm{index}_{\mathbb{Z}/p} S(W_p^{\oplus(d+1)}) \not\subseteq \mathrm{index}_{\mathbb{Z}/p}(\Delta_{p-1,p}^{*(d+1)} * [p]),$$

implying that a $\mathbb{Z}/p$-equivariant map $\Delta_{p-1,p}^{*(d+1)} * [p] \to S(W_p^{\oplus(d+1)})$ cannot exist. □

A direct corollary of Theorems 6.3 and 6.11 is that the Bárány–Larman conjecture holds for all integers r such that $r + 1$ is a prime [17, Cor. 2.3].

Corollary 6.12 (The Bárány–Larman conjecture for primes−1) *Let $r \geq 2$ and $d \geq 1$ be integers such that $r + 1 =: p$ is a prime. Then $n(d, r) = (d + 1)r$ and $tt(d, r) = r$.*

Using the pigeonhole principle and the index computation for the chessboards we can in addition prove that in the case when p is an odd prime the Bárány–Larman function $n(d, p)$ is finite.

Theorem 6.13 *Let p be an odd prime. Then $n(d, p) \leq (d + 1)(2p - 2) + 1$.*

Proof Let $n = (d + 1)(2p - 2) + 1$, and let $(C_1, \dots, C_{d+1})$ be a coloring of the vertex set of the simplex Δ_{n-1} by $d + 1$ colors with each color class of size at least p. Then by the pigeonhole principle at least one of the colors, let say C_{d+1}, has to be of the size at least $2p - 1$. According to Corollary 4.8: If we can prove that there is no $\mathfrak{S}_p$- or $\mathbb{Z}/p$-equivariant map

$$\Delta_{|C_0|,p} * \cdots * \Delta_{|C_{d+1}|,p} \cong (R_{(C_1,\dots,C_{d+1})})^{*p}_{\Delta(2)} \to S(W_p^{\oplus(d+1)}),$$

then for every continuous map $f : \Delta_{n-1} \to \mathbb{R}^d$ there are p pairwise disjoint rainbow faces $\sigma_1, \dots, \sigma_r$ of Δ_{n-1} whose f-images overlap, that is $f(\sigma_1) \cap \cdots \cap f(\sigma_p) \neq \emptyset$. Thus we will now prove that there is no $\mathbb{Z}/r$-equivariant map

$$\Delta_{|C_0|,p} * \cdots * \Delta_{|C_{d+1}|,p} \to S(W_p^{\oplus(d+1)}).$$

Again, using (17) and the join property for the spheres, we have that

$$\operatorname{index}_{\mathbb{Z}/p} S(W_p^{\oplus(d+1)}) = \langle t^{(d+1)(p-1)/2} \rangle = H^{\geq (d+1)(p-1)}(\mathrm{B}\mathbb{Z}/p; \mathbb{F}_p).$$

Since $|C_0| \geq p, \dots, |C_d| \geq p$ and $|C_{d+1}| \geq 2p - 1$, there is a $\mathbb{Z}/p$-equivariant inclusion

$$\Delta_{p-1,p} * \cdots * \Delta_{p-1,p} * \Delta_{2p-1,p} \to \Delta_{|C_0|,p} * \cdots * \Delta_{|C_d|,p} * \Delta_{|C_{d+1}|,p}.$$

Thus the monotonicity property of the Fadell–Husseini index and Corollary 6.7 (iii) imply that

$$\begin{aligned} H^{\geq (d+1)(p-1)+1}(\mathrm{B}\mathbb{Z}/p; \mathbb{F}_p) &= \operatorname{index}_{\mathbb{Z}/p}(\Delta_{p-1,p} * \cdots * \Delta_{p-1,p} * \Delta_{2p-1,p}) \\ &\supseteq \operatorname{index}_{\mathbb{Z}/p}(\Delta_{|C_0|,p} * \cdots * \Delta_{|C_d|,p} * \Delta_{|C_{d+1}|,p}). \end{aligned}$$

Therefore,

$$\operatorname{index}_{\mathbb{Z}/p} S(W_p^{\oplus(d+1)}) \not\subseteq \operatorname{index}_{\mathbb{Z}/p}(\Delta_{|C_0|,p} * \cdots * \Delta_{|C_d|,p} * \Delta_{|C_{d+1}|,p}),$$

and consequently there is no $\mathbb{Z}/p$-equivariant map $\Delta_{|C_0|,p} * \cdots * \Delta_{|C_{d+1}|,p} \to S(W_p^{\oplus(d+1)})$. This concludes the proof of the theorem. □

While focusing on the Bárány–Larman conjecture and the corresponding function $n(d, r)$, we almost overlooked that the index computations for the chessboards establish a considerable strengthening of the topological Tverberg theorem that is known as the optimal colored Tverberg theorem [17, Thm. 2.1].

Theorem 6.14 (The optimal colored Tverberg theorem) *Let $d \geq 1$ be an integer, let p be a prime, $N \geq (d+1)(p-1)$, and let $f : \Delta_N \to \mathbb{R}^d$ be a continuous map. If the vertices of the simplex Δ_N are colored by m colors, where each color class has cardinality at most $p-1$, then there are p pairwise disjoint rainbow faces $\sigma_1, \dots, \sigma_p$ of Δ_N whose f-images overlap,*

$$f(\sigma_1) \cap \cdots \cap f(\sigma_p) \neq \emptyset.$$

7 Dictionary

7.1 Borel Construction

References [1, 29, 44]. Let G be a finite group and let X be a (left) G-space. The *Borel construction* of X is the space given by $\mathrm{E}G \times_G X := (\mathrm{E}G \times X)/G$, where $\mathrm{E}G$ is a free, contractible right G-space and G acts on the product by $g \cdot (e, x) = (e \cdot g^{-1}, g \cdot x)$. The projection $\mathrm{E}G \times X \to \mathrm{E}G$ induces the following fiber bundle

$$X \to \mathrm{E}G \times_G X \to \mathrm{B}G.$$

This fiber bundle is called the *Borel construction fiber bundle*. The Serre spectral sequence associated to the Borel construction fiber bundle has the E_2-term given by

$$E_2^{r,s} = H^r(\mathrm{B}G; \mathcal{H}^s(X; R)) \cong H^r(G; H^s(X; R)),$$

where the coefficients are local and determined by the action of $\pi_1(\mathrm{B}G) \cong G$ on the cohomology of X. Moreover, each row of the spectral sequence has the structure of an $H^*(\mathrm{B}G; R)$-module, while all differentials are $H^*(\mathrm{B}G; R)$-module morphisms.

The Borel construction and the associated fibration are natural with respect to equivariant maps, that is, any G-equivariant map $f : X \to Y$ between G-spaces induces the following morphism of fiber bundles

$$\begin{tikzcd}
\mathrm{E}G \times_G X \arrow[r, "\mathrm{id} \times_G f"] \arrow[d] & \mathrm{E}G \times_G Y \arrow[d] \\
\mathrm{B}G \arrow[r, "="] & \mathrm{B}G.
\end{tikzcd}$$

This morphism of fiber bundle induces a morphism of associated Serre spectral sequences

$$E_t^{r,s}(f) : E_t^{r,s}(\mathrm{E}G \times_G Y) \to E_t^{r,s}(\mathrm{E}G \times_G X),$$

such that

$$E_2^{r,0}(f) : E_2^{r,0}(\mathrm{E}G \times_G Y) \to E_2^{r,0}(\mathrm{E}G \times_G X)$$

is the identity.

7.2 BG

References [1, 29]. For a finite group G the *classifying space* is the quotient space $\mathrm{B}G = \mathrm{E}G/G$. The projection $\mathrm{E}G \to \mathrm{B}G$ is the universal principal G-bundle, that is, the set of all homotopy classes of maps $[X, \mathrm{B}G]$ is in bijection with the set of all isomorphism classes of principal G bundles over X.

7.3 Borsuk–Ulam Theorem

Reference [34]. Let S^n and S^m be free $\mathbb{Z}/2$-spaces. Then a continuous $\mathbb{Z}/2$-equivariant map $S^m \to S^n$ exists if and only if $m \leq n$.

7.4 Cohomology of a Group (Algebraic Definition)

References [1, 20]. Let G be a finite group, and let M be a (left) G-module. Consider a projective resolution $(P_n, d_n)_{n\geq 0}$ of the trivial (left) G module $\mathbb{Z}$, that is, an exact sequence

$$\cdots \longrightarrow P_{n+1} \xrightarrow{d_n} P_n \xrightarrow{d_{n-1}} \cdots \xrightarrow{d_0} P_0 \xrightarrow{\pi} \mathbb{Z} \longrightarrow 0,$$

where each P_n is a projective (left) G-module. The *group cohomology* of G with coefficients in the module M is the cohomology of the following cochain complex

$$\cdots \xleftarrow{d_n^*} \hom_G(P_n, M) \xleftarrow{d_{n-1}^*} \cdots \xleftarrow{d_0^*} \hom_G(P_0, M) \longleftarrow 0.$$

7.5 Cohomology of Group (Topological Definition)

References [1, 20]. Let G be a finite group, and let M be a (left) G-module. The *group cohomology* of G with coefficients in the module M is the cohomology of $\mathrm{B}G$ with local coefficients in the $\pi_1(\mathrm{B}G) \cong G$-module $\mathcal{M}$, that is

$$H^*(G; M) := H^*(\mathrm{B}G; \mathcal{M}).$$

7.6 Connectedness

References [19, 34]. Let $n \geq -1$ be an integer. A topological space X is *n-connected* if any continuous map $f : S^k \to X$, where $-1 \leq k \leq n$, can be continuously extended to a continuous map $g : B^{k+1} \to X$, that is $g|_{\partial B^{k+1}=S^k} = f$. Here B^{k+1} denotes a $(k+1)$-dimensional closed ball whose boundary is the sphere S^k. A topological space X is (-1)-connected if it is non-empty; it is 0-connected if and only if it is path-connected. If the space X is n-connected and Y is m-connected, then the join $X * Y$ is $(n+m+2)$-connected.

If the space X is n-connected, but not $(n+1)$-connected, we write $\operatorname{conn}(X) = n$. Then

$$\operatorname{conn}(X * Y) \geq \operatorname{conn}(X) + \operatorname{conn}(Y) + 2.$$

7.7 Chessboard Complex

References [30, 34]. The *$m \times n$ chessboard complex* $\Delta_{m,n}$ is the simplicial complex whose vertex set is $[m] \times [n]$, and where the set of vertices $\{(i_0, j_0), \ldots, (i_k, j_k)\}$ spans a k-simplex if and only if $\prod_{0 \leq a < b \leq k}(i_a - i_b)(j_a - j_b) \neq 0$. For example, $\Delta_{2,3} \cong S^1$, $\Delta_{3,4} \cong S^1 \times S^1$. The chessboard complex $\Delta_{m,n}$ is an $(\mathfrak{S}_m \times \mathfrak{S}_n)$-space by

$$(\pi_1, \pi_2) \cdot \{(i_0, j_0), \ldots, (i_k, j_k)\} = \{(\pi_1(i_0), \pi_2(j_0)), \ldots, (\pi_1(i_k), \pi_2(j_k))\},$$

where $(\pi_1, \pi_2) \in \mathfrak{S}_m \times \mathfrak{S}_n$, and $\{(i_0, j_0), \ldots, (i_k, j_k)\}$ is a simplex in $\Delta_{m,n}$. The connectivity of the chessboard complex $\Delta_{m,n}$ is

$$\operatorname{conn}(\Delta_{m,n}) = \min\left\{m, n, \left\lfloor \tfrac{m+n+1}{3} \right\rfloor\right\} - 2.$$

For $n \geq 3$, the chessboard complex $\Delta_{n-1,n}$ is a connected, orientable pseudomanifold of dimension $n-2$. Therefore, $H_{n-2}(\Delta_{n-1,n}; \mathbb{Z}) = \mathbb{Z}$ and an orientation homology class is given by the chain

$$z_{n-1,n} = \sum_{\pi \in \mathfrak{S}_n} (\operatorname{sgn} \pi)\langle (1, \pi(1)), \ldots, (n-1, \pi(n-1))\rangle.$$

The symmetric group $\mathfrak{S}_n \cong 1 \times \mathfrak{S}_n \subseteq \mathfrak{S}_{n-1} \times \mathfrak{S}_n$ acts on $\Delta_{n-1,n}$ by the restriction action. Then $\pi \cdot z_{n,n-1} = (\operatorname{sgn} \pi)\, z_{n-1,n-1}$.

7.8 Deleted Join

Reference [34]. Let K be a simplicial complex, let $n \geq 2$, $k \geq 2$ be integers, and let $[n] := \{1, \dots, n\}$. The *n-fold k-wise deleted join* of the simplicial complex K is the simplicial complex

$$K^{*n}_{\Delta(k)} := \{\lambda_1 x_1 + \cdots + \lambda_n x_n \in \sigma_1 * \cdots * \sigma_n \subset K^{*n} : \\ (\forall I \subset [n])\, \mathrm{card}\, I \geq k \Rightarrow \bigcap_{i \in I} \sigma_i = \emptyset\},$$

where $\sigma_1, \dots, \sigma_n$ are faces of K, including the empty face. The symmetric group $\mathfrak{S}_n$ acts on $K^{*n}_{\Delta(k)}$ by

$$\pi \cdot (\lambda_1 x_1 + \cdots + \lambda_n x_n) := \lambda_{\pi^{-1}(1)} x_{\pi^{-1}(1)} + \cdots + \lambda_{\pi^{-1}(n)} x_{\pi^{-1}(n)},$$

for $\pi \in \mathfrak{S}_n$ and $\lambda_1 x_1 + \cdots + \lambda_n x_n \in K^{*n}_{\Delta(k)}$.

7.9 Deleted Product

Reference [34]. Let K be a simplicial complex, let $n \geq 2$, $k \geq 2$ be integers, and let $[n] := \{1, \dots, n\}$. The *n-fold k-wise deleted product* of the simplicial complex K is the cell complex

$$K^{\times n}_{\Delta(k)} := \{(x_1, \dots, x_n) \in \sigma_1 \times \cdots \times \sigma_n \subset K^{\times n} : \\ (\forall I \subset [n])\, \mathrm{card}\, I \geq k \Rightarrow \bigcap_{i \in I} \sigma_i = \emptyset\},$$

where $\sigma_1, \dots, \sigma_n$ are non-empty faces of K. The symmetric group $\mathfrak{S}_n$ acts on $K^{\times n}_{\Delta(k)}$ by

$$\pi \cdot (x_1, \dots, x_n) := (x_{\pi^{-1}(1)}, \dots, x_{\pi^{-1}(n)}),$$

for $\pi \in \mathfrak{S}_n$ and $(x_1, \dots, x_n) \in K^{\times n}_{\Delta(k)}$.

7.10 Dold's Theorem

Reference [34]. Let G be a non-trivial finite group. For an n-connected G-space X and an at most n-dimensional free G-CW complex Y there is no continuous G-equivariant map $X \to Y$.

7.11 EG

References [1, 29, 44]. For a finite group G any contractible free G-CW complex equipped with the right G cellular action is a model for an EG space. Milnor's model is given by $\mathrm{E}G = \operatorname{colim}_{n\in\mathbb{N}} G^{*n}$ where G stands for a 0-dimensional free G-simplicial complex whose vertices are indexed by the elements of the group G and the action on G is given by the right translation, and G^{*n} is an n-fold join of the 0-dimensional simplicial complex with induced diagonal (right) action.

7.12 Equivariant Cohomology (via the Borel Construction)

References [1, 29, 44]. Let G be a finite group and let X be a (left) G-space. The singular or Čech cohomology of the Borel construction $\mathrm{E}G \times_G X$ of the space X is called the *equivariant cohomology* of X and is denoted by $H_G(X;R)$. Here R denotes a group, or a ring of coefficients.

7.13 Equivariant Cohomology of a relative G-CW complex

Reference [44]. Let G be a finite group, let (X,A) be a relative G-CW complex with a free action on $X\backslash A$, and let $C_*(X,A;\mathbb{Z})$ denote the integral cellular chain complex. The cellular free G-action on every skeleton of $X\backslash A$ induces a free G-action on the chain complex $C_*(X,A;\mathbb{Z})$. Thus $C_*(X,A;\mathbb{Z})$ is a chain complex of free $\mathbb{Z}G$-modules.

For a $\mathbb{Z}G$-module M consider

- the G-equivariant chain complex

$$\mathcal{C}_*^G(X,A;M) = C_*(X,A;\mathbb{Z}) \otimes_{\mathbb{Z}G} M,$$

  and define the *equivariant homology* $\mathcal{H}_*^G(X,A;M)$ of (X,A) with coefficients in M to be the homology of the chain complex $\mathcal{C}_*^G(X,A;M)$;
- the G-equivariant cochain complex

$$\mathcal{C}_G^*(X,A;M) = \mathrm{Hom}_{\mathbb{Z}G}(C_*(X,A;\mathbb{Z}),M),$$

  and define the *equivariant cohomology* $\mathcal{H}_G^*(X,A;M)$ of (X,A) with coefficients in M to be the cohomology of the cochain complex $\mathcal{C}_G^*(X,A;M)$.

7.14 Exact Obstruction Sequence

Reference [44]. Let G be a finite group, let $n \geq 1$ be an integer and let Y be a path-connected n-simple G-space. For every relative G-CW complex (X, A) with a free action of G on the complement $X \backslash A$, there exists the obstruction exact sequence

$$[\mathrm{sk}_{n+1} X, Y]_G \to \mathrm{im}\big([\mathrm{sk}_n X, Y]_G \to [\mathrm{sk}_{n-1} X, Y]_G\big) \xrightarrow{[\mathfrak{o}_G^{n+1}]} \mathcal{H}_G^{n+1}(X, A; \pi_n Y),$$

The sequence is natural in X and Y. This should be understood as follows:

- A G-equivariant map $f : \mathrm{sk}_{n-1} X \to Y$ that can be equivariantly extended to the n-skeleton $f' : \mathrm{sk}_n X \to Y$, that is $f'|_{\mathrm{sk}_{n-1} X} = f$, defines a unique element $[\mathfrak{o}_G^{n+1}(f)]$ living in $\mathcal{H}_G^{n+1}(X, A; \pi_n Y)$, called the *obstruction element* associated to the map f;
- The exactness of the sequence means that the obstruction element $[\mathfrak{o}_G^{n+1}(f)]$ is zero if and only if there is a G-equivariant map $f' : \mathrm{sk}_n X \to Y$ whose restriction is in the G homotopy class of the restriction of f, that is $f'|_{\mathrm{sk}_{n-1} X} \simeq_G f|_{\mathrm{sk}_{n-1} X}$, which extends to the $(n+1)$-skeleton $\mathrm{sk}_{n+1} X$.

The obstruction element $[\mathfrak{o}_G^{n+1}(f)]$ associated with the homotopy class $[f] \in [\mathrm{sk}_n X, Y]_G$ can be introduced on the cochain level as well. Let $h : (D^{n+1}, S^n) \to (\mathrm{sk}_{n+1} X, \mathrm{sk}_n X)$ be an attaching map and $e \in C_{n+1}(X, A; \mathbb{Z})$ the corresponding generator. The *obstruction cochain* $\mathfrak{o}_G^{n+1}(f) \in \mathcal{C}_G^{n+1}(X, A; \pi_n Y)$ of the map f is defined on e by

$$\mathfrak{o}_G^{n+1}(h)(e) = [f \circ h] \in [S^n, Y].$$

The cohomology class of the obstruction cocycle coincides with the obstruction element defined via the exact sequence.

7.15 Fadell–Husseini Index

Reference [22]. Let G be a finite group and R be a commutative ring with unit. For a G-space X and a ring R, the *Fadell–Husseini index* of X is defined to be the kernel ideal of the map in equivariant cohomology induced by the G-equivariant map $p_X : X \to \mathrm{pt}$:

$$\mathrm{index}_G(X; R) = \ker\Big(H^*(\mathrm{B}G; R) \to H^*(\mathrm{E}G \times_G X; R)\Big).$$

Some basic properties of the index are:

- *Monotonicity*: If $X \to Y$ is a G-equivariant map then

$$\mathrm{index}_G(X; R) \supseteq \mathrm{index}_G(Y; R).$$

- *Additivity*: If $(X_1 \cup X_2, X_1, X_2)$ is an excisive triple of G-spaces, then
$$\mathrm{index}_G(X_1; R) \cdot \mathrm{index}_G(X_2; R) \subseteq \mathrm{index}_G(X_1 \cup X_2; R).$$
- *Join:* Let X and Y be G-spaces, then
$$\mathrm{index}_G(X; R) \cdot \mathrm{index}_G(Y; R) \subseteq \mathrm{index}_G(X * Y).$$
- *Generalized Borsuk–Ulam–Bourgin–Yang theorem:* Let $f : X \to Y$ be a G-equivariant map, and let $Z \subseteq Y$ be a closed G-invariant subspace. Then
$$\mathrm{index}_G(f^{-1}(Z); R) \cdot \mathrm{index}_G(Y \backslash Z; R) \subseteq \mathrm{index}_G(X; R).$$
- Let U and V be finite dimensional real G-representations. If $H^*(S(U), R)$ and $H^*(S(V), R)$ are trivial G-modules, $\mathrm{index}_G(S(U); R) = \langle f \rangle$ and $\mathrm{index}_G(S(V); R) = \langle g \rangle$, then
$$\mathrm{index}_G(S(U \oplus V); R) = \langle f \cdot g \rangle \subseteq H^*(\mathrm{B}G; R).$$

7.16 G-Action

Let G be a group and let X be a non-empty set. A *(left) G-action* on X is a function $G \times X \to X$, $(g, x) \longmapsto g \cdot x$ with the property that:

$$g \cdot (h \cdot x) = (gh) \cdot x \qquad \text{and} \qquad 1 \cdot x = x,$$

for every $g, h \in G$ and $x \in X$. A set X with a G-action is called a *G-set*. Let G and X in addition be topological spaces. Then a G-action is *continuous* if the function $G \times X \to X$ is continuous with respect to the product topology on $G \times X$. A topological space equipped with a continuous G-action is called a *G-space*.

7.17 G-equivariant Map

Let X and Y be G-sets (spaces). A (continuous) map $f : X \to Y$ is a *G-equivariant map* if $f(g \cdot x) = g \cdot f(x)$ for all $x \in X$ and all $g \in G$.

7.18 G-CW Complex

References [18, 44]. Let G be a finite group. A CW-complex X is a *G-CW complex* if the group G acts on X by cellular maps and for every $g \in G$ the subspace $\{x \in X : g \cdot x = x\}$ is a CW-subcomplex of X.

Let X be a G-CW complex, and let A be a subcomplex of X that is invariant with respect to the action of the group G and consequently a G-CW complex in its own right. The pair of G-CW complex (X, A) is a *relative G-CW complex*.

7.19 Localization Theorem

References [29, 44]. The following result is a consequence of the localization theorem for elementary abelian groups: Let p be a prime, $G = (\mathbb{Z}/p)^n$ for $n \geq 1$, and let X be a finite G-CW complex. The fixed points set X^G of the space X is non-empty if and only the map in cohomology $H^*(\mathrm{B}G; \mathbb{F}_p) \to H^*(\mathrm{E}G \times_G X; \mathbb{F}_p)$, induced by the projection $\mathrm{E}G \times_G X \to \mathrm{B}G$, is a monomorphism.

7.20 n-Simple

Reference [19]. A topological space X is *n-simple* if the fundamental group $\pi_1(X, x_0)$ acts trivially on the n-th homotopy group $\pi_n(X, x_0)$ for every $x_0 \in X$.

7.21 Nerve of a Family of Subsets

Let X be a set and let $\mathcal{X} := \{X_i : i \in I\}$ be a family of subsets of X. The *nerve* of the family $\mathcal{X}$ is the simplical complex $N_{\mathcal{X}}$ with the vertex set I, and a finite subset $\sigma \subseteq I$ is a face of the complex if and only if $\bigcap\{X_i : i \in \sigma\} \neq \emptyset$.

7.22 Nerve Theorem

Reference [10]. Let K be a finite simplicial complex, or a regular CW-complex, and let $\mathcal{K} := \{K_i : i \in I\}$ be a cover of K by a family of subcomplexes, that is $K = \bigcup\{K_i : i \in I\}$.

(1) If for every face σ of the nerve $N_{\mathcal{K}}$ the intersection $\bigcap\{K_i : i \in \sigma\}$ is contractible, then K and $N_{\mathcal{K}}$ are homotopy equivalent, that is $K \simeq N_{\mathcal{K}}$.

(2) If for every face σ of the nerve $N_{\mathcal{K}}$ the intersection $\bigcap\{K_i : i \in \sigma\}$ is $(k-|\sigma|+1)$-connected, then the complex K is k-connected if and only if the nerve $N_{\mathcal{K}}$ is k-connected.

7.23 Primary Obstruction

References [18, 44]. Let G be a finite group, let $n \geq 1$ be an integer and let Y be an $(n-1)$-connected and n-simple G-space. Furthermore, let (X, A) be a relative G-CW complex with the free G action on $X \backslash A$, and let $f : A \to Y$ be a G-equivariant map. Then

- there exists a G-equivariant map $f' : \mathrm{sk}_n X \to Y$ extending f, that is $f'|_A = f$,
- every two G-equivariant extensions $f', f'' : \mathrm{sk}_n X \to Y$ of f are G-homotopic, relative to A, on $\mathrm{sk}_{n-1} X$, that is

$$\mathrm{im}\big([\mathrm{sk}_n X, Y]_G \to [\mathrm{sk}_{n-1} X, Y]_G\big) = \{\mathrm{pt}\},$$

- if $H : A \times I \to Y$ is a G-equivariant homotopy between G-equivariant maps $f : A \to Y$ and $f' : A \to Y$, and if $h : \mathrm{sk}_n X \to Y$ and $h' : \mathrm{sk}_n X \to Y$ are G-equivariant extensions of f and f', then there exists a G-equivariant homotopy $K : \mathrm{sk}_{n-1} X \times I \to Y$ between $h|_{\mathrm{sk}_{n-1} X}$ and $h'|_{\mathrm{sk}_{n-1} X}$ that extends H.

In the case when $\mathrm{im}\big([\mathrm{sk}_n X, Y]_G \to [\mathrm{sk}_{n-1} X, Y]_G\big) = \{\mathrm{pt}\}$ the obstruction sequence becomes

$$[\mathrm{sk}_{n+1} X, Y]_G \to \{\mathrm{pt}\} \overset{[\mathfrak{o}_G^{n+1}]}{\to} \mathcal{H}_G^{n+1}(X, \pi_n Y).$$

The obstruction element $[\mathfrak{o}_G^{n+1}(\mathrm{pt})] \in \mathcal{H}_G^{n+1}(X, \pi_n Y)$ is called the *primary obstruction* and does not depend on the choice of a G-equivariant map on the n-th skeleton of X.

7.24 Restriction and Transfer

References [14, 20]. Let G be a finite group and let $H \subseteq G$ be its subgroup. Consider a $\mathbb{Z}G$-chain complex $C_* = (C_n, c_n)$ and a $\mathbb{Z}G$-module M. Denote by res the restriction from G to H. For every integer n there exists a homomorphism

$$\mathrm{res} : H^n\big(\hom_{\mathbb{Z}G}(C_*, M)\big) \to H^n\big(\hom_{\mathbb{Z}H}(\mathrm{res}\, C_*, \mathrm{res}\, M)\big)$$

that we call the *restriction* from G to H, and a homomorphism

$$\mathrm{tr} : H^n\big(\hom_{\mathbb{Z}H}(\mathrm{res}\, C_*, \mathrm{res}\, M)\big) \to H^n\big(\hom_{\mathbb{Z}G}(C_*, M)\big)$$

that is called the *transfer* from H to G, with the property

$$\mathrm{tr} \circ \mathrm{res} = [G : H] \cdot \mathrm{id}\,.$$

Acknowledgements We are grateful to Alexander Engström and Florian Frick for excellent observations on drafts of this paper and many useful comments. We want to express our gratitude to Peter Landweber for his continuous help and support in improving this manuscript.

References

1. A. Adem, R.J. Milgram, *Cohomology of Finite Groups*, 2nd edn. Grundlehren der Mathematischen Wissenschaften, vol. 309 (Springer, Berlin, 2004)
2. V.I. Arnold, *Experimental Mathematics*. MSRI Mathematical Circles Library, vol. 16 (MSRI, Berkeley/American Mathematical Society, Providence, 2015)
3. S. Avvakumov, I. Mabillard, A. Skopenkov, U. Wagner, *Eliminating higher-multiplicity intersections, III. Codimension 2*, Preprint, 16 pages, arXiv:1511.03501. Nov 2015
4. E.G. Bajmóczy, I. Bárány, On a common generalization of Borsuk's and Radon's theorem. Acta Math. Hungar. **34**, 347–350 (1979)
5. I. Bárány, P.V.M. Blagojević, G.M. Ziegler, Tverberg's theorem at 50: extensions and counterexamples. Not. Am. Math. Soc. **73**(7), 732–739 (2016)
6. I. Bárány, Z. Füredi, L. Lovász, On the number of halving planes. Combinatorica **10**, 175–183 (1990)
7. I. Bárány, D.G. Larman, A colored version of Tverberg's theorem. J. Lond. Math. Soc. **2**, 314–320 (1992)
8. I. Bárány, S.B. Shlosman, A. Szűcs, On a topological generalization of a theorem of Tverberg. J. Lond. Math. Soc. **23**, 158–164 (1981)
9. B.J. Birch, On 3*N* points in a plane. Math. Proc. Camb. Philos. Soc. **55**, 289–293 (1959)
10. A. Björner, *Topological Methods*. Handbook of Combinatorics, vol. 2 (Elsevier, Amsterdam, 1995), pp. 1819–1872
11. A. Björner, L. Lovász, S. Vrećica, R. Živaljević, Chessboard complexes and matching complexes. J. Lond. Math. Soc. **49**, 25–39 (1994)
12. P.V.M. Blagojević, F. Frick, G.M. Ziegler, Tverberg plus constraints. Bull. Lond. Math. Soc. **46**, 953–967 (2014)
13. P.V.M. Blagojević, F. Frick, G.M. Ziegler, Barycenters of polytope skeleta and counterexamples to the topological Tverberg conjecture, via constraints. J. Eur. Math. Soc. (JEMS) (2015, to appear). Preprint, 6 pages, arXiv:1510.07984
14. P.V.M. Blagojević, W. Lück, G.M. Ziegler, Equivariant topology of configuration spaces. J. Topol. **8**, 414–456 (2015)
15. P.V.M. Blagojević, B. Matschke, G.M. Ziegler, Optimal bounds for a colorful Tverberg–Vrećica type problem. Adv. Math. **226**, 5198–5215 (2011)
16. P.V.M. Blagojević, B. Matschke, G.M. Ziegler, A tight colored Tverberg theorem for maps to manifolds. Topol. Appl. **158**, 1445–1452 (2011)
17. P.V.M. Blagojević, B. Matschke, G.M. Ziegler, Optimal bounds for the colored Tverberg problem. J. Eur. Math. Soc. (JEMS) **17**, 739–754 (2015)
18. G.E. Bredon, *Equivariant Cohomology Theories*. Lecture Notes in Mathematics, vol. 34 (Springer, Berlin/New York, 1967)
19. G.E. Bredon, *Topology and Geometry*. Graduate Texts in Mathematics, vol. 139 (Springer, New York, 1993)
20. K.S. Brown, *Cohomology of Groups*. Graduate Texts in Mathematics, vol. 87 (Springer, New York, 1994)
21. A. Dold, Simple proofs of some Borsuk–Ulam results, in *Proceedings of the Northwestern Homotopy Theory Conference*, ed. by H.R. Miller, S.B. Priddy. Contemporary Mathematics, vol. 19 (1983), pp. 65–69
22. E. Fadell, S. Husseini, An ideal-valued cohomological index theory with applications to Borsuk–Ulam and Bourgin–Yang theorems. Ergod. Theory Dynam. Syst. **8**, 73–85 (1988)

23. A. Flores, Über n-dimensionale Komplexe, die im R_{2n+1} absolut selbstverschlungen sind, Ergebnisse eines Math. Kolloquiums **6**, 4–7 (1932/1934)
24. A. Fomenko, D. Fuchs, *Homotopical Topology*, 2nd edn. Graduate Texts in Mathematics, vol. 273 (Springer, Cham, 2016)
25. F. Frick, Counterexamples to the topological Tverberg conjecture. Oberwolfach Rep. **12**, 318–322 (2015)
26. M. Gromov, *Singularities, expanders and topology of maps. II: from combinatorics to topology via algebraic isoperimetry*. Geom. Funct. Anal. (GAFA) **20**, 416–526 (2010)
27. P.M. Gruber, R. Schneider, Problems in geometric convexity, in *Contributions to Geometry (Proceedings of the Geometry Symposium, Siegen, 1978)*, ed. by J. Tölke, J. Wills (Birkhäuser, Basel/Boston, 1979), pp. 255–278
28. B. Grünbaum, *Convex Polytopes*. Graduate Texts in Mathematics, vol. 221 (Springer, New York, 2003). Second edition prepared by V. Kaibel, V. Klee, G.M. Ziegler (Original edition: Interscience, London, 1967)
29. W.Y. Hsiang, *Cohomology Theory of Topological Transformation Groups* (Springer, New York/Heidelberg, 1975). Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 85
30. J. Jonsson, *Simplicial Complexes of Graphs*. Lecture Notes in Mathematics, vol. 1928 (Springer, Berlin, 2008)
31. I. Mabillard, U. Wagner, Eliminating Tverberg points, I. An analogue of the Whitney trick, in *Proceedings of 30th Annual Symposium on Computational Geometry (SoCG)*, Kyoto, June 2014 (ACM, 2014), pp. 171–180
32. I. Mabillard, U. Wagner, Eliminating higher-multiplicity intersections, I. A Whitney trick for Tverberg-type problems, Preprint, 46 pages, Aug 2015, arXiv:1508.02349
33. B.M. Mann, R.J. Milgram, On the Chern classes of the regular representations of some finite groups. Proc. Edinb. Math. Soc. (2) **25**, 259–268 (1982)
34. J. Matoušek, *Using the Borsuk–Ulam Theorem*. Lectures on Topological Methods in Combinatorics and Geometry (Universitext, Springer, Heidelberg, 2003). Second corrected printing 2008
35. J. McCleary, *A User's Guide to Spectral Sequences*, 2nd edn. Cambridge Studies in Advanced Mathematics, vol. 58 (Cambridge University Press, Cambridge, 2001)
36. M. Özaydin, Equivariant maps for the symmetric group, Preprint, 17 pages (1987), http://digital.library.wisc.edu/1793/63829
37. J. Radon, Mengen konvexer Körper, die einen gemeinsamen Punkt enthalten. Math. Ann. **83**, 113–115 (1921)
38. K.S. Sarkaria, A generalized van Kampen–Flores theorem. Proc. Am. Math. Soc. **11**, 559–565 (1991)
39. J. Shareshian, M. Wachs, Torsion in the matching complex and chessboard complex. Adv. Math. **212**, 525–570 (2007)
40. A. Skopenkov, *A user's guide to disproof of topological Tverberg conjecture*, Preprint 2016, arXiv:1605.05141
41. S. Smale, *A Vietoris mapping theorem for homotopy*. Proc. Am. Math. Soc. **8**, 604–610 (1957)
42. P. Soberón, Equal coefficients and tolerance in coloured Tverberg partitions, in *Proceedings of 29th Annual Symposium on Computational Geometry (SoCG)*, Rio de Janeiro, June 2013 (ACM, 2013), pp. 91–96
43. P. Soberón, Equal coefficients and tolerance in coloured Tverberg partitions. Combinatorica **35**, 235–252 (2015)
44. T. tom Dieck, *Transformation Groups*. Studies in Mathematics, vol. 8 (Walter de Gruyter, Berlin, 1987)
45. H. Tverberg, A generalization of Radon's theorem. J. Lond. Math. Soc. **41**, 123–128 (1966)
46. E.R. Van Kampen, Komplexe in euklidischen Räumen. Abh. Math. Semin. Univ. Hamburg **9**, 72–78 (1933)
47. A.Yu. Volovikov, On a topological generalization of Tverberg's theorem. Math. Notes **59**(3), 454–456 (1996)
48. A.Yu. Volovikov, On the van Kampen-Flores theorem. Math. Notes **59**(5), 477–481 (1996)

49. S. Vrećica, R.T. Živaljević, New cases of the colored Tverberg theorem, in *Jerusalem Combinatorics'93*, Jerusalem, ed. by H. Barcelo, G. Kalai. Contemporary Mathematics, vol. 178 (American Mathematical Society, 1994), pp. 325–325
50. H. Whitney, The self-intersections of a smooth *n*-manifold in 2*n*-space. Ann. Math. **45**, 220–246 (1944)
51. G.M. Ziegler, 3*N* colored points in a plane. Not. Am. Math. Soc. **58**(4), 550–557 (2011)
52. R.T. Živaljević, User's guide to equivariant methods in combinatorics II. Publications de l'Institut Mathématique **64**(78), 107–132 (1998)
53. R.T. Živaljević, S. Vrećica, The colored Tverberg's problem and complexes of injective functions. J. Combin. Theory Ser. A **61**, 309–318 (1992)

One-Sided Epsilon-Approximants

Boris Bukh and Gabriel Nivasch

In memory of a great teacher,
Jirka Matoušek

Abstract Given a finite point set $P \subset \mathbb{R}^d$, we call a multiset A a *one-sided weak ε-approximant* for P (with respect to convex sets), if $|P \cap C|/|P| - |A \cap C|/|A| \leq \varepsilon$ for every convex set C.

We show that, in contrast with the usual (two-sided) weak ε-approximants, for every set $P \subset \mathbb{R}^d$ there exists a one-sided weak ε-approximant of size bounded by a function of ε and d.

1 Introduction

A common theme in mathematics is approximation of large, complicated objects by smaller, simpler objects. This paper proposes a new notion of approximation in combinatorial geometry, which we call one-sided ε-approximants. It is a notion of approximation that is in strength between ε-approximants and ε-nets. We recall these two notions first.

Let $P \subset \mathbb{R}^d$ be a finite set, and $\mathcal{F} \subset 2^{\mathbb{R}^d}$ a family of sets in $\mathbb{R}^d$. In applications, the family $\mathcal{F}$ is usually a geometrically natural family, such as the family of all halfspaces, the family of all simplices, or the family of all convex sets. A finite set

Research was supported in part by Churchill College, Cambridge, by U.S. taxpayers through NSF grant DMS-1301548, and by Alfred P. Sloan Foundation through Sloan Research Fellowship.

B. Bukh (✉)
Department of Mathematical Sciences, Carnegie Mellon University, 15213 Pittsburgh, PA, USA
e-mail: bbukh@math.cmu.edu

G. Nivasch
Department of Computer Science, Ariel University, Ariel, Israel
e-mail: gabrieln@ariel.ac.il

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_12

$A \subset \mathbb{R}^d$ is called an *ε-approximant for P with respect to $\mathcal{F}$* if

$$\left| \frac{|C \cap P|}{|P|} - \frac{|C \cap A|}{|A|} \right| \leq \varepsilon \qquad \text{for all } C \in \mathcal{F}.$$

The notion of an ε-approximant was introduced by Vapnik and Chervonenkis [23] in the context of statistical learning theory. They associated to each family $\mathcal{F}$ a number VC-dim$(\mathcal{F}) \in \{1, 2, 3, \dots, \infty\}$, which has become known as *VC dimension*, and proved that if VC-dim$(\mathcal{F}) < \infty$, then every set P admits an ε-approximant A of size $|A| \leq C_{\text{VC-dim}(\mathcal{F})}\varepsilon^{-2}$, a bound which does not depend on the size of P. The ε-approximants that they constructed had the additional property that $A \subset P$. Following tradition, we say that A is a *strong ε-approximant* if $A \subset P$. When we wish to emphasize that our ε-approximants are not necessarily subsets of P, we call them *weak ε-approximants*. The bound has been improved to $|A| \leq C_{\text{VC-dim}(\mathcal{F})}\varepsilon^{-2+2/(\text{VC-dim}(\mathcal{F})+1)}$ (see [18, Theorem 1.2] and [19, Exercise 5.2.7]) which is optimal [1].

In a geometric context, Haussler and Welzl [17] introduced ε-nets. With P and $\mathcal{F}$ as above, a set N is called an *ε-net for P with respect to $\mathcal{F}$* if

$$\frac{|C \cap P|}{|P|} > \varepsilon \implies C \cap N \neq \emptyset \qquad \text{for all } C \in \mathcal{F}.$$

An ε-approximant is an ε-net, but not conversely. While an ε-net is a weaker notion of approximation, its advantage over an ε-approximant is that every set P admits an ε-net of size only $C_{\text{VC-dim}(\mathcal{F})}\varepsilon^{-1}\log\varepsilon^{-1}$, which is smaller than the bound for the ε-approximants. The ε-nets constructed by Haussler and Welzl are also strong, i.e., they satisfy $N \subset P$.

Most geometrically important families $\mathcal{F}$ have a bounded VC dimension. A notable exception is the family $\mathcal{F}_{\text{conv}}$ of all convex sets. Indeed, it is easy to see that a set of n points in convex position does not admit any strong ε-net of size smaller than $(1 - \varepsilon)n$ with respect to $\mathcal{F}_{\text{conv}}$. Alon, Bárány Füredi, and Kleitman [3] showed that for every $P \subset \mathbb{R}^d$ there exists a (weak) ε-net of size bounded solely by a function of ε and d. No extension of their result to ε-approximants is possible.

Proposition 1 *If $P \subset \mathbb{R}^2$ is a set of n points in convex position, then every ε-approximant with respect to $\mathcal{F}_{\text{conv}}$ has size at least $n(\frac{1}{4} - \varepsilon/2)$.*

Proof Let $p_1, p_2, \dots, p_n$ be the enumeration of the vertices of P in clockwise order along the convex hull of P. For $i = 1, \dots, \lfloor (n-1)/2 \rfloor$ write T_i for the triangle $p_{2i-1}, p_{2i}, p_{2i+1}$. Suppose $A \subset \mathbb{R}^2$ is an ε-approximant for P. Let $I \overset{\text{def}}{=} \{i : T_i \cap A = \emptyset\}$. Note that $|I| \geq n/2 - 2|A| - 1$ since each point of A lies in at most two triangles. Define $S \overset{\text{def}}{=} \{p_1, p_3, p_5, \dots\}$ to be the odd-numbered points, and let $S' \overset{\text{def}}{=} S \cup \{p_{2i} : i \in I\}$. Let $C \overset{\text{def}}{=} \operatorname{conv} S$ and $C' \overset{\text{def}}{=} \operatorname{conv} S'$. Then $C \cap A = C' \cap A$, but $|C' \cap P|/|P| - |C \cap P|/|P| = |I|/|P| > \varepsilon$ if $|A| < |P|(\frac{1}{4} - \varepsilon/2)$. □

In light of Proposition 1, we introduce a new concept. A multiset[1] $A \subset \mathbb{R}^d$ is a *one-sided ε-approximant for P with respect to the family $\mathcal{F}$* if

$$\frac{|C \cap P|}{|P|} - \frac{|C \cap A|}{|A|} \leq \varepsilon \qquad \text{for all } C \in \mathcal{F}.$$

In other words, if $C \in \mathcal{F}$, then C might contain many more points of A than expected, but never much fewer. It is clear that an ε-approximant is a one-sided ε-approximant, and that a one-sided ε-approximant is an ε-net.

Our main result shows that allowing one-sided errors is enough to sidestep the pessimistic Proposition 1.

Theorem 2 *Let $P \subset \mathbb{R}^d$ be a finite set, and let $\varepsilon \in (0, 1]$ be a real number. Then P admits a one-sided ε-approximant with respect to $\mathcal{F}_{\text{conv}}$ of size at most $g(\varepsilon, d)$, for some g that depends only on ε and on d.*

Unfortunately, due to the use of a geometric Ramsey theorem, our bound on g is very weak:

$$g(\varepsilon, d) \leq \text{tw}_d \left(\varepsilon^{-c}\right)$$

for some constant $c > 1$ that depends only on d, where the tower function is given by $\text{tw}_1(x) \overset{\text{def}}{=} x$ and $\text{tw}_{i+1}(x) \overset{\text{def}}{=} 2^{\text{tw}_i(x)}$. We believe this bound to be very far from sharp.

In the rest of the paper we omit the words "with respect to $\mathcal{F}_{\text{conv}}$" when referring to one-sided approximants.

2 Outline of the Construction and of the Paper

At a high level, the proof of Theorem 2 can be broken into three steps:

1. We replace the given set P by a bounded-size set $\hat{P}$. The price of this replacement is an extra condition that a one-sided ε-approximant A for $\hat{P}$ would need to satisfy to be a one-sided ε-approximant for P. Namely, A must be a one-sided ε-approximant for a *semialgebraic reason*.
2. We break $\hat{P}$ into long *orientation-homogeneous* subsequences $S_1, S_2, \ldots, S_m$.
3. For each S_i we give an explicit one-sided ε-approximant A_i satisfying the semialgebraicity condition. The union of $A_1, \ldots, A_m$ is then the desired ε-approximant.

Step 1 relies on the semialgebraic regularity lemma from [16], which we recall in Sect. 3.3. Given a fixed set Φ of semialgebraic predicates, this lemma permits us

[1] In this paper we allow A to be a multiset. While the results of this paper continue to hold if we require A to be a set, the proofs become more technical. We sketch the necessary changes in the final section.

to replace P with a constant-sized $\hat{P}$ that behaves similarly to P with respect to the predicates in Φ. Since in steps 2 and 3 we employ only one predicate, in our case we have $|\Phi| = 1$. We define that predicate in Sect. 3.

In step 1 we lose some control on the interaction between parts of $\hat{P}$. To remedy this, we use a well-known hypergraph Turán theorem, discussed in Sect. 4, to extract well-behaved chunks from $\hat{P}$.

The construction of A_i in step 3 consists of Tverberg points of a certain family $\mathcal{F}$ of subsets of S_i. The property that $\mathcal{F}$ needs to satisfy is most naturally described in terms of interval chains, which are introduced in Sect. 5. The actual construction of requisite interval chains is based on the idea behind the regularity lemma for words from [5, 14]. To obtain a better quantitative bound, we eschew using the lemma directly and provide an alternative argument. This is also done in Sect. 5.

All the ingredients are put together in Sect. 6.

The paper concludes with several remarks and open problems.

3 Geometric Preliminaries

The *convex hull* of a point set P is denoted $\operatorname{conv} P$, and its *affine hull* is denoted $\operatorname{ahull} P$.

Tverberg's theorem (see, e.g., [20, p. 200]) asserts that any set $Q \subset \mathbb{R}^d$ of $(s-1)(d+1)+1$ points can be partitioned into s pairwise disjoint subsets whose convex hulls intersect. We denote by $\operatorname{Tver}_s(Q)$ an arbitrary point in such an intersection. A special case of Tverberg's theorem is the case $s = 2$, which is due to Radon [21]. In that case, if Q is in general position (no $d+1$ points are affinely dependent), then the partition is unique and $\operatorname{Tver}_2(Q)$ is also unique.

A *(geometric) predicate* of arity k is a property that a k-tuple of points $p_1, \ldots, p_k$ might or might not satisfy. A predicate is *semialgebraic* if it is a Boolean combination of expressions of the form $f(p_1, \ldots, p_k) \geq 0$, where the f's are polynomials. Predicates that depend on the sign of a single polynomial are especially useful, we call then *polynomial predicates*. For brevity we will identify polynomial predicates with the underlying polynomials.

An important polynomial predicate is the *orientation* of a $(d+1)$-tuple of points in $\mathbb{R}^d$. The orientation of $p_0, \ldots, p_d \in \mathbb{R}^d$ is given by

$$\operatorname{orient}(p_0, \ldots, p_d) \overset{\text{def}}{=} \operatorname{sgn} \det \begin{bmatrix} p_0 & \cdots & p_d \\ 1 & \cdots & 1 \end{bmatrix}.$$

We have $\operatorname{orient}(p_0, \ldots, p_d) = 0$ if and only if the points are affinely dependent.

3.1 Orientation-Homogeneous Sequences

We will call a sequence of points in $\mathbb{R}^d$ *orientation-homogeneous* if all its $(d+1)$-tuples have the same nonzero orientation. It is well known that every orientation-homogeneous sequence is in convex position, and that the convex hull of such a sequence is combinatorially equivalent to a cyclic polytope (see, e.g., [24] for background).

Let $P = (p_1, \dots, p_n)$ be an orientation-homogeneous sequence. For a set $I = \{i_1 < \cdots < i_m\}$, define the subsequence of P indexed by I by $P_I \overset{\text{def}}{=} (p_{i_1}, \dots, p_{i_m})$. If $|I| = d$, then the d points P_I span a hyperplane $H_I \overset{\text{def}}{=} \operatorname{ahull} P_I$ in $\mathbb{R}^d$. It is simple to tell to which side of H_I a point $p_j \in P \setminus P_I$ belongs: The index set I partitions $[n] \setminus I$ into $d+1$ intervals (some of which might be empty). The side of H_I to which p_j belongs depends only on the parity of the interval number to which p_j belongs. In other words, p_j is on one side if $j \in (-\infty, i_1) \cup (i_2, i_3) \cup \cdots$, and on the other side if $j \in (i_1, i_2) \cup (i_3, i_4) \cup \cdots$. Hence, two points p_j and $p_{j'}$ with $j < j'$ lie on the same side of H_I if and only if $[j, j'] \cap I$ is of even size.

Of particular interest to us are sets I of size $d+2$. We define, for such a set $I = \{i_1 < i_2 < \cdots < i_{d+2}\}$, a partition $I = I_{\text{odd}} \cup I_{\text{even}}$, where $I_{\text{odd}} \overset{\text{def}}{=} \{i_1, i_3, \dots\}$ and $I_{\text{even}} \overset{\text{def}}{=} \{i_2, i_4, \dots\}$.

Lemma 3 *If P is orientation-homogeneous and $|I| = d+2$, then the convex sets* $\operatorname{conv} P_{I_{\text{odd}}}$ *and* $\operatorname{conv} P_{I_{\text{even}}}$ *intersect.*

Proof Indeed, suppose they are disjoint, and hence there exists a hyperplane H that separates $P_{I_{\text{odd}}}$ from $P_{I_{\text{even}}}$. Then H can be perturbed into a hyperplane H' that goes through some d points of P_I, i.e., $H' = H_J$ for some $J \subset I$, $|J| = d$. The set $P_{I \setminus J}$ consists of two points, say $p_i, p_{i'}$ with $i < i'$, and they belong to the same part of the partition $P = P_{I_{\text{odd}}} \cup P_{I_{\text{even}}}$ precisely when $[i, i'] \cap J$ is of odd size. This is in contradiction with the criterion for H_J to separate p_i from $p_{i'}$. □

By Ramsey's theorem, there is a number $\mathrm{OT}_d(n)$ such that each sequence of $\mathrm{OT}_d(n)$ points in general position contains an orientation-homogeneous subsequence of length n. The growth rate of $\mathrm{OT}_d(n)$ is known quite precisely: For all $d \geq 2$ we have $\mathrm{tw}_d(c'_d n) \leq \mathrm{OT}_d(n) \leq \mathrm{tw}_d(c_d n)$ for positive constants $c'_d < c_d$. The upper bound is due to Suk [22], and the lower bound is due to Bárány, Matoušek and Pór [6], which is based on an earlier work by Eliáš, Matoušek, Roldán-Pensado and Safernová [13].

3.2 Point Selection

The following lemma is a minor variation on Lemma 2.2 from [4]:

Lemma 4 *Let $s \overset{\text{def}}{=} \lfloor d/2 \rfloor + 1$, and let $D \overset{\text{def}}{=} (s-1)(d+1)+1$. Let $(p_1, p_2, \dots, p_{2D+1})$ be an orientation-homogeneous sequence of $2D+1$ points in $\mathbb{R}^d$. Let $Q =$*

$\{p_2, p_4, \dots, p_{2D}\}$ and $R = \{p_1, p_3, \dots, p_{2D+1}\}$ (so $|Q| = D$ and $|R| = D + 1$). Then $\mathrm{Tver}_s(Q) \in \operatorname{conv} R$.

Proof Let $x \overset{\text{def}}{=} \mathrm{Tver}_s(Q)$. If $x \notin \operatorname{conv} R$, then there exists a hyperplane H separating x from R. There must be at least s points of Q on the same side of H as x (at least one from each part in the Tverberg partition). Let Q' be any s of these points. Pick any set $R' \subset R$ of $\lceil d/2 \rceil + 1$ points that interleaves Q'. By Lemma 3, the sets $\operatorname{conv} Q'$ and $\operatorname{conv} R'$ intersect, contradicting the fact that H separates Q' from R. □

3.3 A Regularity Lemma for Semialgebraic Predicates

We shall use a regularity lemma of Fox–Pach–Suk [16], which is a quantitative improvement over the prior version due to Fox–Gromov–Lafforgue–Naor–Pach [15]. The improvement is due to the use of the efficient cuttings of Chazelle–Friedman [9] and Clarkson [10].

Consider a polynomial $f \in \mathbb{R}[\vec{x}_1, \dots, \vec{x}_k]$, where each $\vec{x}_i$ is a vector of d indeterminates. The *degree* of f in $\vec{x}_i$ is the degree of f as a polynomial in $\vec{x}_i$ while regarding $\vec{x}_j$ for $i \neq j$ as constants. We say that f is of *complexity* at most D if it is of degree at most D in each of $\vec{x}_1, \dots, \vec{x}_k$.

Lemma 5 (Theorem 1.3 in [16]) *For any $k, d, t, D \in \mathbb{N}$ there exists a constant $c = c(k, d, t, D) > 0$ with the following property. Let $0 < \gamma < 1/2$, let $P \subset \mathbb{R}^d$ be a finite multiset, and let $f_1, \dots, f_t \in \mathbb{R}[\vec{x}_1, \dots, \vec{x}_k]$ be t polynomials of complexity at most D each. Then there exists a partition $P = P_1 \cup \dots \cup P_M$ of P into at most $M \le (1/\gamma)^c$ parts, and a small set $\mathcal{E} \subset [M]^k$ of "exceptional" k-tuples, satisfying the following:*

1. *The exceptions are few: $|\mathcal{E}| \le \gamma M^k$,*
2. *Almost all k-tuples are regular: whenever $(i_1, \dots, i_k) \notin \mathcal{E}$ and $p_1 \in P_{i_1}, \dots, p_k \in P_{i_k}$, then the sign of*

$$f_j(p_1, \dots, p_k)$$

   depends only on j and on the tuple $(i_1, \dots, i_k)$ but not on the actual choice of the points $p_1, \dots, p_k$. (Note that the elements $i_1, \dots, i_k$ of the tuple need not be distinct nor in increasing order.)
3. *The partition is an equipartition: For all i, j the cardinalities of P_i and of P_j differ by at most one.*

(The statement appearing in [16] is slightly different: In part (2) instead of claiming that the signs of all f_j are constant, the original merely states that an arbitrary fixed Boolean formula in signs of f_j is constant. However, their proof actually establishes the stronger statement above. Alternatively, one may refine the partition $\mathcal{P}$ by iterative application of the original statement to each f_j in turn. The only minor drawback is that instead of a true equipartition one would then obtain a partition whose parts differ by as much as t, the number of polynomials.)

The main point of Lemma 5 is that the number M of parts is independent of $|P|$ (otherwise we could trivially partition P into parts of size 1). The price for this independence is the small set $\mathcal{E}$ which indexes "irregular" tuples.

Invoking Lemma 5 with the orientation predicate, we obtain the following result, which is what we actually need:

Corollary 6 *For each d there exists a constant $c = c(d) > 0$ with the following property. Let $0 < \gamma < 1/2$, and let $P \subset \mathbb{R}^d$ be a finite point set in general position. Then there exists a partition $P = P_1 \cup \cdots \cup P_M$ of P into M parts, with $1/\gamma \leq M \leq 2(1/\gamma)^c$, and a small hypergraph $\mathcal{H} \subset \binom{[M]}{d+1}$ of "exceptional" $(d + 1)$-sets, satisfying the following:*

1. $|\mathcal{H}| \leq \gamma\binom{M}{d+1}$,
2. *Whenever $\{i_0, i_1, \ldots, i_d\} \in \binom{[M]}{d+1} \setminus \mathcal{H}$ and $p_0 \in P_{i_0}, p_1 \in P_{i_1}, \ldots, p_d \in P_{i_d}$, then the sign of* $\operatorname{orient}(p_0, p_1, \ldots, p_d)$ *depends only on the tuple $(i_0, \ldots, i_d)$ but not on the actual choice of the points $p_0, \ldots, p_d$. (The sign of* orient *obviously does depend on the permutation of the elements $i_0, \ldots, i_d$.)*
3. *For all i, j the cardinalities of P_i and of P_j differ by at most one.*

Proof If $|P| \leq 2(1/\gamma)^c$, then simply partition P into parts of size 1. So, assume $|P| \geq 2(1/\gamma)^c$. We apply Lemma 5 with $t \overset{\text{def}}{=} 1$ and $f_1 \overset{\text{def}}{=}$ orient. We obtain a partition of P into $M \leq (1/\gamma)^c$ parts, each of size at least 2, and a set $\mathcal{E} \subset [M]^{d+1}$ of size at most γM^{d+1}.

We now show that all tuples $(i_0, \ldots, i_d) \in [M]^{d+1}$ that contain repeated elements must belong to $\mathcal{E}$. Indeed, consider one such tuple, and say $i_j = i_{j'}$. Since the part P_{i_j} has at least two elements, say p and q, swapping p and q causes orient to flip its nonzero sign (recall that P is in general position). Hence, the tuple $(i_0, \ldots, i_d)$ is not regular, i.e., it does not satisfy property 2 above.

This consideration implies the lower bound for M: We have $\gamma M^{d+1} \geq |\mathcal{E}| \geq M^{d+1} - (d + 1)!\binom{M}{d+1} \geq M^{d+1} - M^d(M - 1) = M^d$, and hence $M \geq 1/\gamma$.

Finally, we let $\mathcal{H}$ consist of all tuples in $\mathcal{E}$ whose elements are pairwise distinct (this definition makes sense since, in our case, $\mathcal{E}$ is invariant under permutations). Since $\mathcal{E}$ contains *all* the tuples with repeated elements, it can contain at most a γ-fraction of the remaining tuples. Therefore, the same is true for $\mathcal{H}$. □

4 Independent Sets in Hypergraphs

We will also need the following bound on hypergraph Turán numbers. We give a simple probabilistic proof based on [2, Theorem 3.2.1], though a stronger bound can be found in [12].

Lemma 7 *Let $r \geq 2$, and suppose $\mathcal{H}$ is an r-uniform hypergraph on n vertices with βn^r edges, where $n \geq \frac{1}{2}\beta^{-1/(r-1)}$. Then $\mathcal{H}$ contains an independent set on at least $\frac{1}{4}\beta^{-1/(r-1)}$ vertices.*

Proof Let $p \overset{\text{def}}{=} \beta^{-1/(r-1)}/(2n)$. Note that $p \leq 1$ by the assumption on n. Let $S \subseteq V(\mathcal{H})$ be a random set where $\Pr[v \in S] = p$ for each $v \in V(\mathcal{H})$ independently. Then the expected number of edges spanned by S is $p^r \beta n^r$. For each edge in S we may remove one vertex to obtain an independent set. Hence, $\mathcal{H}$ contains an independent set of size at least $\mathbb{E}[I] = pn - p^r\beta n^r \geq \frac{1}{2}\beta^{-1/(r-1)} - \frac{1}{2^r}\beta^{1-r/(r-1)}/n^r \cdot n^r$. □

5 Interval Chains

We will reduce the geometric problem of constructing one-sided ε-approximants to a combinatorial problem about interval chains. Let $[i, j]$ denote the interval of integers $\{i, i+1, \dots, j\}$. We still write $[t]$ for $\{1, 2, \dots, t\}$. An *interval chain* of size k (also called k-chain) in $[t]$ is a sequence of k consecutive, disjoint, nonempty intervals

$$I \overset{\text{def}}{=} [a_1, a_2 - 1][a_2, a_3 - 1] \cdots [a_k, a_{k+1} - 1],$$

where $1 \leq a_1 < a_2 < \cdots < a_{k+1} \leq t + 1$. Interval chains were introduced by Condon and Saks [11]. They were subsequently used by Alon, Kaplan, Nivasch, Sharir and Smorodinsky [4] and by Bukh, Matoušek, Nivasch [8] to obtain bounds for weak ε-nets for orientation-homogeneous point sets.

A D-tuple of integers $(x_1, \dots x_D)$ is said to *stab* a k-chain I if each x_i lies in a different interval of I.

The problem considered in [4] was to build, for given D, k, and t, a small-sized family $\mathcal{F}$ of D-tuples that stab all k-chains in $[t]$. Phrased differently, for each interval chain I with *at least* k intervals, there should be *at least* one D-tuple in $\mathcal{F}$ that stabs I.

In contrast, here we will consider the following problem: Given D, ε, and t, we want to build a small-sized family (multiset) $\mathcal{F}$ of D-tuples such that, for each interval chain I in $[t]$, if αt is the number of intervals in I, then at least an $(\alpha - \varepsilon)$-fraction of the D-tuples in $\mathcal{F}$ stab I. We call such an $\mathcal{F}$ an *ε-approximating family.*

Our construction of ε-approximating families is similar to the statement of the regularity lemma for words, due to Axenovich, Person and Puzynina [5]. The lemma, which was also independently discovered by Feige, Koren and Tennenholtz [14] under the name of 'local repetition lemma', can be used directly to construct ε-approximating families. Doing so yields a family whose size is exponential in $1/\varepsilon$. In contrast, we avoid using the full strength of the regularity lemma and obtain a construction of polynomial size.

Lemma 8 *Suppose $D \geq 2$ and $0 < \varepsilon < 1$. Let $K \overset{\text{def}}{=} \lceil (D-1)\ln(4/\varepsilon) \rceil$ and $t \overset{\text{def}}{=} m(D-1)^K$ for some integer $m \geq 4/\varepsilon$. Then there exists an ε-approximating family $\mathcal{F}$ of D-tuples in $[t]$, of size $|\mathcal{F}| \leq t$.*

Proof The argument is more conveniently phrased in the "dual" setting, in which D-tuples become $(D-1)$-interval chains and ℓ-interval chains become $(\ell+1)$-tuples. Namely, the D-tuple $(x_1, \dots, x_D)$ becomes the interval chain $[x_1+1, x_2] \cdots [x_{D-1}+1, x_D]$, and the ℓ-interval chain $[a_1, a_2-1] \cdots [a_\ell, a_{\ell+1}-1]$ becomes the tuple $(a_1, a_2, \dots, a_{\ell+1})$. Then a $(D-1)$-chain C "stabs" a tuple T if T contains points on both sides of C, as well as inside each interval of C.

For each $k = 0, 1, \dots, K-1$, we partition $[t]$ into disjoint intervals of length $(D-1)^k$, by letting $B_{k,i} \overset{\text{def}}{=} \left[(i-1)(D-1)^k+1, i(D-1)^k\right]$ for $1 \le i \le t/(D-1)^k$. Then we group these intervals into disjoint $(D-1)$-chains, by letting

$$\mathcal{F}_k \overset{\text{def}}{=} \{B_{k,(i-1)(D-1)+1} \cdots B_{k,i(D-1)} : 1 \le i \le t/(D-1)^{k+1}\}.$$

We call each $\mathcal{F}_k$ a *layer*. Note that each chain in $\mathcal{F}_k$ fits exactly in an interval of layer $k+1$.

Then we define the multiset $\mathcal{F}$ by taking w_k copies of $\mathcal{F}_k$ for each $0 \le k \le K-1$, where

$$w_k \overset{\text{def}}{=} (D-2)^k.$$

Hence, letting $E \overset{\text{def}}{=} (D-2)/(D-1)$, we have $|\mathcal{F}| = \sum_{k=0}^{K-1} w_k |\mathcal{F}_k| = t(1-E^K)$. Therefore, by the choice of K,

$$t/2 \le |\mathcal{F}| < t$$

Let J be a subset of $[t]$, and let αt be the size of J. We claim that at least an $(\alpha - \varepsilon)$-fraction of the chains in $\mathcal{F}$ stab J.

Call a $(D-1)$-chain $C \in \mathcal{F}$ *empty* if J does not intersect any interval of C, and *occupied* otherwise. If C is occupied, then call it *fully occupied* if J intersects all intervals of C, and *partially occupied* otherwise.

For each $0 \le k \le K-1$, let β_k denote the fraction of chains of $\mathcal{F}_k$ that are occupied by J, and let $\gamma_k \le \beta_k$ denote the fraction of chains of $\mathcal{F}_k$ that are partially occupied by J.

Claim 1 *For each k we have $\beta_k \ge \alpha + (\gamma_0 + \cdots + \gamma_k)/(D-1)$.*

Proof For each layer j, Let $\mathcal{F}'_j$ be the set of occupied chains of $\mathcal{F}_j$, and let $\mathcal{F}''_j \subset \mathcal{F}'_j$ be the set of those that are only partially occupied. Hence, $\mathcal{F}''_j$ covers a γ_j-fraction of $[t]$. From each chain $C \in \mathcal{F}''_j$ choose an empty interval, and let $\mathcal{B}_j$ be the union of these empty intervals. Hence, $\mathcal{B}_j$ covers a $(\gamma_j/(D-1))$-fraction of $[t]$. Furthermore, since each chain in $\mathcal{F}''_j$ contains a point of J, the sets $\mathcal{B}_0, \dots, \mathcal{B}_k$ must be pairwise disjoint, as well as disjoint from J, and their union $\mathcal{U} \overset{\text{def}}{=} \mathcal{B}_0 \cup \cdots \mathcal{B}_k \cup J$ must be completely contained in the union of $\mathcal{F}'_k$. Hence, $\mathcal{F}'_k$ covers at least an $\big(\alpha + (\gamma_0 + \cdots + \gamma_k)/(D-1)\big)$-fraction of $[t]$, and the claim follows. □

Let us now derive a lower bound on the number of fully occupied chains in $\mathcal{F}$. By some tedious calculations we obtain:

$$\begin{aligned}\sum_{k=0}^{K-1}(\beta_k-\gamma_k)w_k|\mathcal{F}_k| &\geq \sum_{k=0}^{K-1}\Big(\alpha+\frac{\gamma_0+\cdots+\gamma_k}{D-1}-\gamma_k\Big)w_k|\mathcal{F}_k| \\ &= \alpha|\mathcal{F}|+\sum_{k=0}^{K-1}\gamma_k\left(\frac{1}{D-1}\sum_{j=k}^{K-1}w_j|\mathcal{F}_j|-w_k|\mathcal{F}_k|\right) \\ &= \alpha|\mathcal{F}|-t\frac{E^K}{D-1}\sum_{k=0}^{K-1}\gamma_k \geq \alpha|\mathcal{F}|-tE^K\beta_{K-1} \\ &\geq \left(\alpha-2E^K\right)|\mathcal{F}| \geq (\alpha-\varepsilon/2)|\mathcal{F}|;\end{aligned}$$

where the upper bound for $\sum\gamma_k$ was obtained from Claim 1.

Finally, note that in each layer $\mathcal{F}_k$ there are at most two fully occupied chains that do not stab J. Since $|\mathcal{F}_k| \geq m \geq 4/\varepsilon$, the said chains constitute at most an $(\varepsilon/2)$-fraction of $\mathcal{F}$. □

6 Construction of the One-Sided Approximants

In this section we prove Theorem 2.

Let s and D be as in Lemma 4. Then let t be as small as possible to satisfy the condition of Lemma 8 with $\varepsilon/2$ in place of ε (so t is polynomial in $1/\varepsilon$). Then define

$$u \overset{\text{def}}{=} \lceil 4/\varepsilon\rceil, \qquad n_0 \overset{\text{def}}{=} tu, \qquad N \overset{\text{def}}{=} \mathrm{OT}_d(n_0), \qquad \beta \overset{\text{def}}{=} (4N)^{-d}, \qquad \gamma \overset{\text{def}}{=} \beta(\varepsilon/5)^{d+1}; \tag{1}$$

where the function $\mathrm{OT}_d(n_0)$ is defined at the end of Sect. 3.1. Invoking Lemma 8, let $\mathcal{F}$ be an $(\varepsilon/2)$-approximating family of D-tuples in $[t]$, of size $|\mathcal{F}| \leq t$.

Let $P \subset \mathbb{R}^d$ be a given finite point set, and let $n \overset{\text{def}}{=} |P|$. We will construct a one-sided ε-approximant multiset A for P. If $n \leq 40/(\varepsilon\gamma^c)$ for the constant c of Corollary 6, then simply let $A \overset{\text{def}}{=} P$. Hence, assume $n \geq 40/(\varepsilon\gamma^c)$. In this case, our multiset A will consist of Tverberg points of certain D-tuples of points of P.

We first handle the case when P is in general position; then we handle degeneracies with a simple perturbation argument. Hence, suppose the point set $P \subset \mathbb{R}^d$ is in general position (no $d+1$ points are affinely dependent).

We start by invoking Corollary 6 on P and the parameter γ given in (1). We obtain a partition of P into $1/\gamma \leq M \leq 2(1/\gamma)^c$ almost-equal-sized *parts* $P_1,\ldots,P_M$, and a corresponding hypergraph $\mathcal{H} \subseteq \binom{[M]}{d+1}$ of size $|\mathcal{H}| \leq \gamma\binom{M}{d+1}$.

We make all parts have exactly the same size by discarding at most one point from each part. Hence we discard at most $M \leq 2(1/\gamma)^c$ points. Since $n \geq 40/(\varepsilon\gamma^c)$, we discarded at most an $(\varepsilon/20)$-fraction of the points of P. By a slight abuse of notation, we denote the new parts by the same names $P_1, \dots, P_M$. We will consider $P_1, \dots, P_M$ as an *ordered* sequence (where the order was chosen arbitrarily).

Let $\widehat{P} = (p_1, \dots, p_M)$, where $p_i \in P_i$ for all i, be an arbitrarily chosen sequence of representatives from the parts. We will now repeatedly "fish out" equal-length orientation-homogeneous subsequences from $\widehat{P}$, until there are too few points left to continue the process. For this purpose, let $\widehat{P}_1 \overset{\text{def}}{=} \widehat{P}$, and let $i \leftarrow 1$. Repeat the following: If $\left|\widehat{P}_i\right| < \varepsilon M/5$ then stop. Otherwise, $\widehat{P}_i$ is large enough so that the number of edges of $\mathcal{H}$ spanned by $\widehat{P}_i$ is at most

$$|\mathcal{H}| \leq \gamma \binom{M}{d+1} \leq \gamma M^{d+1} = \beta(\varepsilon M/5)^{d+1} \leq \beta \left|\widehat{P}_i\right|^{d+1}.$$

In view of $M \geq 1/\gamma$, we also have $\varepsilon M/5 \geq (5/\varepsilon)^d/\beta \geq \frac{1}{2}\beta^{-1/d}$. Hence, we can apply Lemma 7 on $\widehat{P}_i$ with $r = d + 1$. We conclude that $\widehat{P}_i$ has an independent set of size N. By the definition of N, that independent set has an orientation-homogeneous subsequence S_i of length n_0. Let $\widehat{P}_{i+1} \overset{\text{def}}{=} \widehat{P}_i \setminus S_i$, increase i by 1, and return to the beginning of the loop.

At the end of this process, we get orientation-homogeneous sequences $S_1, S_2, \dots, S_m$ for some $m \leq M/n_0$, and a leftover sequence $S^* \overset{\text{def}}{=} \widehat{P}_{m+1}$ of size at most $\varepsilon M/5$. From each S_i we will now construct a multiset A_i of Tverberg points; their union will be our desired multiset A.

So fix i, and denote $S_i = (q_0, q_1, q_2, \dots, q_{n_0-1})$. Let $v_j \overset{\text{def}}{=} q_{(j-1)u}$ for all $1 \leq j \leq t$. We will call the elements v_j *separators*. Let $\mathfrak{v} \overset{\text{def}}{=} (v_1, \dots, v_t)$. For each $j = 1, \dots, t$, define the *block* $b_j \overset{\text{def}}{=} (q_{(j-1)u+1}, \dots, q_{ju-1})$, which contains the elements of S_i between separators v_j and v_{j+1}. Let

$$\mathcal{B}_j \overset{\text{def}}{=} \bigcup_{p_k \in b_j} P_k$$

be the union of all the parts that correspond to points of block b_j.

To each D-tuple $\overline{x} = (x_1, \dots, x_D) \in \mathcal{F}$, associate the D-tuple of separators $Q_{\overline{x}} \overset{\text{def}}{=} \{v_{x_1}, \dots, v_{x_D}\}$. Then define the multiset

$$A_i \overset{\text{def}}{=} \{\mathrm{Tver}_s(Q_{\overline{x}}) : \overline{x} \in \mathcal{F}\}.$$

Lemma 9 *Let $C \subseteq \mathbb{R}^d$ be a convex set. Take the set of indices $J \overset{def}{=} \{j : \mathcal{B}_j \cap C \neq \emptyset\}$. List the elements of J in increasing order as $J = \{j_1, j_2, \dots, j_\ell\}$. Let I be the $(\ell - 1)$-interval chain:*

$$I \overset{def}{=} [j_1 + 1, j_2][j_2 + 1, j_3] \cdots [j_{\ell-1} + 1, j_\ell].$$

Then, if the D-tuple $\overline{x} \in \mathcal{F}$ *stabs I, then C contains the corresponding Tverberg point* $\mathrm{Tver}_s(Q_{\overline{x}})$.

Proof Suppose $\overline{x} = (x_1, \dots, x_D)$ stabs I. Then there exists a subset $J' \overset{\text{def}}{=} (j'_0, \dots, j'_D) \subset J$ such that $j'_0 < x_1 \le j'_q < \cdots < x_D \le j'_D$. For each $j \in J$ there is a part $P(j)$ whose representative point $p(j)$ belongs to the block b_j, and such that C contains some point $p'(j) \in P(j)$. The sequence of representatives $p(j'_0), v_{x_1}, p(j'_1), \dots, v_{x_D}, p(j'_D)$, being a subsequence of S_i, is orientation-homogeneous. Therefore, by regularity, and since S_i avoids the hypergraph $\mathcal{H}$, the sequence $p'(j'_0), v_{x_1}, p'(j'_1), \dots, v_{x_D}, p'(j'_D)$ is also orientation-homogeneous. Therefore, by Lemma 4, we have

$$\mathrm{Tver}_s(Q_{\overline{x}}) \in \operatorname{conv}\{p'(j'_0), \dots, p'(j'_D)\} \subset C,$$

as desired. □

Let $\mathcal{S}_i \overset{\text{def}}{=} \bigcup_{p_j \in S_i} P_j$ be the union of all the parts whose representative points belong to S_i.

Corollary 10 *Let* $C \subseteq \mathbb{R}^d$ *be a convex set, and let* α *be the fraction of the points of* $\mathcal{S}_i$ *contained in C. Then C contains at least an* $(\alpha - 3\varepsilon/4)$*-fraction of the points of* A_i.

Proof Since $|\mathfrak{v}| = t \le \varepsilon n_0/4$, and since all the parts $P_1, \dots, P_M$ have equal size, the set C meets at least an $(\alpha - \varepsilon/4)$-fraction of the sets $\mathcal{B}_j$. The desired conclusion follows from Lemma 9 since $\mathcal{F}$ is $(\varepsilon/2)$-approximating. □

Finally, let

$$A \overset{\text{def}}{=} \bigcup_{i=1}^{m} A_i.$$

With some patience, we can use (1) and the bound $\mathrm{OT}_d(n) \le \mathrm{tw}_d(c_d n)$ mentioned above to obtain the bound $|A| = M|\mathcal{F}| \le \mathrm{tw}_d\left(\varepsilon^{-c'}\right)$ for some constant $c' = c'(d) > 1$.

Note that at most $(\varepsilon/20)n + (\varepsilon/5)n = \varepsilon n/4$ points of P were either discarded in making $P_1, \dots, P_M$ equal or were relegated to the "leftover" $S^* = \widehat{P} \setminus (S_1 \cup \cdots \cup S_m)$. So, if a convex set C contains an α-fraction of the points of P, and an α_i-fraction of the points of $\mathcal{S}_i$ for each i, then $\operatorname{avg}_i \alpha_i \ge \alpha - \varepsilon/4$.

By Corollary 10, C contains at least an $(\alpha_i - 3\varepsilon/4)$-fraction of the points of A_i. Hence, averaging again, C contains an $(\alpha - \varepsilon)$-fraction of the points of A.

This concludes the proof of Theorem 2 for the case when P is in general position.

If $P = \{p_1, \dots, p_n\}$ is not in general position, take an arbitrarily small continuous perturbation $P(t) \overset{\text{def}}{=} \{p_1(t), \dots, p_n(t)\}$ such that $P(0) = P$ and $P(t)$ is in general position for all $0 < t \le 1$. For each $t > 0$ we apply the above argument on $P(t)$; we get a family $\mathcal{I}(t) \subset \binom{[n]}{D+1}$ such that multiset $A(t) \overset{\text{def}}{=} \{\mathrm{Tver}_s(P(t)_I) : I \in \mathcal{I}(t)\}$ is a one-sided ε-approximant for $P(t)$. Since P is finite, there are only a finitely many possible values for $\mathcal{I}(t)$, so one of them occurs infinitely often for $t = t_1, t_2, t_3, \dots$

with $\lim t_i = 0$. Then, by a standard argument, the limit multiset $\lim_{i\to\infty} A(t_i)$ exists and is a one-sided ε-approximant for P.

7 Problems and Remarks

- The main problem is to prove reasonable upper bounds on $g(\varepsilon, d)$. The only known lower bound on $g(\varepsilon, d)$ is of the form $c_d(1/\varepsilon)\log^{d-1}(1/\varepsilon)$. It is a consequence of the lower bounds on the size of weak ε-nets [8] and the fact that every one-sided ε-approximant is an ε-net.
- Much smaller one-sided approximants can be constructed if P is orientation-homogeneous: We apply the same construction that was applied to individual sets S_i in Sect. 6 to the set P (with $u \overset{\text{def}}{=} |P|/t$ instead of $u = n_0/t$), obtaining one-sided ε-approximants of size polynomial in $1/\varepsilon$. While this bound is much better than the general bound on $g(\varepsilon, d)$ from Theorem 2, it is still far from the known bounds for ε-nets: Every orientation-homogeneous set admits an ε-net of size only $O\big(\varepsilon^{-1}\alpha(\varepsilon^{-1})\big)$ in the plane and of size only $\varepsilon^{-1}2^{\alpha(\varepsilon^{-1})^{O(1)}}$ in $\mathbb{R}^d$ for $d \geq 3$, where α is the inverse Ackermann function [4].
- The *diagonal of the stretched grid* is a specific orientation-homogeneous sequence considered in [7] and in [8]. Denote it D. The authors in [8] obtained a lower bound for ε-nets for D from the lower bound for the interval chains problem considered in [4]. Similarly, a lower bound for the interval chains problem discussed in Sect. 5 would yield a lower bound for ε-approximants for D.
- In Theorem 2 it is possible to assure that the one-sided approximant A is a genuine set rather than a multiset. It is easy to do so if P is in general position, as we may simply perturb each point of A slightly. In general, we cannot ensure that each sequence S_i is orientation-homogeneous, but we can ensure that each S_i is orientation-homogeneous inside the affine subspace $\operatorname{ahull} S_i$. That can be done by using Ramsey's theorem to extract subsequences of $\widehat{P}$ that lie in a proper affine subspace, and then using the induction on the dimension. We can then perturb the points of A_i within $\operatorname{ahull} S_i$. The rest of the argument remains the same.

Acknowledgements We thank the referees for pointing omissions and typos. We are also grateful to Po-Shen Loh and the late Jirka Matoušek for discussions.

References

1. R. Alexander, Geometric methods in the study of irregularities of distribution. Combinatorica **10**(2), 115–136 (1990)
2. N. Alon, J.H. Spencer, *The Probabilistic Method*, 2nd edn. Wiley-Interscience Series in Discrete Mathematics and Optimization (Wiley, New York, 2000). With an appendix on the life and work of Paul Erdős

3. N. Alon, I. Bárány, Z. Füredi, D.J. Kleitman, Point selections and weak ϵ-nets for convex hulls. Combin. Probab. Comput. **1**(3), 189–200 (1992). http://www.tau.ac.il/~nogaa/PDFS/abfk3.pdf
4. N. Alon, H. Kaplan, G. Nivasch, M. Sharir, S. Smorodinsky, Weak ϵ-nets and interval chains. J. ACM **55**(6), Art. 28, 32 (2008). http://www.gabrielnivasch.org/academic/publications/interval_chains.pdf
5. M. Axenovich, Y. Person, S. Puzynina, A regularity lemma and twins in words. J. Combin. Theory Ser. A **120**(4), 733–743 (2013). arXiv:1204.2180
6. I. Bárány, J. Matoušek, A. Pór, Curves in $\mathbb{R}^d$ intersecting every hyperplane at most $d + 1$ times, in *Computational Geometry (SoCG'14)* (ACM, New York, 2014), pp. 565–571. arXiv:1309.1147
7. B. Bukh, J. Matoušek, G. Nivasch, Stabbing simplices by points and flats. Discret. Comput. Geom. **43**(2), 321–338 (2010). arXiv:0804.4464
8. B. Bukh, J. Matoušek, G. Nivasch, Lower bounds for weak epsilon-nets and stair-convexity. Israel J. Math. **182**, 199–208 (2011). arXiv:0812.5039
9. B. Chazelle, J. Friedman, A deterministic view of random sampling and its use in geometry. Combinatorica **10**(3), 229–249 (1990)
10. K.L. Clarkson, A randomized algorithm for closest-point queries. SIAM J. Comput. **17**(4), 830–847 (1988)
11. A. Condon, M. Saks, A limit theorem for sets of stochastic matrices. Linear Algebra Appl. **381**, 61–76 (2004)
12. D. de Caen, Extension of a theorem of Moon and Moser on complete subgraphs. Ars Combin. **16**, 5–10 (1983)
13. M. Eliáš, J. Matoušek, E. Roldán-Pensado, Z. Safernová, Lower bounds on geometric Ramsey functions. SIAM J. Discret. Math. **28**(4), 1960–1970 (2014). arXiv:1307.5157
14. U. Feige, T. Koren, M. Tennenholtz, Chasing ghosts: competing with stateful policies, in *55th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2014* (IEEE Computer Society, Los Alamitos, 2014), pp. 100–109. arXiv:1407.7635
15. J. Fox, M. Gromov, V. Lafforgue, A. Naor, J. Pach, Overlap properties of geometric expanders. J. Reine Angew. Math. **671**, 49–83 (2012). arXiv:1005.1392
16. J. Fox, J. Pach, A. Suk, A polynomial regularity lemma for semi-algebraic hypergraphs and its applications in geometry and property testing, in *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms* (2015). arXiv:1502.01730v1
17. D. Haussler, E. Welzl, ε-nets and simplex range queries. Discret. Comput. Geom. **2**(2), 127–151 (1987)
18. J. Matoušek, Tight upper bounds for the discrepancy of half-spaces. Discret. Comput. Geom. **13**(3–4), 593–601 (1995)
19. J. Matoušek, *Geometric Discrepancy: An Illustrated Guide*. Volume 18 of Algorithms and Combinatorics (Springer, Berlin, 1999)
20. J. Matoušek, *Lectures on Discrete Geometry*. Volume 212 of Graduate Texts in Mathematics (Springer, New York, 2002)
21. J. Radon, Mengen konvexer Körper, die einen gemeinsamen Punkt enthalten. Math. Ann. **83**(1–2), 113–115 (1921). http://gdz.sub.uni-goettingen.de/dms/load/img/?PID=GDZPPN002267888
22. A. Suk, A note on order-type homogeneous point sets. Mathematika **60**, 37–42 (2014). arXiv:1305.5934
23. V.N. Vapnik, A.Ja. Červonenkis, The uniform convergence of frequencies of the appearance of events to their probabilities. Teor. Verojatnost. i Primenen. **16**, 264–279 (1971). http://mi.mathnet.ru/tvp2146
24. G.M. Ziegler, *Lectures on Polytopes*. Volume 152 of Graduate Texts in Mathematics (Springer, New York, 1995)

A Note on Induced Ramsey Numbers

David Conlon, Domingos Dellamonica, Steven La Fleur, Vojtěch Rödl, and Mathias Schacht

Dedicated to the memory of Jirka Matoušek

Abstract The induced Ramsey number $r_{\text{ind}}(F)$ of a k-uniform hypergraph F is the smallest natural number n for which there exists a k-uniform hypergraph G on n vertices such that every two-coloring of the edges of G contains an induced monochromatic copy of F. We study this function, showing that $r_{\text{ind}}(F)$ is bounded above by a reasonable power of $r(F)$. In particular, our result implies that $r_{\text{ind}}(F) \leq 2^{2^{ct}}$ for any 3-uniform hypergraph F with t vertices, mirroring the best known bound for the usual Ramsey number. The proof relies on an application of the hypergraph container method.

1 Introduction

The *Ramsey number* $r(F; q)$ of a k-uniform hypergraph F is the smallest natural number n such that every q-coloring of the edges of $K_n^{(k)}$, the complete k-uniform hypergraph on n vertices, contains a monochromatic copy of F. In the particular

The first author was supported by a Royal Society University Research Fellowship.

The fourth author was partially supported by NSF grants DMS-1102086 and DMS-1301698.

The fifth author was supported through the Heisenberg-Programme of the DFG.

D. Conlon
Mathematical Institute, University of Oxford, Oxford, UK
e-mail: david.conlon@maths.ox.ac.uk

D. Dellamonica • S. La Fleur • V. Rödl (✉)
Department of Mathematics and Computer Science, Emory University, Atlanta, GA, USA
e-mail: ddellam@emory.edu; slafleu@emory.edu; rodl@emory.edu

M. Schacht
Fachbereich Mathematik, Universität Hamburg, Hamburg, Germany
e-mail: schacht@math.uni-hamburg.de

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_13

case when $q = 2$, we simply write $r(F)$. The existence of $r(F; q)$ was established by Ramsey in his foundational paper [17] and there is now a large body of work studying the Ramsey numbers of graphs and hypergraphs. For a recent survey, we refer the interested reader to [5].

In this paper, we will be concerned with a well-known refinement of Ramsey's theorem, the induced Ramsey theorem. We say that a k-uniform hypergraph F is an *induced subgraph* of another k-uniform hypergraph G if $V(F) \subset V(G)$ and any k vertices in F form an edge if and only if they also form an edge in G. The *induced Ramsey number* $r_{\text{ind}}(F; q)$ of a k-uniform hypergraph F is then the smallest natural number n for which there exists a k-uniform hypergraph G on n vertices such that every q-coloring of the edges of G contains an induced monochromatic copy of F. Again, in the particular case when $q = 2$, we simply write $r_{\text{ind}}(F)$.

For graphs, the existence of induced Ramsey numbers was established independently by Deuber [6], Erdős, Hajnal, and Pósa [9], and Rödl [18], while for k-uniform hypergraphs with $k \geq 3$ their existence was shown independently by Nešetřil and Rödl [16] and Abramson and Harrington [1]. The bounds that these original proofs gave on $r_{\text{ind}}(F; q)$ were enormous. However, at that time it was noted by Rödl (unpublished) that for bipartite graphs F the induced Ramsey numbers are exponential in the number of vertices. Moreover, it was conjectured by Erdős [7] that there exists a constant c such that every graph F with t vertices satisfies $r_{\text{ind}}(F) \leq 2^{ct}$. If true, the complete graph would show that this is best possible up to the constant c. A result of Conlon, Fox, and Sudakov [3], building on earlier work by Kohayakawa, Prömel, and Rödl [13], comes close to establishing this conjecture, showing that

$$r_{\text{ind}}(F) \leq 2^{ct \log t}.$$

However, the method used to prove this estimate only works in the 2-color case. For $q \geq 3$, the best known bound, due to Fox and Sudakov [11], is $r_{\text{ind}}(F; q) \leq 2^{ct^3}$, where c depends only on q.

In this note, we study the analogous question for hypergraphs, showing that the induced Ramsey number is never significantly larger than the usual Ramsey number. Our main result is the following.

Theorem 1 *Let F be a k-uniform hypergraph with t vertices and ℓ edges. Then there are positive constants c_1, c_2, and c_3 such that*

$$r_{\text{ind}}(F; q) \leq 2^{c_1 k\ell^3 \log(qt\ell)} R^{c_2 k\ell^2 + c_3 t\ell},$$

where $R = r(F; q)$ is the classical q-color Ramsey number of F.

Define the tower function $t_k(x)$ by $t_1(x) = x$ and, for $i \geq 1$, $t_{i+1}(x) = 2^{t_i(x)}$. A seminal result of Erdős and Rado [8] says that

$$r(K_t^{(k)}; q) \leq t_k(ct),$$

where c depends only on k and q. This yields the following immediate corollary of Theorem 1.

Corollary 1 *For any natural numbers $k \geq 3$ and $q \geq 2$, there exists a constant c such that if F is a k-uniform hypergraph with t vertices, then*

$$r_{\mathrm{ind}}(F; q) \leq t_k(ct).$$

A result of Erdős and Hajnal (see, for example, Chapter 4.7 in [12] and [4]) says that

$$r(K_t^{(k)}; 4) \geq t_k(c't),$$

where c' depends only on k. Therefore, the Erdős–Rado bound is sharp up to the constant c for $q \geq 4$. By taking $F = K_t^{(k)}$, this also implies that Corollary 1 is tight up to the constant c for $q \geq 4$. Whether it is also sharp for $q = 2$ and 3 depends on whether $r(K_t^{(k)}) \geq t_k(c't)$, though determining if this is the case is a famous, and seemingly difficult, open problem.

The proof of Theorem 1 relies on an application of the hypergraph container method of Saxton and Thomason [20] and Balogh, Morris, and Samotij [2]. In Ramsey theory, the use of this method was pioneered by Nenadov and Steger [14] and developed further by Rödl, Ruciński, and Schacht [19] in order to give an exponential-type upper bound for Folkman numbers. Our modest results are simply another manifestation of the power of this beautiful method.

2 Proof of Theorem 1

In order to state the container theorem we first need some definitions. Recall that the degree $d(\sigma)$ of a set of vertices σ in a hypergraph H is the number of edges of H containing σ, while the average degree is the average of $d(v) := d(\{v\})$ over all vertices v.

Definition 2 Let H be an ℓ-uniform hypergraph of order N with average degree d. Let $\tau > 0$. Given $v \in V(H)$ and $2 \leq j \leq \ell$, let

$$d^{(j)}(v) = \max\left\{d(\sigma) \ : \ v \in \sigma \subset V(H), |\sigma| = j\right\}.$$

If $d > 0$, define δ_j by the equation

$$\delta_j \tau^{j-1} N d = \sum_v d^{(j)}(v).$$

The *codegree function* $\delta(H,\tau)$ is then defined by

$$\delta(H,\tau) = 2^{\binom{\ell}{2}-1} \sum_{j=2}^{\ell} 2^{-\binom{j-1}{2}} \delta_j.$$

If $d = 0$, define $\delta(H,\tau) = 0$.

The precise lemma we will need is a slight variant of Corollary 3.6 from Saxton and Thomason's paper [20]. A similar version was already used in the work of Rödl, Ruciński, and Schacht [19] and we refer the interested reader to that paper for a thorough discussion.

Lemma 3 *Let H be an ℓ-uniform hypergraph on N vertices with average degree d. Let $0 < \varepsilon < 1/2$. Suppose that τ satisfies $\delta(H,\tau) \le \varepsilon/12\ell!$ and $\tau \le 1/144\ell!^2\ell$. Then there exists a collection $\mathcal{C}$ of subsets of $V(H)$ such that*

(i) for every set $I \subset V(H)$ such that $e(H[I]) \le \varepsilon\tau^{\ell}e(H)$, there is $C \in \mathcal{C}$ with $I \subset C$,
(ii) $e(H[C]) \le \varepsilon e(H)$ for all $C \in \mathcal{C}$,
(iii) $\log|\mathcal{C}| \le 1000\ell!^3\ell \log(1/\varepsilon)N\tau\log(1/\tau)$.

Before we give the proof of Theorem 1, we first describe the ℓ-uniform hypergraph H to which we will apply Lemma 3.

Construction 4 Given a k-uniform hypergraph F with ℓ edges, we construct an auxiliary hypergraph H by taking

$$V(H) = \binom{[n]}{k} \qquad \text{and} \qquad E(H) = \left\{ E \in \binom{V(H)}{\ell} : E \cong F \right\}.$$

In other words, the vertices of H are the k-tuples of $[n]$ and the edges of H are copies of F in $\binom{[n]}{k}$.

Proof of Theorem 1 Recall that $R = r(F;q)$, the q-color Ramsey number of F, and suppose that F has t vertices and ℓ edges. Let us fix the following numbers:

$$\begin{aligned} \tau &= n^{-\frac{1}{2\ell}}, \qquad p = 1000R^k q\alpha, \qquad \alpha = n^{-\frac{1}{2\ell}+\frac{1}{4\ell(\ell+1)}}, \\ \varepsilon &= 1/(2qR^t), \qquad n = \ell^{40\ell^2(\ell+1)}(1000q)^{8\ell(\ell+1)}R^{4k\ell(\ell+1)+4t\ell}\binom{t}{k}^{4\ell}. \end{aligned} \tag{1}$$

Remark 5 Note that n is bounded above by an expression of the form

$$2^{c_1k\ell^3\log(qt\ell)}R^{c_2k\ell^2+c_3t\ell},$$

as required.

Obviously, $R \geq t$ and one can check that p and n satisfy the following conditions, which we will make use of during the course of the proof:

$$p \leq 1, \tag{2}$$

$$n \geq (24 \cdot 2^{\binom{\ell}{2}} t^t q \ell! R^t)^2, \tag{3}$$

$$n > (144 \ell!^2 \ell)^{2\ell}, \tag{4}$$

$$n > \ell^{40\ell^2(\ell+1)}, \tag{5}$$

$$n > (1000q)^{8\ell(\ell+1)} R^{4k\ell(\ell+1)+4t\ell} \binom{t}{k}^{4\ell}. \tag{6}$$

We will show that, with positive probability, a random hypergraph $G \in \mathbb{G}^{(k)}(n,p)$ has the property that every q-coloring of its edges contains an induced monochromatic copy of F. The proof proceeds in two stages. First, we use Lemma 3 to show that, with probability $1 - o(1)$, G has the property that any q-coloring of its edges yields many monochromatic copies of F. Then we show that some of these monochromatic copies must be induced.

More formally, let X be the event that there is a q-coloring of the edges of G which contains at most

$$M := \frac{\varepsilon \tau^{\ell} (n)_t}{\mathrm{aut}(F)}$$

monochromatic copies of F in each color, and let Y be the event that G contains at least M noninduced copies of F. Note that if $\overline{X} \cap \overline{Y}$ happens, then, in any q-coloring, there are more monochromatic copies of F in one of the q colors than there are noninduced copies of F in G. Hence, that color class must contain an induced copy of F.

We now proceed to show that the probability $\mathbf{P}(X)$ tends to zero as n tends to infinity. In order to apply Lemma 3, we need to check that τ and ε satisfy the requisite assumptions with respect to the ℓ-uniform hypergraph H defined in Construction 4. Let $\sigma \subset V(H)$ be arbitrary and define

$$V_\sigma = \bigcup_{v \in \sigma} v \subset [n].$$

For an arbitrary set $W \subset [n] \setminus V_\sigma$ with $|W| = t - |V_\sigma|$, let $\mathrm{emb}_F(\sigma, W)$ denote the number of copies $\widetilde{F}$ of F with $V(\widetilde{F}) = W \cup V_\sigma$ and $\sigma \subset E(\widetilde{F})$. Observe that this number does not actually depend on the choice of W, so we will simply use $\mathrm{emb}_F(\sigma)$ from now on.

Since there are clearly $\binom{n-|V_\sigma|}{t-|V_\sigma|}$ choices for the set W, we arrive at the following claim.

Claim 1 *For any $\sigma \subset V(H)$,*

$$d(\sigma) = \binom{n-|V_\sigma|}{t-|V_\sigma|}\mathrm{emb}_F(\sigma). \qquad \square$$

Let us denote by t_j the minimum number of vertices of F which span j edges. From Claim 1, it follows that for any $\sigma \subset V(H)$ with $|\sigma| = j$, we have

$$d(\sigma) = \binom{n-|V_\sigma|}{t-|V_\sigma|}\mathrm{emb}_F(\sigma) \le \binom{n-t_j}{t-t_j}\mathrm{emb}_F(\sigma).$$

On the other hand, for a singleton $\sigma_1 \subset V(H)$, we have $|V_{\sigma_1}| = k$ and therefore $d = d(\sigma_1)$ is such that

$$\frac{d(\sigma)}{d} \le \frac{\binom{n-t_j}{t-t_j}}{\binom{n-k}{t-k}}\frac{\mathrm{emb}_F(\sigma)}{\mathrm{emb}_F(\sigma_1)} \le \frac{\binom{n-t_j}{t-t_j}}{\binom{n-k}{t-k}} < \left(\frac{n}{t}\right)^{k-t_j}.$$

It then follows from Definition 2 and (1) that

$$\delta_j < \frac{(n/t)^{k-t_j}}{\tau^{j-1}} < t^t n^{k-t_j+(j-1)/(2\ell)}. \tag{7}$$

Since t_j is increasing with respect to j, $t_2 \ge k+1$, and $j \le \ell$, we have $k - t_j + \frac{j-1}{2\ell} \le -1/2$. Thus, in view of (7), we have

$$\delta_j < t^t n^{k-t_j+(j-1)/(2\ell)} \le t^t n^{-1/2} \tag{8}$$

for all $2 \le j \le \ell$.

Using Definition 2 and inequality (8), we can now bound the codegree function $\delta(H,\tau)$ by

$$\delta(H,\tau) = 2^{\binom{\ell}{2}-1}\sum_{j=2}^{\ell} 2^{-\binom{j-1}{2}}\delta_j \le 2^{\binom{\ell}{2}-1} t^t n^{-1/2}\sum_{j=2}^{\ell} 2^{-\binom{j-1}{2}} \le 2^{\binom{\ell}{2}} t^t n^{-1/2}. \tag{9}$$

Since n satisfies (3), inequality (9) implies that

$$\delta(H,\tau) \le 2^{\binom{\ell}{2}} t^t n^{-1/2} \le \frac{\varepsilon}{12\ell!}.$$

That is, $\delta(H,\tau)$ satisfies the condition in Lemma 3.

Finally, (4) implies that τ satisfies the condition

$$\tau = n^{-1/(2\ell)} < \frac{1}{144\ell!^2\ell}.$$

Therefore, the assumptions of Lemma 3 are met and we can let $\mathcal{C}$ be the collection of subsets from $V(H)$ obtained from applying Lemma 3. Denote the elements of $\mathcal{C}$ by $C_1, C_2, \ldots, C_{|\mathcal{C}|}$.

For every choice of $1 \le a_1, \ldots, a_q \le |\mathcal{C}|$ (not necessarily distinct), let $E_{a_1,\ldots,a_q}$ be the event that $G \subseteq C_{a_1} \cup \cdots \cup C_{a_q}$. Next we will show the following claim.

Claim 2

$$\mathbf{P}(X) \le \mathbf{P}\Bigg(\bigvee_{a_1,\ldots,a_q} E_{a_1,\ldots,a_q} \Bigg) \le \sum_{a_1,\ldots,a_q} \mathbf{P}(E_{a_1,\ldots,a_q}). \tag{10}$$

Proof Suppose that $G \in X$. By definition, there exists a q-coloring of the edges of G, say with colors $1, 2, \ldots, q$, which contains at most M copies of F in each color. For any color class j, let I_j denote the set of vertices of H which correspond to edges of color j in G. Since each edge in $H[I_j]$ corresponds to a copy of F in color j, we have $e(H[I_j]) \le M$. Note that

$$M = \varepsilon \tau^{\ell} e(H),$$

which means that each I_j satisfies the condition (i) of Lemma 3. Therefore, for each color class j, there must be a set $C_{a_j} \in \mathcal{C}$ such that $C_{a_j} \supset I_j$. Since $G = \bigcup_j I_j$, this implies that $G \in E_{a_1,\ldots,a_q}$. Since $G \in X$ was arbitrary, the bound (10) follows and the claim is proved. □

Owing to Claim 2, we now bound $\mathbf{P}(E_{a_1,\ldots,a_q})$. Recalling the definition of the event $E_{a_1,\ldots,a_q}$, we note that

$$\mathbf{P}(E_{a_1,\ldots,a_q}) = (1-p)^{|V(H) \smallsetminus (C_{a_1} \cup \cdots \cup C_{a_q})|}. \tag{11}$$

Hence, we shall estimate $|V(H) \smallsetminus (C_{a_1} \cup \cdots \cup C_{a_q})|$ to derive a bound for $\mathbf{P}(X)$ by (10).

Claim 3 *For all choices* $1 \le a_1, \ldots, a_q \le |\mathcal{C}|$ *we have*

$$|V(H) \smallsetminus (C_{a_1} \cup \cdots \cup C_{a_q})| \ge \frac{1}{2}\left(\frac{n}{R}\right)^k.$$

Proof Let $a_1, \ldots, a_q$ be fixed and set

$$\mathcal{A} = \left\{ A \in \binom{[n]}{R} : \binom{A}{k} \subset C_{a_1} \cup \cdots \cup C_{a_q} \right\}. \tag{12}$$

By the definition of $R = r(F; q)$, for each set $A \in \mathcal{A}$ there is an index $j = j(A) \in [q]$ such that C_{a_j} contains a copy of F with vertices from A. The element $e \in E(C_{a_j})$ that corresponds to this copy of F satisfies $e \subset \binom{A}{k}$ and, thus, $\bigcup_{x \in e} x \subset A$. We now give an upper bound for $|\mathcal{A}|$ by counting the number of pairs in

$$\mathcal{P} = \left\{ (e, A) \in \bigcup_{i=1}^{q} E\big(C_{a_i}\big) \times \mathcal{A} \text{ with } \bigcup_{x \in e} x \subset A \right\}.$$

On the one hand, we have already established that $|\mathcal{P}| \geq |\mathcal{A}|$. On the other hand, for any fixed $e \in E(H)$, we have $|\bigcup_{x\in e} x| = |V(F)| = t$ and, therefore, there are at most $\binom{n-t}{R-t}$ sets $A \supset \bigcup_{x\in e} x$. It follows that

$$\begin{aligned} |\mathcal{A}| \leq |\mathcal{P}| \leq \left| \bigcup_{i=1}^{q} E\big(C_{a_i}\big) \right| \binom{n-t}{R-t} &\overset{(ii)}{\leq} q\varepsilon e(H) \binom{n-t}{R-t} \\ &\overset{(1)}{=} \frac{e(H)}{2R^t} \binom{n-t}{R-t} \leq \frac{(n)_t}{2R^t} \binom{n-t}{R-t} \leq \frac{1}{2} \binom{n}{R}. \end{aligned} \tag{13}$$

By definition, each $A \in \binom{[n]}{R} \setminus \mathcal{A}$ satisfies $\binom{A}{k} \not\subset C_{a_1} \cup \cdots \cup C_{a_q}$. Hence, $V(H) \setminus (C_{a_1} \cup \cdots \cup C_{a_q})$ intersects $\binom{A}{k}$. Since an element of $V(H)$ can appear in at most $\binom{n-k}{R-k}$ sets A, it follows from (13) that there are at least

$$\frac{1}{2}\binom{n}{R} \Big/ \binom{n-k}{R-k} \geq \frac{1}{2}\left(\frac{n}{R}\right)^k$$

elements in $V(H) \setminus (C_{a_1} \cup \cdots \cup C_{a_q})$, as required. □

In view of Claim 3, our choice of $p = 1000R^k q\alpha$, where $\alpha = n^{-1/2\ell + 1/4\ell(\ell+1)}$, and (11), we have, for any $C_{a_1}, \ldots, C_{a_q} \in \mathcal{C}$,

$$\begin{aligned} \mathbf{P}(E_{a_1,\ldots,a_q}) &\leq (1-p)^{(n/R)^k/2} \\ &\leq \exp\big(-pn^k/2R^k\big) = \exp\big(-(1000R^k q\alpha)n^k/2R^k\big) \\ &= e^{-500q\alpha n^k} \leq e^{-1000q\alpha N}, \end{aligned} \tag{14}$$

where, in the last step, we used $N = \binom{n}{k} \leq \frac{n^k}{2}$. Therefore, (10) and (14) together with the bound on $|\mathcal{C}|$ given by Lemma 3(*iii*) imply that

$$\begin{aligned} \mathbf{P}(X) \leq \sum_{C_{a_1},\ldots,C_{a_q} \in \mathcal{C}} \mathbf{P}(E_{a_1,\ldots,a_q}) &\leq |\mathcal{C}|^q e^{-1000q\alpha N} \\ &\leq \exp\big(1000q\ell!^3\ell \log(1/\varepsilon) N\tau \log(1/\tau) - 1000q\alpha N\big) \\ &= \exp\big(1000qN\tau(\ell!^3\ell \log(1/\varepsilon)\log(1/\tau) - \alpha/\tau)\big) \\ &\leq \exp\big(1000qN\tau(\ell!^3 \log^2 n - n^{1/(4\ell(\ell+1))})\big) \leq 1/4, \end{aligned}$$

where we used that n satisfies (5).

Now, by Markov's inequality, with probability at least $1/2$, the number of noninduced copies of F in G will be at most twice the expected number of copies, which is fewer than

$$2p^{\ell+1}\frac{(n)_t}{\mathrm{aut}(F)}\binom{t}{k} = 2(1000q)^{\ell+1}R^{k(\ell+1)}n^{-1/2-1/(4\ell)}\frac{(n)_t}{\mathrm{aut}(F)}\binom{t}{k}$$
$$< \frac{1}{2qR^t}(n^{-1/(2\ell)})^{\ell}\frac{(n)_t}{\mathrm{aut}(F)} = \varepsilon\tau^{\ell}\frac{(n)_t}{\mathrm{aut}(F)} = M,$$

where the inequality above follows from (6). In other words, $\mathbf{P}(\overline{Y}) \geq 1/2$ and, therefore, $\mathbf{P}(\overline{X} \cap \overline{Y}) \geq 1/4$, so there exists a graph G such that $\overline{X} \cap \overline{Y}$ holds. By our earlier observations, this completes the proof.

3 Concluding Remarks

Beginning with Fox and Sudakov [10], much of the recent work on induced Ramsey numbers for graphs has used pseudorandom rather than random graphs for the target graph G. The results of this paper rely very firmly on using random hypergraphs. It would be interesting to know whether comparable bounds could be proved using pseudorandom hypergraphs.

It would also be interesting to prove comparable bounds for the following variant of the induced Ramsey theorem, first proved by Nešetřil and Rödl [15]: for every graph F, there exists a graph G such that every q-coloring of the triangles of G contains an induced copy of F all of whose triangles receive the same color. By taking $F = K_t$ and $q = 4$, we see that $|G|$ may need to be double exponential in $|F|$. We believe that a matching double-exponential upper bound should also hold.

References

1. F.G. Abramson, L.A. Harrington, Models without indiscernibles. J. Symb. Log. **43**(3), 572–600 (1978). doi:10.2307/2273534. MR503795
2. J. Balogh, R. Morris, W. Samotij, Independent sets in hypergraphs. J. Am. Math. Soc. **28**(3), 669–709 (2015). doi:10.1090/S0894-0347-2014-00816-X MR3327533
3. D. Conlon, J. Fox, B. Sudakov, On two problems in graph Ramsey theory. Combinatorica **32**(5), 513–535 (2012). doi:10.1007/s00493-012-2710-3. MR3004807
4. D. Conlon, J. Fox, B. Sudakov, An improved bound for the stepping-up lemma. Discret. Appl. Math. **161**(9), 1191–1196 (2013). doi:10.1016/j.dam.2010.10.013. MR3030610
5. D. Conlon, J. Fox, B. Sudakov, *Recent Developments in Graph Ramsey Theory*. Surveys in Combinatorics 2015, London Mathematical Society Lecture Note Series, vol. 424 (Cambridge University Press, Cambridge, 2015), pp. 49–118. doi:10.1017/CBO9781316106853.003
6. W. Deuber, *Generalizations of Ramsey's Theorem*. Infinite and Finite Sets (Colloquium, Keszthely, 1973; dedicated to P. Erdős on his 60th birthday), vol. I (North-Holland,

Amsterdam, 1975), pp. 323–332. Colloquium Mathematical Society János Bolyai, vol. 10. MR0369127

7. P. Erdős, *On Some Problems in Graph Theory, Combinatorial Analysis and Combinatorial Number Theory*. Graph Theory and Combinatorics (Cambridge, 1983) (Academic, London, 1984), pp. 1–17. MR777160
8. P. Erdős, R. Rado, Combinatorial theorems on classifications of subsets of a given set. Proc. Lond. Math. Soc. (3) **2**, 417–439 (1952). MR0065615
9. P. Erdős, A. Hajnal, L. Pósa, *Strong Embeddings of Graphs Into Colored Graphs*. Infinite and Finite Sets (Colloquium, Keszthely, 1973; Dedicated to P. Erdős on his 60th birthday), vol. I (North-Holland, Amsterdam, 1975), pp. 585–595. Colloquium Mathematical Society János Bolyai, vol. 10. MR0382049
10. J. Fox, B. Sudakov, Induced Ramsey-type theorems. Adv. Math. **219**(6), 1771–1800 (2008). doi:10.1016/j.aim.2008.07.009. MR2455625
11. J. Fox, B. Sudakov, Density theorems for bipartite graphs and related Ramsey-type results. Combinatorica **29**(2), 153–196 (2009). MR2520279
12. R.L. Graham, B.L. Rothschild, J.H. Spencer, *Ramsey Theory*, 2nd edn. Wiley-Interscience Series in Discrete Mathematics and Optimization (Wiley, New York, 1990). A Wiley-Interscience Publication. MR1044995
13. Y. Kohayakawa, H.J. Prömel, V. Rödl, Induced Ramsey numbers. Combinatorica **18**(3), 373–404 (1998). doi:10.1007/PL00009828. MR1721950
14. R. Nenadov, A. Steger, A short proof of the random Ramsey theorem. Combin. Probab. Comput. **25**(1), 130–144 (2016). doi:10.1017/S0963548314000832. MR3438289
15. J. Nešetřil, V. Rödl, *Partitions of Subgraphs*. Recent Advances in Graph Theory (Proceedings of the Second Czechoslovak Symposium, Prague, 1974) (Academia, Prague, 1975), pp. 413–423. MR0429655
16. J. Nešetřil, V. Rödl, Partitions of finite relational and set systems. J. Combin. Theory Ser. A **22**(3), 289–312 (1977). MR0437351
17. F.P. Ramsey, On a problem of formal logic. Proc. Lond. Math. Soc. (2) **30**(1), 264–286 (1930). doi:10.1112/plms/s2-30.1.264
18. V. Rödl, The dimension of a graph and generalized Ramsey theorems. Master's thesis, Charles University (1973)
19. V. Rödl, A. Ruciński, M. Schacht, An exponential-type upper bound for Folkman numbers. Combinatorica. doi:10.1007/s00493-015-3298-1. To appear
20. D. Saxton, A. Thomason, Hypergraph containers. Invent. Math. **201**(3), 925–992 (2015). doi:10.1007/s00222-014-0562-8. MR3385638

ARRIVAL: A Zero-Player Graph Game in NP ∩ coNP

Jérôme Dohrau, Bernd Gärtner, Manuel Kohler, Jiří Matoušek, and Emo Welzl

Abstract Suppose that a train is running along a railway network, starting from a designated origin, with the goal of reaching a designated destination. The network, however, is of a special nature: every time the train traverses a switch, the switch will change its position immediately afterwards. Hence, the next time the train traverses the same switch, the other direction will be taken, so that directions alternate with each traversal of the switch.

Given a network with origin and destination, what is the complexity of deciding whether the train, starting at the origin, will eventually reach the destination?

It is easy to see that this problem can be solved in exponential time, but we are not aware of any polynomial-time method. In this short paper, we prove that the problem is in NP ∩ coNP. This raises the question whether we have just failed to find a (simple) polynomial-time solution, or whether the complexity status is more subtle, as for some other well-known (two-player) graph games (Halman, Algorithmica 49(1):37–50, 2007).

1 Introduction

In this paper, a *switch graph* is a directed graph G in which every vertex has at most two outgoing edges, pointing to its *even* and to its *odd* successor. Formally, a switch graph is a 4-tuple $G = (V, E, s_0, s_1)$, where $s_0, s_1 : V \to V$, $E = \{(v, s_0(v)) :$

This research was done in the 2014 undergraduate seminar *Wie funktioniert Forschung*? (How does research work?).

J. Dohrau • B. Gärtner (✉) • M. Kohler • E. Welzl
Department of Computer Science, Institute of Theoretical Computer Science, ETH Zürich,
CH-8092 Zürich, Switzerland
e-mail: gaertner@inf.ethz.ch

J. Matoušek
Department of Computer Science, Institute of Theoretical Computer Science, ETH Zürich,
CH-8092 Zürich, Switzerland

Department of Applied Mathematics, Charles University, Malostranské nám. 25, 118 00 Prague,
Czech Republic

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_14

$v \in V\} \cup \{(v, s_1(v)) : v \in V\}$, with loops (v, v) allowed. Here, $s_0(v)$ is the even successor of v, and $s_1(v)$ the odd successor. We may have $s_0(v) = s_1(v)$ in which case v has just one outgoing edge. We always let $n = |V|$; for $v \in V$, $E^+(v)$ denotes the set of outgoing edges at v, while $E^-(v)$ is the set of incoming edges.

Given a switch graph $G = (V, E, s_0, s_1)$ with origin and destination $o, d \in V$, the following procedure describes the train run that we want to analyze; our problem is to decide whether the procedure terminates. For the procedure, we assume arrays `s_curr` and `s_next`, indexed by V, such that initially `s_curr`$[v] = s_0(v)$ and `s_next`$[v] = s_1(v)$ for all $v \in V$.

```
procedure Run($G, o, d$)
  $v := o$
  while $v \neq d$ do
    $w :=$ s_curr[$v$]
    swap (s_curr[$v$], s_next[$v$])
    $v := w$                                    ▷ traverse edge $(v, w)$
  end while
end procedure
```

Definition 1 Problem *ARRIVAL* is to decide whether procedure Run(G, o, d) terminates for a given switch graph $G = (V, E, s_0, s_1)$ and $o, d \in V$.

Theorem 1 *Problem* ARRIVAL *is decidable.*

Proof The deterministic procedure Run can be interpreted as a function that maps the current *state* (v, `s_curr`, `s_next`) to the next state. We can think of the state as the current location of the train, and the current positions of all the switches. As at most $n2^n$ different states may occur, Run either terminates within this many iterations, or some state repeats, in which case Run enters an infinite loop. Hence, to decide ARRIVAL, we have to go through at most $n2^n$ iterations of Run. □

Figure 1 shows that a terminating run may indeed take exponential time.

Existing research on switch graphs (with the above, or similar definitions) has mostly focused on actively controlling the switches, with the goal of attaining some desired behavior of the network (e.g. reachability of the destination); see e.g. [5]. The question we address here rather fits into the theory of cellular automata. It is motivated by the online game *Looping Piggy* (`https://scratch.mit.edu/projects/1200078/`) that the second author has written for the *Kinderlabor*, a

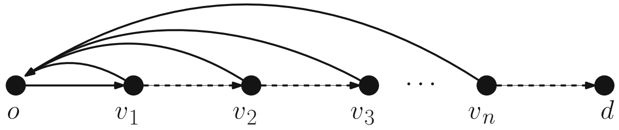


Fig. 1 Switch graph G with $n+2$ vertices on which Run(G, o, d) traverses an exponential number of edges. If we encode the current positions of the switches at $v_n, \ldots, v_1$ with an n-bit binary number (0: even successor is next; 1: odd successor is next), then the run counts from 0 to $2^n - 1$, resets the counter to 0, and terminates. *Solid edges* point to even or unique successors, *dashed edges* to odd successors

Swiss-based initiative to educate children at the ages 4–12 in natural sciences and computer science (http://kinderlabor.ch).

In Sects. 2 and 3, we prove that ARRIVAL is in NP as well as in coNP; Sect. 4 shows that a terminating run can be interpreted as the unique solution of a flow-type integer program with balancing conditions whose LP relaxation may have only fractional optimal solutions.

2 ARRIVAL Is in NP

A natural candidate for an NP-certificate is the *run profile* of a terminating run. The run profile assigns to each edge the number of times it has been traversed during the run. The main difficulty is to show that fake run profiles cannot fool the verifier. We start with a necessary condition for a run profile: it has to be a *switching flow*.

Definition 2 Let $G = (V, E, s_0, s_1)$ be a switch graph, and let $o, d \in V$, $o \neq d$. A *switching flow* is a function $\mathbf{x} : E \to \mathbb{N}_0$ (where $\mathbf{x}(e)$ is denoted as x_e) such that the following two conditions hold for all $v \in V$.

$$\sum_{e \in E^+(v)} x_e - \sum_{e \in E^-(v)} x_e = \begin{cases} 1, & v = o, \\ -1, & v = d, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

$$0 \leq x_{(v,s_1(v))} \leq x_{(v,s_0(v))} \leq x_{(v,s_1(v))} + 1. \tag{2}$$

Observation 1 *Let $G = (V, E, s_0, s_1)$ be a switch graph, and let $o, d \in V$, $o \neq d$, such that* Run(G, o, d) *terminates. Let $\mathbf{x}(G, o, d) : E \to \mathbb{N}_0$ (the run profile) be the function that assigns to each edge the number of times it has been traversed during* Run(G, o, d). *Then $\mathbf{x}(G, o, d)$ is a switching flow.*

Proof Condition (1) is simply flow conservation (if the run enters a vertex, it has to leave it, except at o and d), while (2) follows from the run alternating between successors at any vertex v, with the even successor $s_0(v)$ being first. □

While every run profile is a switching flow, the converse is not always true. Figure 2 shows two switching flows for the same switch graph, but only one of them

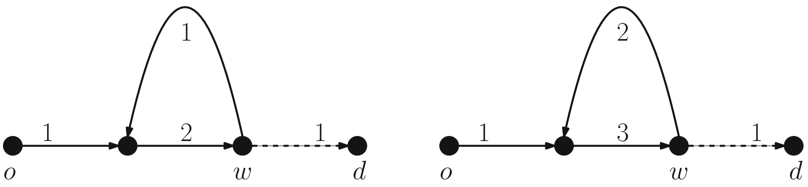


Fig. 2 Run profile (*left*) and fake run profile (*right*); both are switching flows. *Solid edges* point to even or unique successors, *dashed edges* to odd successors

is the actual run profile. The "fake" run results from going to the even successor of w twice in a row, before going to the odd successor d. This shows that the balancing condition (2) fails to capture the strict alternation between even and odd successors. Despite this, and maybe surprisingly, the existence of a switching flow implies termination of the run.

Lemma 1 *Let $G = (V, E, s_0, s_1)$ be a switch graph, and let $o, d \in V$, $o \neq d$. If there exists a switching flow* $\mathbf{x}$*, then* $\textsc{Run}(G, o, d)$ *terminates, and* $\mathbf{x}(G, o, d) \leq \mathbf{x}$ *(componentwise).*

Proof We imagine that for all $e \in E$ we put x_e pebbles on edge e, and then start $\textsc{Run}(G, o, d)$. Every time an edge is traversed, we let the run collect one pebble. The claim is that we never run out of pebbles, which proves termination as well as the inequality for the run profile.

To prove the claim, we first observe two invariants: during the run, flow conservation (w.r.t. to the remaining pebbles) always holds, except at d, and at the current vertex which has one more pebble on its outgoing edges. Moreover, by alternation, starting with the even successor, the numbers of pebbles on $(v, s_0(v))$ and $(v, s_1(v))$ always differ by at most one, for every vertex v.

For contradiction, consider now the first iteration of $\textsc{Run}(G, o, d)$ where we run out of pebbles, and let $e = (v, w)$ be the edge (now holding -1 pebbles) traversed in the offending iteration. By the above alternation invariant, the other outgoing edge at v cannot have any pebbles left, either. Then the flow conservation invariant at v shows that already some incoming edge of v has a deficit of pebbles, so we have run out of pebbles before, which is a contradiction. □

Theorem 2 *Problem* ARRIVAL *is in* NP.

Proof Given an instance (G, o, d), the verifier receives a function $\mathbf{x} : E \to \mathbb{N}_0$, in form of binary encodings of the values x_e, and checks whether it is a switching flow. For a Yes-instance, the run profile of $\textsc{Run}(G, o, d)$ is a witness by Observation 1; the proof of Theorem 1 implies that the verification can be made to run in polynomial time, since every value x_e is bounded by $n2^n$. For a No-instance, the check will fail by Lemma 1. □

3 ARRIVAL Is in coNP

Given an instance (G, o, d) of ARRIVAL, the main idea is to construct in polynomial time an instance $(\bar{G}, o, \bar{d})$ such that $\textsc{Run}(G, o, d)$ terminates if and only if $\textsc{Run}(\bar{G}, o, \bar{d})$ does not terminate. As the main technical tool, we prove that nontermination is equivalent to the arrival at a "dead end".

Definition 3 Let $G = (V, E, s_0, s_1)$ be a switch graph, and let $o, d \in V$, $o \neq d$. A *dead end* is a vertex from which there is no directed path to the destination d in the graph (V, E). A *dead edge* is an edge $e = (v, w)$ whose head w is a dead end.

An edge that is not dead is called *hopeful*; the length of the shortest directed path from its head w to d is called its *desperation*.

By computing the tree of shortest paths to d, using inverse breadth-first search from d, we can identify the dead ends in polynomial time. Obviously, if $\textsc{Run}(G, o, d)$ ever reaches a dead end, it will not terminate, but the converse is also true. For this, we need one auxiliary result.

Lemma 2 *Let* $G = (V, E, s_0, s_1)$ *be a switch graph,* $o, d \in V$, $o \neq d$, *and let* $e = (v, w) \in E$ *be a hopeful edge of desperation k. Then* $\textsc{Run}(G, o, d)$ *will traverse e at most* $2^{k+1} - 1$ *times.*

Proof Induction on the desperation k of $e = (v, w)$. If $k = 0$, then $w = d$, and indeed, the run will traverse e at most $2^1 - 1 = 1$ times. Now suppose $k > 0$ and assume that the statement is true for all hopeful edges of desperation $k - 1$. In particular, one of the two successor edges $(w, s_0(w))$ and $(w, s_1(w))$ is such a hopeful edge, and is therefore traversed at most $2^k - 1$ times. By alternation at w, the other successor edge is traversed at most once more, hence at most 2^k times. By flow conservation, the edges entering w (in particular e) can be traversed at most $2^k + 2^k - 1 = 2^{k+1} - 1$ times. □

Lemma 3 *Let* $G = (V, E, s_0, s_1)$ *be a switch graph, and let* $o, d \in V$, $o \neq d$. *If* $\textsc{Run}(G, o, d)$ *does not terminate, it will reach a dead end.*

Proof By Lemma 2, hopeful edges can be traversed only finitely many times, hence if the run cycles, it eventually has to traverse a dead edge and thus reach a dead end. □

Now we can prove the main result of this section.

Theorem 3 *Problem* ARRIVAL *is in* coNP.

Proof Let (G, o, d) be an instance, $G = (V, E, s_0, s_1)$. We transform (G, o, d) into a new instance $(\bar{G}, o, \bar{d})$, $\bar{G} = (\bar{V}, \bar{E}, \bar{s}_0, \bar{s}_1)$ as follows. We set $\bar{V} = V \cup \{\bar{d}\}$, where $\bar{d}$ is an additional vertex, the new destination. We define $\bar{s}_0, \bar{s}_1$ as follows. For every dead end w, we set

$$\bar{s}_0(w) = \bar{s}_1(w) := \bar{d}. \tag{3}$$

For the old destination d, we install the loop

$$\bar{s}_0(d) = \bar{s}_1(d) := d. \tag{4}$$

For the new destination, $\bar{s}_0(\bar{d})$ and $\bar{s}_1(\bar{d})$ are chosen arbitrarily. In all other cases, $\bar{s}_0(v) := s_0(v)$ and $\bar{s}_1(v) := s_1(v)$. This defines $\bar{E}$ and hence $\bar{G}$.

The crucial properties of this construction are the following:

(i) If $\textsc{Run}(G, o, d)$ reaches the destination d, it has not visited any dead ends, hence s_0 and $\bar{s}_0$ as well as s_1 and $\bar{s}_1$ agree on all visited vertices except d. This means that $\textsc{Run}(\bar{G}, o, \bar{d})$ will also reach d, but then cycle due to the loop that we have installed in (4).

(ii) If $\textsc{Run}(G, o, d)$ cycles, it will at some point reach a first dead end w, by Lemma 3. As s_0 and $\bar{s}_0$ as well as s_1 and $\bar{s}_1$ agree on all previously visited vertices, $\textsc{Run}(\bar{G}, o, \bar{d})$ will also reach w, but then terminate due to the edges from w to $\bar{d}$ that we have installed in (3).

To summarize, $\textsc{Run}(G, o, d)$ terminates if and only if $\textsc{Run}(\bar{G}, o, \bar{d})$ does not terminate. Since $(\bar{G}, o, \bar{d})$ can be constructed in polynomial time, we can verify in polynomial time that (G, o, d) is a No-instance by verifying that $(\bar{G}, o, \bar{d})$ is a Yes-Instance via Theorem 2. □

4 Is ARRIVAL in P?

Observation 1 and Lemma 1 show that ARRIVAL can be decided by checking the solvability of a system of linear (in)equalities (1) and (2) over the nonnegative integers.

The latter is an NP-complete problem in general: many of the standard NP-complete problems, e.g. SAT (satisfiability of boolean formulas) can easily be reduced to finding an integral vector that satisfies a system of linear (in)equalities.

In our case, we have a flow structure, though, and finding integral flows in a network is a well-studied and easy problem [6, Chapter 8]. In particular, if only the flow conservation constraints (1) are taken into account, the existence of a nonnegative integral solution is equivalent to the existence of a nonnegative real solution. This follows from the classical *Integral Flow Theorem*, see [6, Corollary 8.7]. Real solutions to systems of linear (in)equalities can be found in polynomial time through linear programming [6, Chapter 4].

However, the additional balancing constraints (2) induced by alternation at the switches, make the situation more complicated. Figure 3 depicts an instance which has a real-valued "switching flow" satisfying constraints (1) and (2), but no integral one (since the run does not terminate).

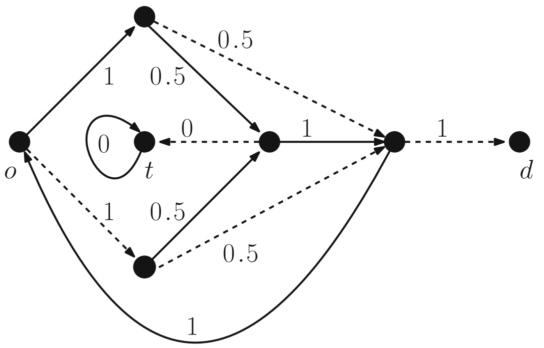


Fig. 3 The run will enter the loop at t and cycle, so there is no (integral) switching flow. But a real-valued "switching flow" (given by the numbers) exists. *Solid edges point* to even or unique successors, *dashed edges* to odd successors

We conclude with a result that summarizes the situation and may be the basis for further investigations.

Theorem 4 *Let $G = (V, E, s_0, s_1)$ be a switch graph, and let $o, d \in V$, $o \neq d$. $\mathrm{RUN}(G, o, d)$ terminates if and only if there exists an integral solution satisfying the constraints* (1) *and* (2)*. In this case, the run profile $\mathbf{x}(G, o, d)$ is the unique integral solution that minimizes the linear objective function $\Sigma(\mathbf{x}) = \sum_{e \in E} x_e$ subject to the constraints* (1) *and* (2)*.*

Proof Observation 1 and Lemma 1 show the equivalence between termination and existence of an integral solution (a switching flow). Suppose that the run terminates with run profile $\mathbf{x}(G, o, d)$. We have $\mathbf{x}(G, o, d) \leq \mathbf{x}$ for every switching flow $\mathbf{x}$, by Lemma 1. In particular, $\Sigma(\mathbf{x}(G, o, d)) \leq \Sigma(\mathbf{x})$, so the run profile has minimum value among all switching flows. A different switching flow $\mathbf{x}$ of the same value would have to be smaller in at least one coordinate, contradicting $\mathbf{x}(G, o, d) \leq \mathbf{x}$. □

Theorem 4 shows that the existence of $\mathbf{x}(G, o, d)$ and its value can be established by solving an integer program [6, Chapter 5]. Moreover, this integer program is of a special kind: its unique optimal solution is at the same time a least element w.r.t. the partial order "≤" over the set of feasible solutions.

5 Conclusion

The main question left open is whether the zero-player graph game ARRIVAL is in P. There are three well-known two-player graph games in NP ∩ coNP for which membership in P is also not established: *simple stochastic games*, *parity games*, and *mean-payoff games*. All three are even in UP ∩ coUP, meaning that there exist efficient verifiers for Yes- and No-instances that accept *unique* certificates [1, 3]. In all three cases, the way to prove this is to assign payoffs to the vertices in such a way that they form a certificate if and only if they solve a system of equations with a unique solution.

It is natural to ask whether also ARRIVAL is in UP ∩ coUP. We do not know the answer. The natural approach suggested by Theorem 4 is to come up with a verifier that does not accept just any switching flow, but only the unique one of minimum norm corresponding to the run profile. However, verifying optimality of a feasible integer program solution is hard in general, so for this approach to work, one would have to exploit specific structure of the integer program at hand. We do not know how to do this.

As problems in NP ∩ coNP cannot be NP-hard (unless NP and coNP collapse), other concepts of hardness could be considered for ARRIVAL. As a first step in this direction, Karthik [4] has shown that a natural search version of ARRIVAL is contained in the complexity class PLS (Polynomial Local Search) which has complete problems not known to be solvable in polynomial time. PLS-hardness of

ARRIVAL would not contradict common complexity theoretic beliefs; establishing such a hardness result would at least provide a satisfactory explanation why we have not been able to find a polynomial-time algorithm for ARRIVAL.

Acknowledgements We thank the referees for valuable comments and Rico Zenklusen for constructive discussions.

References

1. A. Condon, The complexity of stochastic games. Inf. Comput. **96**(2), 203–224 (1992)
2. N. Halman, Simple stochastic games, parity games, mean payoff games and discounted payoff games are all LP-type problems. Algorithmica **49**(1), 37–50 (2007)
3. M. Jurdziński, Deciding the winner in parity games is in UP ∩ co-UP. Inf. Process. Lett. **68**(3), 119–124 (1998)
4. C.S. Karthik, Did the train reach its destination: the complexity of finding a witness (2016). https//arxiv.org/abs/1609.03840
5. B. Katz, I. Rutter, G. Woeginger, An algorithmic study of switch graphs. Acta Inf. **49**(5), 295–312 (2012)
6. B. Korte, J. Vygen, *Combinatorial Optimization: Theory and Algorithms*, 5th edn. (Springer, Berlin, 2012)

Constant-Factor Approximation for TSP with Disks

Adrian Dumitrescu and Csaba D. Tóth

Abstract We revisit the traveling salesman problem with neighborhoods (TSPN) and present the first constant-ratio approximation for disks in the plane: Given a set of n disks in the plane, a TSP tour whose length is at most $O(1)$ times the optimal can be computed in time that is polynomial in n. Our result is the first constant-ratio approximation for a class of planar convex bodies of arbitrary size and arbitrary intersections.

In order to achieve a $O(1)$-approximation, we reduce the traveling salesman problem with disks, up to constant factors, to a minimum weight hitting set problem in a geometric hypergraph. The connection between TSPN and hitting sets in geometric hypergraphs, established here, is likely to have future applications.

1 Introduction

In the Euclidean Traveling Salesman Problem (ETSP), given a set of points in the Euclidean space $\mathbb{R}^d$, $d \geq 2$, one seeks a shortest closed curve (a.k.a. *tour*) that visits each point. In the TSP *with neighborhoods* (TSPN), each point is replaced by a point-set, called *region* or *neighborhood*, and the TSP tour must visit at least one point in each region, i.e., it must intersect each region. The oldest record that we could trace of this variant goes back to Arkin and Hassin [2]. Since the Euclidean TSP is known to be NP-hard in $\mathbb{R}^d$ for every $d \geq 2$ [21, 22, 40], TSPN is also NP-hard for every $d \geq 2$. TSP is recognized as one of the corner-stone

Research by this author was supported in part by the NSF award CCF-1423615.

A. Dumitrescu
Department of Computer Science, University of Wisconsin–Milwaukee, Milwaukee, WI, USA
e-mail: dumitres@uwm.edu

C.D. Tóth (✉)
Department of Mathematics, California State University, Northridge, Los Angeles, CA, USA

Department of Computer Science, Tufts University, Medford, MA, USA
e-mail: cdtoth@acm.org

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_15

problems in combinatorial optimization. Other related problems in geometric network optimization can be found in the two surveys by Mitchell [32, 33].

It is known that the Euclidean TSP admits a polynomial-time approximation scheme in $\mathbb{R}^d$, where $d = O(1)$, due to classic results of Arora [3] and Mitchell [31]. Subsequent running time improvements have been obtained by Rao and Smith [41]; specifically, the running time of their PTAS is $O(f(\varepsilon)\, n \log n)$, where $f(\varepsilon)$ grows exponentially in $1/\varepsilon$. In contrast, TSPN is generally harder to approximate. Typically, somewhat better approximations are available when the neighborhoods are *pairwise disjoint*, or *fat*, or have *comparable sizes*. We briefly review some of the previous work concerning approximation algorithms for TSPN.

Related work Arkin and Hassin [2] gave constant-factor approximations for translates of a connected region, and more generally, for neighborhoods of pairwise *parallel* diameters, where the ratio between the longest and the shortest diameter is bounded by a constant. Dumitrescu and Mitchell [16] extended the above result to connected neighborhoods with comparable diameters. Bodlaender et al. [8] described a PTAS for TSPN with disjoint fat neighborhoods of about the same size in $\mathbb{R}^d$, where d is constant (this includes the case of disjoint unit disks in the plane). Earlier Dumitrescu and Mitchell [16] proposed a PTAS for TSPN with fat neighborhoods of about the same size and bounded depth in the plane, where Spirkl [44] recently reported and filled a gap; see also a follow-up note in [36].

Mata and Mitchell [29] gave a $O(\log n)$-approximation for TSPN with n connected and arbitrarily intersecting neighborhoods in the plane; see also [7]. Elbassioni et al. [19] and Gudmundsson and Levcopoulos [24] improved the running time of the algorithm. The $O(\log n)$-approximation relies on the following early result by Levcopoulos and Lingas [28]: Every (simple) rectilinear polygon P with n vertices, r of which are reflex, can be partitioned in $O(n \log n)$ time into rectangles whose total perimeter is $\log r$ times the perimeter of P.

Using an approximation algorithm due to Slavik [43] for Euclidean group TSP, de Berg et al. [6] obtained constant-factor approximations for disjoint fat convex regions (of arbitrary diameters) in the plane. Subsequently, Elbassioni et al. [19] gave constant-factor approximations for arbitrarily intersecting fat convex regions of comparable size. Preliminary work by Mitchell [34] gave a PTAS for planar regions of bounded depth and arbitrary size, in particular for disjoint fat regions. Chan and Jiang [12] gave a PTAS for fat, weakly disjoint regions in metric spaces of constant doubling dimension (combining an earlier QPTAS by Chan and Elbassioni [11] with a PTAS for TSP in doubling metrics by Bartal et al. [4]).

Disks and balls are undoubtedly among the simplest neighborhood types [2, 16, 25]. TSPN for disks is NP-hard, and it remains so for congruent disks, since when the disk centers are fixed and the radius tends to zero, the problem reduces to TSP for points. Regarding approximations, the case of congruent balls is relatively well understood: Given a set of n congruent (say, unit) balls in $\mathbb{R}^d$, a TSP tour whose length is at most $O(1)$ times the optimal can be computed in polynomial time, when d is constant [18]. However, for disks of arbitrary radii and intersections, no constant-ratio approximation was known. Some of the difficulties with disks of arbitrary radii in the plane where uncovered in [17].

Recent work of Dumitrescu and Tóth [18] focused on unbounded neighborhoods, such as lines or hyperplanes: They gave a constant-factor approximation for TSPN with n hyperplanes in $\mathbb{R}^d$ in $O(n)$ time; and a $O(\log^3 n)$-approximation for n lines in $\mathbb{R}^d$ in time polynomial in n, where d is constant. In contrast, the current paper considers TSPN with arbitrary disks in $\mathbb{R}^2$, which requires quite different approximation techniques and new ideas.

Degree of approximation Regarding the degree of approximation achievable, TSPN with arbitrary neighborhoods is generally APX-hard [6, 15, 42], and it remains so even for segments of nearly the same length [19]. For disconnected neighborhoods, TSPN cannot be approximated within any constant ratio unless $P = NP$ [42]. Further, approximating TSPN for (arbitrary) connected neighborhoods in the plane within a factor smaller than 2 is NP-hard [42]. Delineating the class of neighborhoods for which constant-factor approximations are possible remains mysterious, at least at the moment. It is conjectured that approximating TSPN for disconnected regions in the plane within a $O(\log^{1/2} n)$ factor is intractable unless $P = NP$ [42]. Similarly, it is conjectured that approximating TSPN for connected regions in $\mathbb{R}^3$ within a $O(\log^{1/2} n)$ factor and for disconnected regions in $\mathbb{R}^3$ within a $O(\log^{2/3} n)$ factor [42] are probably intractable.

Our results In this paper we present a polynomial-time (deterministic) algorithm that, given a set of n disks in $\mathbb{R}^2$ (with arbitrary radii and intersections), returns a TSP tour whose length is $O(1)$ times the optimal.

Theorem 1 *Given a set of n disks in the plane, a TSP tour whose length is at most $O(1)$ times the optimal can be computed in time polynomial in n.*

In their seminal paper on TSPN, Arkin and Hassin [2] suggested disks as the most natural type of neighborhood—in which the traveling salesman can meet each potential buyer at some point close to the respective buyer's location. Here the radius of each disk indicates how much each potential buyer is willing to travel to meet the salesman. While constant-ratio approximations for disks of the same (or comparable) radius [2, 16] and for disjoint disks of arbitrary radii [6] have been obtained early on, the case of disks with arbitrary radii and arbitrary intersections has remained open until now.

2 Preliminaries

We achieve a $O(1)$-approximation for TSP with disks by reducing the problem, up to constant factors, to a minimum weight hitting set problem in a geometric hypergraph, for which a constant-factor approximation algorithm was found only recently [13].

Hitting sets Hitting sets are defined in general in terms of hypergraphs (i.e., set systems or range spaces). A hypergraph is a pair $\mathcal{G} = (V, E)$ where V is a finite vertex set and $E \subset 2^V$ is a finite collection of subsets of V (called edges). In a geometric (primal) hypergraph, the vertex set V is a finite set of n points in Euclidean space $\mathbb{R}^d$, and all sets in E are of the form $V \cap Q$ where Q is a certain geometric shape of bounded description complexity, e.g., halfspace, ball, triangle, axis-aligned rectangle, etc. Geometric hypergraphs often have nice properties such as bounded VC-dimension or bounded union complexity; see [20, 39].

A *hitting set* in a hypergraph $\mathcal{G} = (V, E)$ is a subset of vertices $H \subseteq V$ such that every hyperedge in E contains some point in H. The *minimum hitting set* (MHS) problem asks for a hitting set of minimum cardinality in a given hypergraph. The *minimum weight hitting set* (MWHS) problem asks for a hitting set of minimum weight in a given hypergraph with vertex weights $w : V \to \mathbb{R}^+$.

Brönnimann and Goodrich [9] gave a $O(\log \mathsf{OPT})$-approximation for MHS in geometric hypergraphs using LP-relaxations and the fact that geometric hypergraphs have bounded VC-dimension. Clarkson and Varadarajan [14] gave a $O(\log\log n)$-approximation for *some* geometric hypergraphs, by observing a connection between hitting sets and the combinatorial complexity of the union of the corresponding geometric objects. Mustafa and Ray [38] gave a PTAS for MHS with disks and pseudo-disks in the plane using a local search paradigm; see also [1, 10]. However, this method does not seem to extend to the weighted version (MWHS).

Varadarajan [45] gave a $O(\log\log n)$-approximation for MWHS, extending the results from [14]. His approach was further extended by Chan et al. [13] who obtained a randomized polynomial-time $O(1)$-approximation algorithm for MWHS in geometric hypergraphs of linear union complexity, including geometric hypergraphs defined by disks in $\mathbb{R}^2$ [26]; their algorithm can be derandomized [13, Section 3]. Specifically, we use the following result due to Chan et al. [13].

Theorem 2 ([13, Corollary 1.4 and Section 3]) *There is a polynomial-time (deterministic) $O(1)$-approximation algorithm for the minimum weight hitting set problem for disks in $\mathbb{R}^2$.*

Definitions Let $\mathcal{R}$ be a set of regions (neighborhoods) in $\mathbb{R}^2$. An optimal TSP tour for $\mathcal{R}$, denoted by $\mathsf{OPT}(\mathcal{R})$, is a shortest closed curve in the plane that intersects every region in $\mathcal{R}$; when $\mathcal{R}$ is clear from the context, $\mathsf{OPT}(\mathcal{R})$ and OPT are used interchangeably.

The Euclidean length of a curve γ is denoted by $\mathrm{len}(\gamma)$. Similarly, the total (Euclidean) length of the edges of a geometric graph G is denoted by $\mathrm{len}(G)$. The perimeter of a polygon P is denoted by $\mathrm{per}(P)$; the boundary and the interior of a region R are denoted by ∂R and R°, respectively; the convex hull of a planar set S is denoted by $\mathrm{conv}(S)$.

The distance between two planar point sets $S_1, S_2 \subset \mathbb{R}^2$, is $\mathrm{dist}(S_1, S_2) = \inf\{\mathrm{dist}(s_1, s_2) : s_1 \in S_1, s_2 \in S_2\}$. The distance between a point set S_1 and a geometric graph G is defined as $\mathrm{dist}(S_1, G) := \mathrm{dist}(S_1, S_G)$, where S_G is the set of all points at vertices and on the edges of G.

Algorithm Outline Given a set S of n disks in Euclidean plane, we construct a connected geometric graph G that intersects every disk in S and such that $\text{len}(G) = O(\text{len}(\mathsf{OPT}))$. An Eulerian tour of the multi-graph obtained by doubling each edge of G visits each disk and its length is $2\,\text{len}(G) = O(\text{len}(\mathsf{OPT}))$, as desired.

The graph G is the union of three geometric graphs, G_1, G_2 and G_3. The graph G_1 is a $O(1)$-approximation of an optimal tour for a maximal subset of pairwise disjoint disks in S; this step is based on earlier results [6, 19] (Sect. 3). The graph G_2 connects G_1 to nearby disks that are guaranteed to be at distance at most $\text{len}(\mathsf{OPT})/n$ from G_1. The graph G_3 connects any remaining disks to G_1; this step is based on recent results on minimum weight hitting sets due to Chan et al. [13] (Sect. 5).

The interface between TSP and the hitting set problem is established by a quadtree subdivision [5, Ch. 14]. Previously, Arora [3] and Mitchell [35] used quadtrees for approximating Euclidean TSP and TSP with disjoint neighborhoods, respectively. The quadtree variety that we need, a so-called *stratified grid*, was introduced by Mitchell [35] for certain orthogonal polygons. Here we define stratified grids in a more general setting, for arbitrary geometric graphs (Sect. 4).

3 Preprocessing

Let S be a set of n disks in the plane. The algorithm first constructs the graphs G_1 and G_2 as follows (Fig. 1).

1. Select an *independent subset* I, $I \subseteq S$, of pairwise disjoint disks by the following greedy algorithm: Set $I := \emptyset$. Consider the disks in S in increasing order of radius (with ties broken arbitrarily), and successively place a disk $D \in S$ into I if it is disjoint from all previous disks in I.
2. Compute a constant-factor approximate TSP tour ξ_0 for I using the algorithm in [6] or [19]. (A PTAS for disjoint disks in the plane is available [16, 44] but not needed here.) It is clear that $\text{len}(\xi_0) = O(\text{len}(\mathsf{OPT}))$.

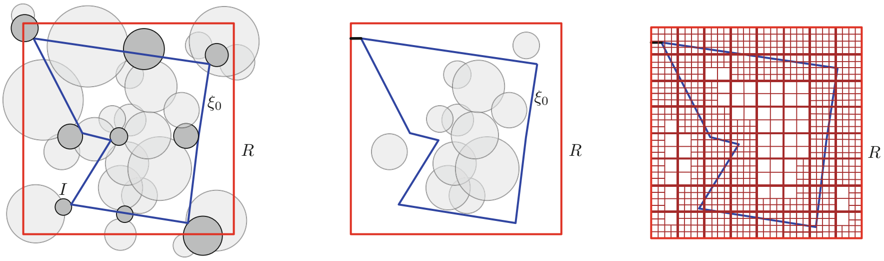


Fig. 1 *Left*: a set of disks in the plane; the independent set I selected by a greedy algorithm is highlighted; a TSP tour ξ_0 for I, and a minimum square R intersecting all disks are shown. *Middle*: G_1 is the union of ξ_0, R, and a shortest line segment connecting them; only the disks in $S_2 \cup S_3$ that do not intersect G_1 are shown. *Right*: a stratified grid for R and G_1

3. Let R be a minimum axis-parallel square such that $\mathrm{conv}(R)$ intersects every disk in S (i.e., every disk intersects ∂R or is contained in R°). The square R is determined by up to 3 disks in S, thus R can be trivially computed in $O(n^4)$ time: there are $O(n^3)$ squares that pairs and triples define, and each can be checked in $O(n)$ time as to whether it intersects all disks. Alternatively, finding R is an LP-type problem of combinatorial dimension 3 that can be solved in $O(n)$ time [30, Section 5]; see also [23] for a modern treatment of LP-type problems and violator spaces. Let r denote its side-length of R; obviously, we have $r \leq \mathrm{len}(\mathsf{OPT})$.
4. Let G_1 be the union of ξ_0, R, and a shortest line segment connecting ξ_0 and R (if disjoint).

The graph G_1 intersects all disks in I and possibly some disks in $S \backslash I$. Our primary interest is in the disks in S that are disjoint from G_1.

Lemma 1 *For every disk $D \in S$, we have* $\mathrm{dist}(D, G_1) \leq \mathrm{diam}(D)$.

Proof Let $D \in S$. If G_1 intersects D, then $\mathrm{dist}(D, G_1) = 0$, and the claim is trivial. Assume that D is disjoint from G_1. Since ξ_0 intersects every disk in I and $\xi_0 \subseteq G_1$, we have $D \in S \setminus I$. By the greedy choice of $I \subseteq S$, the disk D intersects some disk $D' \in I$ of equal or smaller radius, where G_1 intersects D'. Consequently, $\mathrm{dist}(D, G_1) \leq \mathrm{diam}(D') \leq \mathrm{diam}(D)$. □

Connecting nearby disks to G_1 We partition S into three subsets: let S_1 be the set of disks in S that intersect G_1; let S_2 be the set of disks $D \in S \setminus S_1$ such that $\mathrm{dist}(D, G_1) \leq \frac{r}{n}$; and let $S_3 = S \backslash (S_1 \cup S_2)$. Let G_2 be a graph that consists of $|S_2|$ line segments: specifically for every $D \in S_2$, G_2 contains a shortest segment connecting D and G_1. Then $\mathrm{len}(G_2) = \sum_{D \in S_2} \mathrm{dist}(D, G_1) \leq |S_2| \cdot \frac{r}{n} \leq r \leq \mathrm{len}(\mathsf{OPT})$. By construction, we have

$$\mathrm{dist}(D, G_1) > \frac{r}{n} \text{ for every } D \in S_3. \tag{1}$$

By Lemma 1 and inequality (1) we have

Corollary 1 *For every disk $D \in S_3$, we have* $\mathrm{diam}(D) > \frac{r}{n}$.

In the next section, we show how to find a geometric graph G_3 such that G_3 intersects every disk in S_3 and $G_1 \cup G_3$ is connected (note, however, that G_3 need not be connected).

4 Stratified Grids

Recall that we have a geometric graph G_1, and a set S_3 of at most n disks in the interior of an axis-aligned square R of side-length r, $r \leq \mathrm{len}(\mathsf{OPT})$, satisfying (1). Let $\mathsf{OPT}(S_3, G_1)$ denote a geometric graph Γ of minimum length such that $G_1 \cup \Gamma$ is connected and intersects every disk in S_3. Note that $\mathrm{len}(\mathsf{OPT}(S_3, G_1)) \leq \mathrm{len}(\mathsf{OPT}(S_3)) \leq \mathrm{len}(\mathsf{OPT})$, for every G_1.

In Sects. 5 and 6, we use hitting sets to compute a $O(1)$-approximation of $\mathsf{OPT}(S_3, G_1)$. Similarly to a quadtree decomposition, we recursively construct a subdivision of R into squares of side-lengths $r/2^i$, for $i = 0, 1, \ldots, \lceil \log n \rceil$. Refer to Fig. 1(right).

Previously, Mitchell [35] used a similar quadtree decomposition for TSPN with disjoint regions in the plane, coined the term "stratified grid," and derived several basic properties of quadtrees that we rederive here. Specifically, he proved analogues of Lemmas 2 and 5 for the problem studied in [35]. However, Mitchell used stratified grids only for special types of orthogonal polygons, called *histograms* [27]; here we generalize this tool to arbitrary geometric graphs.

The following algorithm subdivides a square Q unless it is too small (i.e., $\mathrm{diam}(Q) < \frac{r}{2n}$) or it is relatively far from G_1 (i.e., $\mathrm{diam}(Q) < \mathrm{dist}(Q, G_1)$).

Stratify $(\mathbf{R}, \mathbf{G_1})$ Let L be a FIFO queue and $\mathcal{Q}$ be a set of axis-aligned squares. Set $L = (R)$ and $\mathcal{Q} = \emptyset$. Repeat the following while L is nonempty. Set $Q \leftarrow \texttt{dequeue}(L)$. If $\mathrm{diam}(Q) \geq \max(\frac{r}{2n}, \mathrm{dist}(Q, G_1))$, then subdivide Q into four congruent axis-aligned squares, and enqueue them onto L. Otherwise, let $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{Q\}$. Return $\mathcal{Q}$.

It is worth noting that $\mathcal{Q}$ does not directly depend on the disks in S_3, but only indirectly, via G_1. By construction, the squares in $\mathcal{Q}$ are interior-disjoint, and every square in $\mathcal{Q}$ has diameter at least $r/(4n)$. Consequently, the number of squares in $\mathcal{Q}$ is $O(n^2)$. Thus the algorithm **Stratify**(R, G_1) runs in polynomial time in n, since $O(n^2)$ squares are enqueued onto L, and $\mathrm{dist}(Q, G_1)$ can be computed in polynomial time for all $Q \in L$. We show that the squares in $\mathcal{Q}$ have a property similar to the disks in S_3 (cf. Lemma 1): only larger squares can be farther from G_1.

Lemma 2 *For every square $Q \in \mathcal{Q}$, we have* $\mathrm{dist}(Q, G_1) \leq 3\,\mathrm{diam}(Q)$.

Proof Put $q = \mathrm{diam}(Q)$. Recall that Q is obtained by subdividing a square Q', $Q \subset Q'$, with $\mathrm{diam}(Q') = 2q$. Since Q' is subdivided by the algorithm, we have $\mathrm{diam}(Q') \geq \max(\frac{r}{2n}, \mathrm{dist}(Q', G_1))$. Since $\mathrm{dist}(p', Q) \leq q$ for every point $p' \in Q'$, the triangle inequality yields $\mathrm{dist}(Q, G_1) \leq 3q$. □

For every square $Q \in \mathcal{Q}$, we define a graph $\gamma(Q)$ that consists of the boundary of Q and a shortest line segment from Q to G_1; see Fig. 2(left). By Lemma 2, we have $\mathrm{len}(\gamma(Q)) \leq (3 + 2\sqrt{2})\,\mathrm{diam}(Q)$; on the other hand, $\mathrm{len}(\gamma(Q)) \geq \mathrm{per}(Q) = 2\sqrt{2}\,\mathrm{diam}(Q)$, and so we have the following.

Corollary 2 *For every $Q \in \mathcal{Q}$, we have* $\mathrm{len}(\gamma(Q)) = \Theta(\mathrm{diam}(Q))$.

The following observation is crucial for reducing the problem of approximating $\mathsf{OPT}(S_3, G_1)$ to a minimum weight hitting set problem.

Lemma 3 *If a square $Q \in \mathcal{Q}$ intersects a disk $D \in S_3$, then*

(i) $\mathrm{diam}(Q) \leq 2\,\mathrm{diam}(D)$, *and*
(ii) *D intersects the boundary of Q (and the graph $\gamma(Q)$ in particular).*

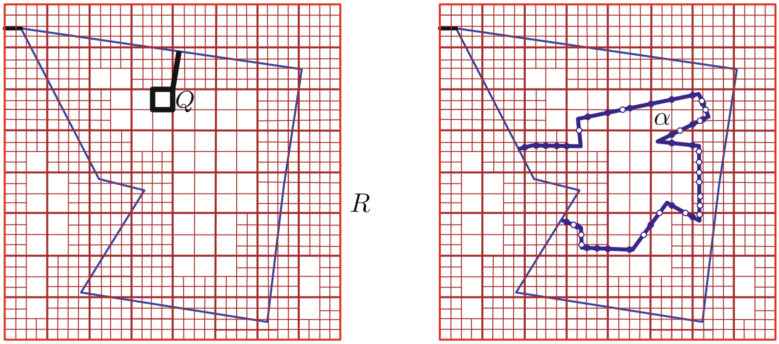


Fig. 2 *Left*: a *square* Q of the stratified grid, and the graph $\gamma(Q)$. *Right*: a polygonal curve α; the intersections of α with *horizontal* (resp., *vertical*) *edges* of the stratified grid are marked with *empty* (resp., *full*) *dots*

Proof

(i) Since $Q \in \mathcal{Q}$, Algorithm **Stratify**(R, G_1) did not subdivide Q, and so we have $\operatorname{diam}(Q) < \frac{r}{2n}$ or $\operatorname{diam}(Q) < \operatorname{dist}(Q, G_1)$. If $\operatorname{diam}(Q) < \frac{r}{2n}$, then Corollary 1 yields

$$\operatorname{diam}(Q) < \frac{r}{2n} < \frac{r}{n} < \operatorname{diam}(D) < 2\operatorname{diam}(D).$$

If $\operatorname{diam}(Q) < \operatorname{dist}(Q, G_1)$, then $\operatorname{dist}(Q, G_1) \leq \operatorname{dist}(D, G_1) + \operatorname{diam}(D)$ follows from the intersection condition and the triangle inequality. Consequently,

$$\operatorname{diam}(Q) < \operatorname{dist}(Q, G_1) \leq \operatorname{dist}(D, G_1) + \operatorname{diam}(D) \leq 2\operatorname{diam}(D),$$

where the last inequality holds by Lemma 1.

(ii) Suppose, to the contrary, that the boundary of Q is disjoint from D, hence D lies in the interior of Q. This immediately implies

$$\operatorname{dist}(Q, G_1) \leq \operatorname{dist}(D, G_1). \tag{2}$$

Since $Q \in \mathcal{Q}$, Algorithm **Stratify**(R, G_1) did not subdivide Q, and so we have $\operatorname{diam}(Q) < \frac{r}{2n}$ or $\operatorname{diam}(Q) < \operatorname{dist}(Q, G_1)$. If $\operatorname{diam}(Q) < \frac{r}{2n}$, Corollary 1 yields $\operatorname{diam}(Q) < \frac{r}{2n} < \frac{r}{n} < \operatorname{diam}(D)$. If $\operatorname{diam}(Q) < \operatorname{dist}(Q, G_1)$, then the combination of (2) and Lemma 1 yield

$$\operatorname{diam}(Q) < \operatorname{dist}(Q, G_1) \leq \operatorname{dist}(D, G_1) < \operatorname{diam}(D).$$

In both cases, we have shown that $\operatorname{diam}(Q) < \operatorname{diam}(D)$. Therefore D cannot lie in the interior of Q, which contradicts the assumption. □

Recall that the squares in $\mathcal{Q}$ can only intersect at common boundary points; we call such squares *adjacent*.

Lemma 4 *If two squares $Q_1, Q_2 \in \mathcal{Q}$ are adjacent and* $\mathrm{diam}(Q_1) \leq \mathrm{diam}(Q_2)$*, then*

$$\frac{1}{2}\mathrm{diam}(Q_2) \leq \mathrm{diam}(Q_1) \leq \mathrm{diam}(Q_2) \leq 2\ \mathrm{diam}(Q_1).$$

Proof If $\mathrm{diam}(Q_1) = \mathrm{diam}(Q_2)$, the inequalities are satisfied. We may thus assume that $\mathrm{diam}(Q_1) < \mathrm{diam}(Q_2) = q$. Then Algorithm **Stratify**(R, G_1) subdivided a square Q_1' such that $Q_1 \subset Q_1'$, $\mathrm{diam}(Q_1') \leq q$, and $Q_1' \cap Q_2 \neq \emptyset$. The algorithm subdivided Q_1' but did not subdivide Q_2. This implies

$$q \geq \max\left(\frac{r}{2n}, \mathrm{dist}(Q_1', G_1)\right) \text{ and } q < \max\left(\frac{r}{2n}, \mathrm{dist}(Q_2, G_1)\right).$$

The first inequality yields $q \geq \max(\frac{r}{2n}, \mathrm{dist}(Q_1', G_1)) \geq \frac{r}{2n}$, and then the second inequality yields $q < \max(\frac{r}{2n}, \mathrm{dist}(Q_2, G_1)) = \mathrm{dist}(Q_2, G_1)$. Consequently, $\mathrm{dist}(Q_1', G_1) \leq q < \mathrm{dist}(Q_2, G_1)$.

Since Q_1' and Q_2 intersect, and their diameters are at most q, the triangle inequality yields $|\mathrm{dist}(Q_1', G_1) - \mathrm{dist}(Q_2, G_1)| \leq q$. It follows that

$$q < \mathrm{dist}(Q_2, G_1) \leq \mathrm{dist}(Q_1', G_1) + q \leq 2q.$$

Similarly, since Q_1 and Q_2 intersect, $\mathrm{dist}(Q_1, G_1) \geq \mathrm{dist}(Q_2, G_1) - \mathrm{diam}(Q_1) > q - \mathrm{diam}(Q_1)$. Combining with Lemma 2, we get

$$3\,\mathrm{diam}(Q_1) \geq \mathrm{dist}(Q_1, G_1) > q - \mathrm{diam}(Q_1),$$

that is, $\mathrm{diam}(Q_1) > q/4$. Finally, recall that the ratio between the diameters of any two squares in $\mathcal{Q}$ is a power of 2. Therefore $q/4 < \mathrm{diam}(Q_1) < q$ yields $\mathrm{diam}(Q_1) = q/2$, as required. □

5 Hitting Sets for Squares and Disks

For the graph G_1 and the set of disks S_3, we define a hypergraph $\mathcal{G} = (\mathcal{Q}, E)$, where the vertex set is the set $\mathcal{Q}$ of squares in the stratified grid; and for every disk $D \in S_3$, the set of squares in $\mathcal{Q}$ that intersect D forms a hyperedge in E. Thus, a subset $\mathcal{H} \subseteq \mathcal{Q}$ of squares is a *hitting set* in the hypergraph $\mathcal{G}$ if and only if every disk in S_3 intersects some square in $\mathcal{H}$.

For every hitting set $\mathcal{H}$, the geometric graph $\Gamma = \cup_{Q \in \mathcal{H}} \gamma(Q)$ intersects every disk in S_3 by Lemma 3, and $G_1 \cup \Gamma$ is connected by construction. Let the *weight* of a square $Q \in \mathcal{Q}$ be $w(Q) = \mathrm{diam}(Q)$. In this section, we show that the minimum-weight hitting set for $\mathcal{G} = (\mathcal{Q}, E)$ is a $O(1)$-approximation for $\mathsf{OPT}(S_3, G_1)$. The following technical lemma considers a single curve (i.e., a Jordan arc). For a curve α, let $\mathcal{Q}(\alpha)$ denote the set of squares in $\mathcal{Q}$ that intersect α. Refer to Fig. 2(right).

Lemma 5 *Let α be a directed polygonal curve whose start and end points lie on G_1. If α intersects at least one disk in S_3, then* $\operatorname{len}(\alpha) = \Omega(\sum_{Q\in\mathcal{Q}(\alpha)} \operatorname{diam}(Q))$.

Proof By (1), we have $\operatorname{dist}(D, G_1) > \frac{r}{n}$ for every disk $D \in S_3$. Since α intersects at least one disk in S_3, we have $\operatorname{len}(\alpha) \geq \frac{2r}{n}$.

Let $A = (Q_0, Q_1, \ldots, Q_m)$ be the sequence of distinct squares that intersect α in the order in which they are first encountered by α (with no repetitions and ties broken arbitrarily). Since Q_0 intersects G_1, we have $\operatorname{diam}(Q_0) < \max(\frac{r}{2n}, 0) = \frac{r}{2n}$, and consequently

$$\operatorname{len}(\alpha) \geq \frac{2r}{n} \geq 4\operatorname{diam}(Q_0). \tag{3}$$

Let $B = (Q_{\sigma(0)}, Q_{\sigma(1)}, \ldots, Q_{\sigma(\ell)})$ be the subsequence of A such that $\sigma(0) = 0$ and a square Q_i, $1 \leq i \leq m$, is added to B if it is disjoint from Q_j for all $0 \leq j < i$. By construction, B consists of pairwise disjoint squares, and every square in A is either in B or adjacent to some square in B. By Lemma 4, the sizes of adjacent squares in $\mathcal{Q}$ differ by a factor of at most 2. Consequently, each square in $\mathcal{Q}$ is adjacent to at most 12 squares in $\mathcal{Q}$ (at most two along each side and at most one at each corner). It follows that

$$\sum_{i=0}^{m} \operatorname{diam}(Q_i) = \Theta\left(\sum_{j=0}^{\ell} \operatorname{diam}(Q_{\sigma(j)})\right). \tag{4}$$

For $j = 0, \ldots, \ell$, let $p_{\sigma(j)}$ be the first intersection point of α with $Q_{\sigma(j)}$. For two points $p, q \in \alpha$, denote by $\alpha(p, q)$ the portion of α between p and q. Since the squares in B are pairwise disjoint, and the sizes of adjacent squares differ by at most a factor of 2 (Lemma 4), we have

$$\operatorname{len}\left(\alpha(p_{\sigma(j)}, p_{\sigma(j+1)})\right) \geq |p_{\sigma(j)}p_{\sigma(j+1)}| \geq \frac{1}{2\sqrt{2}} \max\left(\operatorname{diam}(Q_{\sigma(j)}), \operatorname{diam}(Q_{\sigma(j+1)})\right)$$

for $j = 0, \ldots, \ell - 1$. Consequently, if $\ell \geq 1$, we have

$$\operatorname{len}(\alpha) = \sum_{j=0}^{\ell-1} \operatorname{len}\left(\alpha(p_{\sigma(j)}, p_{\sigma(j+1)})\right) = \Omega\left(\sum_{j=0}^{\ell} \operatorname{diam}(Q_{\sigma(j)})\right). \tag{5}$$

The combination of (3), (4), and (5) yields $\operatorname{len}(\alpha) = \Omega(\sum_{i=0}^{m} \operatorname{diam}(Q_i))$, as required. □

Lemma 6 *If Γ is a geometric graph such that Γ intersects every disk in S_3 and $G_1 \cup \Gamma$ is connected, then there is a hitting set $\mathcal{H} \subseteq \mathcal{Q}$ for $\mathcal{G} = (\mathcal{Q}, E)$ such that*

$$\operatorname{len}(\Gamma) = \Omega\left(\sum_{Q\in\mathcal{H}} \operatorname{diam}(Q)\right). \tag{6}$$

Proof Let $\mathcal{H}$ be the set of squares in $\mathcal{Q}$ that intersect Γ, and observe that $\mathcal{H}$ is a hitting set for $\mathcal{G}$. For each connected component C of Γ, let α be a directed polygonal curve that starts and ends at some points in $G_1 \cap C$ and traverses every edge of C at least once and at most twice. Then $\text{len}(C) \geq \frac{1}{2}\text{len}(\alpha)$, and $\text{len}(\alpha) = \Omega(\sum_{Q\in\mathcal{Q}(\alpha)} \text{diam}(Q))$ by Lemma 5. Summation over all the components of Γ yields (6). □

Recall that $\mathsf{OPT}(S_3, G_1)$ is a geometric graph Γ of minimum length such that $G_1 \cup \Gamma$ is connected and intersects every disk in S_3. Let W_0 denote the minimum weight of a hitting set in the hypergraph $\mathcal{G}$. The main result of this section is the following.

Corollary 3 *We have* $W_0 = O(\text{len}(\mathsf{OPT}(S_3, G_1)))$.

Proof Invoke Lemma 6 with $\Gamma = \mathsf{OPT}(S_3, G_1)$. Then $\mathcal{G}$ has a hitting set $\mathcal{H} \subset \mathcal{Q}$ of weight $\sum_{Q\in\mathcal{H}} \text{diam}(Q) = O(\text{len}(\Gamma)) = O(\text{len}(\mathsf{OPT}(S_3, G_1)))$. This is clearly an upper bound on the minimum weight W_0 of a hitting set in $\mathcal{G}$. □

6 Hitting Sets for Points and Disks

In Sect. 5 we defined a hypergraph $\mathcal{G} = (\mathcal{Q}, E)$ for squares and disks; that is, the vertices are squares in $\mathcal{Q}$ and the hyperedges are the squares intersecting a disk in S_3. In order to apply Theorem 2 by Chan et al. [13], we reduce the problem to a traditional geometric hypergraph problem, where the vertices are points in $\mathbb{R}^2$ and a hyperedge corresponds to the set of points contained in a disk $D \in S_3$.

For each square $Q \in \mathcal{Q}$, we define a set of 25 *sentinel* points, and show (Lemma 7) that if a disk $D \in S_3$ intersects Q, then D contains one of the sentinel points of Q. A constant number of sentinels suffice if none of the disks intersecting Q is too small, and indeed, Lemma 3 has shown that this is the case.

For a square $Q \in \mathcal{Q}$, where $Q = [a, a+h] \times [b, b+h]$, let the 25 sentinel points be $(a + ih/2, b + jh/2)$ for all $i, j \in \{-1, 0, 1, 2, 3\}$; see Fig. 3(left).

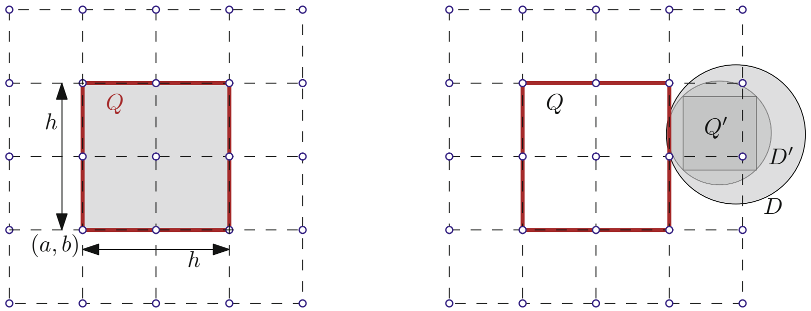


Fig. 3 *Left*: the set of 25 sentinel points for a *square* Q. *Right*: a disk $D \in S_3$ intersects a *square* $Q \in \mathcal{Q}$

Lemma 7 *If a disk $D \in S_3$ intersects a square $Q \in \mathcal{Q}$, then D contains a sentinel point corresponding to Q.*

Proof Assume that a disk $D \in S_3$ intersects a square $Q \in \mathcal{Q}$ of side length h; refer to Fig. 3(right). By Lemma 3, $\mathrm{diam}(Q) \leq 2\,\mathrm{diam}(D)$. By scaling down D from an arbitrary center in $D \cap Q$, we find a disk D' intersecting Q with $\mathrm{diam}(D') = \frac{1}{2}\,\mathrm{diam}(Q)$. The inscribed axis-aligned square Q' of D' has $\mathrm{diam}(Q') = \frac{1}{2}\,\mathrm{diam}(Q)$. That is, the side-length of Q' is $h/2$, and $\mathrm{dist}(Q', Q) \leq (\sqrt{2}-1)h$. Since the sentinels of Q form a section of a square lattice of (the same) side-length $h/2$, within distance $\frac{\sqrt{2}}{2}\,h$ from Q, some sentinel of Q lies in Q', and hence in $D' \subseteq D$, as claimed. □

We define a new weighted hypergraph $\mathcal{G}' = (V', E')$, where V' is the union of sentinel point sets for all $Q \in \mathcal{Q}$ that lie in R (sentinels in the exterior of R are discarded); and each hyperedge in E' is the set of sentinels in V' contained in a disk $D \in S_3$. Note that a sentinel $s \in V'$ may correspond to several squares in $\mathcal{Q}$. Let the *weight* of a sentinel $s \in V'$ be the sum of the diameters of the squares $Q \in \mathcal{Q}$ that correspond to s. Hence the total weight of all sentinels is at most $25 \sum_{Q\in\mathcal{Q}} \mathrm{diam}(Q)$. We next derive a bound on the weight of each sentinel.

Lemma 8 *For every $Q \in \mathcal{Q}$, the weight of every sentinel corresponding to Q is* $O(\mathrm{diam}(Q))$.

Proof By Lemma 4, the side-lengths of adjacent squares of the stratified grid differ by at most a factor of 2. Consequently, every sentinel in V' corresponding to a square $Q \in \mathcal{Q}$ is contained in Q or in a square of $\mathcal{Q}$ adjacent to Q.

Let $s \in V'$ be a sentinel. Then s may correspond to all squares in $\mathcal{Q}$ that contain s, and to adjacent squares in $\mathcal{Q}$. Every point is contained in at most 4 squares of $\mathcal{Q}$, whose side-lengths differ by a factor of at most 2; and they are each adjacent to $O(1)$ additional squares whose side-lengths differ by another factor of at most 2. Overall, s corresponds to $O(1)$ squares in $\mathcal{Q}$ whose side-lengths differ by a factor of $\Theta(1)$. Therefore, the weight of s is $O(\mathrm{diam}(Q))$ for every square $Q \in Q$ corresponding to s. □

By Theorem 2, there is a polynomial-time $O(1)$-approximation algorithm for MWHS on $\mathcal{G}' = (V', E')$. It remains to show that a $O(1)$-approximation for MWHS on the hypergraph $\mathcal{G}' = (V', E')$ provides a $O(1)$-approximation for MWHS on the hypergraph $\mathcal{G} = (\mathcal{Q}, E)$.

Lemma 9

1. *For every hitting set $\mathcal{H} \subseteq \mathcal{Q}$ for $\mathcal{G}$, the set H' of sentinels in V' corresponding to the squares $Q \in \mathcal{H}$ is a hitting set for $\mathcal{G}'$ of weight* $O(\sum_{Q\in\mathcal{H}} \mathrm{diam}(Q))$.
2. *For every hitting set $H' \subseteq V'$ for $\mathcal{G}'$, the set $\mathcal{H}$ of squares $Q \in \mathcal{Q}$ that contain the sentinel points in H' is a hitting set for $\mathcal{G}$ of weight $O(\sum_{s\in H'} w(s))$.*

Proof
(1) If $\mathcal{H}$ is a hitting set for $\mathcal{G}$, then every disk $D \in S_3$ intersects some square $Q \in \mathcal{H}$. By Lemma 7, D contains one of the sentinels of Q. Consequently, every disk $D \in S_3$ contains a sentinel in H'. Every square $Q \in \mathcal{H}$ corresponds to 25 sentinels, each of weight $O(\mathrm{diam}(Q)$ by Lemma 8. The weight of H' is $O(\sum_{Q\in\mathcal{H}} \mathrm{diam}(Q))$.

(2) If H' is a hitting set for $\mathcal{G}'$, then every disk $D \in S_3$ contains some point $s \in H'$. The point s lies in a square $Q \in \mathcal{Q}$ of the stratified grid, which is in $\mathcal{H}$. Consequently, every disk $D \in S_3$ intersects some square $Q \in \mathcal{H}$. By construction, the weight of each sentinel s is the sum of weights of the corresponding squares in $\mathcal{Q}$, including all squares in $\mathcal{Q}$ that contain s. Therefore, the weight of $\mathcal{H}$ is at most $\sum_{s \in H'} w(s)$, as required. □

We are now ready to prove Theorem 1 by analyzing the constructed graph $G = G_1 \cup G_2 \cup G_3$.

Proof of Theorem 1. Let S be a set of n disks in $\mathbb{R}^2$. Compute an independent set $I \subset S$ as described in Sect. 3, and a TSP tour ξ_0 for I with $\text{len}(\xi_0) = O(\text{len}(\mathsf{OPT}))$ (as in [6] or [19]). Compute the graph G_1 with $\text{len}(G_1) = O(\text{len}(\mathsf{OPT}))$, and the partition $S = S_1 \cup S_2 \cup S_3$ as described in Sect. 3. The graph G_1 intersects the disks in S_1.

Construct the graph G_2, which contains a shortest segment between G_1 and every disk $D \in S_2$. The length of this graph is $\text{len}(G_2) = \sum_{D \in S_2} \text{dist}(D, G_1) \le |S_2| \cdot \frac{r}{n} \le r \le \text{len}(\mathsf{OPT})$.

Compute the stratified grid $\mathcal{Q}$, and construct the weighted hypergraph $\mathcal{G}' = (V', E')$, where V' is the set of sentinel points for all squares $Q \in \mathcal{Q}$, the weight of a sentinel s is the sum of diameters of the corresponding squares $Q \in \mathcal{Q}$, and for every disk $D \in S_3$, the set of sentinels lying in D forms a hyperedge in E'. Use the algorithm by Chan et al. [13] to compute a hitting set H' for $\mathcal{G}'$ whose weight is $O(1)$ times the minimum. Let $\mathcal{H}$ be the set of squares in $\mathcal{Q}$ containing the sentinels in H'. By Lemma 9, $\mathcal{H}$ is a hitting set for the hypergraph $\mathcal{G} = (\mathcal{Q}, E)$ whose weight is at most $O(1)$ times the minimum W_0. Put $G_3 = \cup_{Q \in \mathcal{H}} \gamma(Q)$. Then G_3 intersects every disk in S_3 by Lemma 3, and $\text{len}(G_3) = \Theta(\sum_{Q \in \mathcal{H}} \text{diam}(Q)) = \Theta(W_0)$ by Corollary 2. Finally, Corollary 3 yields $\text{len}(G_3) = O(\text{len}(\mathsf{OPT}(S_3, G_1)))$. By the definition of $\mathsf{OPT}(S_3, G_1))$, we have $\text{len}(\mathsf{OPT}(S_3, G_1)) \le \text{len}(\mathsf{OPT}(S_3)) \le \text{len}(\mathsf{OPT})$, and consequently $\text{len}(G_3) = O(\text{len}(\mathsf{OPT}))$.

Note that the graph $G = G_1 \cup G_2 \cup G_3$ is connected by construction, it intersects every disk in $S = S_1 \cup S_2 \cup S_3$, and $\text{len}(G) = \text{len}(G_1) + \text{len}(G_2) + \text{len}(G_3) = O(\text{len}(\mathsf{OPT}))$. Consequently, an Eulerian tour of the multi-graph containing each edge of G twice visits each disk and its length is $2\,\text{len}(G) = O(\text{len}(\mathsf{OPT}))$, as required.

Since the above steps as well as algorithm **Stratify**(R, G_1) all run in time that is polynomial in n, the constant-factor approximation algorithm for TSP with disks runs in polynomial time. □

7 Conclusions

In this paper, we obtained the first constant-ratio approximation for TSP with disks in the plane. This is the first result of this kind for a class of planar convex bodies of arbitrary size that can intersect in an arbitrary fashion. In light of the connection we established between TSPN and MWHS in geometric hypergraphs, the following question emerges:

1. Besides regions of linear union complexity (e.g., disks and pseudo-disks[1]), what other types of regions admit a constant-factor approximation for the minimum weight hitting set problem?

Obviously, a constant-factor approximation for MWHS with a certain type of neighborhoods does not automatically imply a constant-factor approximation for TSPN with the same type of neighborhoods. We conclude with a few, perhaps the simplest still unsolved questions on TSPN that we could identify:

2. Is there a constant-factor approximation algorithm for TSP with a set of objects of linear union complexity, e.g., pseudo-disks?
3. Is there a constant-factor approximation algorithm for TSP with convex bodies in the plane?[2]
4. Is TSP with disks in the plane APX-hard? Is TSP with convex bodies in the plane APX-hard?
5. Is there a constant-factor approximation algorithm for TSP with balls (with arbitrary radii and intersections) in $\mathbb{R}^d$, in fixed dimension $d \geq 3$?

References

1. P.K. Agarwal, E. Ezra, M. Sharir, Near-linear approximation algorithms for geometric hitting sets. Algorithmica **63**(1–2), 1–25 (2012)
2. E.M. Arkin, R. Hassin, Approximation algorithms for the geometric covering salesman problem. Discret. Appl. Math. **55**(3), 197–218 (1994)
3. S. Arora, Polynomial time approximation schemes for Euclidean traveling salesman and other geometric problems. J. ACM **45**(5), 753–782 (1998)
4. Y. Bartal, L.-A. Gottlieb, R. Krauthgamer, The traveling salesman problem: low-dimensionality implies a polynomial time approximation scheme. SIAM J. Comput. **45**(4), 1563–1581 (2016)
5. M. de Berg, O. Cheong, M. van Kreveld, M. Overmars, *Computational Geometry*, 3rd edn. (Springer, Heidelberg, 2008)
6. M. de Berg, J. Gudmundsson, M.J. Katz, C. Levcopoulos, M.H. Overmars, A.F. van der Stappen, TSP with neighborhoods of varying size. J. Algorithms **57**(1), 22–36 (2005)
7. M. Bern, D. Eppstein, Approximation algorithms for geometric problems, in *Approximation Algorithms for NP-Hard Problems*, ed. by D.S. Hochbaum (PWS Publishing Company, Boston, 1997), pp. 296–345

[1] A set of regions $\mathcal{R}$ consists of *pseudo-disks*, if every pair of regions $\omega_1, \omega_2 \in \mathcal{R}$ satisfies the *pseudo-disk property*: the sets $\omega_1 \setminus \omega_2$ and $\omega_2 \setminus \omega_1$ are connected [5, p. 293]. Equivalently, the boundaries $\partial\omega_1$ and $\partial\omega_2$ have at most two proper intersection points.

[2] Very recently, Mitchell [37] proposed a constant-factor approximation algorithm for this variant. However, no complete proof is available at the time of this writing. In fact, we believe that TSP with planar convex bodies is much harder to approximate than TSP with disks.

8. H.L. Bodlaender, C. Feremans, A. Grigoriev, E. Penninkx, R. Sitters, T. Wolle, On the minimum corridor connection problem and other generalized geometric problems. Comput. Geom. Theory Appl. **42**(9), 939–951 (2009)
9. H. Brönnimann, M.T. Goodrich, Almost optimal set covers in finite VC-dimension. Discret. Comput. Geom. **14**(4), 463–479 (1995)
10. N. Bus, S. Garg, N.H. Mustafa, S. Ray, Improved local search for geometric hitting set, in *Proceedings of 32nd Symposium on Theoretical Aspects of Computer Science (STACS).* Volume 30 of LIPIcs (Schloss Dagstuhl, 2015), pp. 184–196
11. T.-H.H. Chan, K. Elbassioni, A QPTAS for TSP with fat weakly disjoint neighborhoods in doubling metrics. Discret. Comput. Geom. **46**(4), 704–723 (2011)
12. T.-H.H. Chan, S.H.-C. Jiang, Reducing curse of dimensionality: improved PTAS for TSP (with neighborhoods) in doubling metrics, in *Proceedings of 27th ACM-SIAM Symposium on Discrete Algorithms (SODA)* (SIAM, 2016), pp. 754–765
13. T.M. Chan, E. Grant, J. Könemann, M. Sharpe, Weighted capacitated, priority, and geometric set cover via improved quasi-uniform sampling, in *Proceedings of 23rd ACM-SIAM Symposium on Discrete Algorithms (SODA)* (SIAM, 2012), pp. 1576–1585
14. K.L. Clarkson, K.R. Varadarajan, Improved approximation algorithms for geometric set cover. Discret. Comput. Geom. **37**(1), 43–58 (2007)
15. M. Dror, J.B. Orlin, Combinatorial optimization with explicit delineation of the ground set by a collection of subsets. SIAM J. Discret. Math. **21**(4), 1019–1034 (2008)
16. A. Dumitrescu, J.S.B. Mitchell, Approximation algorithms for TSP with neighborhoods in the plane. J. Algorithms **48**(1), 135–159 (2003)
17. A. Dumitrescu, Cs.D. Tóth, On the total perimeter of homothetic convex bodies in a convex container. Beiträge zur Algebra Geom. **56**(2), 515–532 (2015)
18. A. Dumitrescu, Cs.D. Tóth, The traveling salesman problem for lines, balls and planes. ACM Trans. Algorithms **12**(3), article 43 (2016)
19. K.M. Elbassioni, A.V. Fishkin, R. Sitters, Approximation algorithms for the Euclidean traveling salesman problem with discrete and continuous neighborhoods. Int. J. Comput. Geom. Appl. **19**(2), 173–193 (2009)
20. G. Even, D. Rawitz, S. Shahar, Hitting sets when the VC-dimension is small. Inf. Process. Lett. **95**(2), 358–362 (2005)
21. M.R. Garey, R. Graham, D.S. Johnson, Some NP-complete geometric problems, in *Proceedings of 8th ACM Symposium on Theory of Computing (STOC)* (ACM, 1976), pp. 10–22
22. M.R. Garey, D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (W.H. Freeman and Company, New York, 1979)
23. B. Gärtner, J. Matoušek, L. Rüst, P. Škovroň, Violator spaces: structure and algorithms. Discret. Appl. Math. **156**(11), 2124–2141 (2008)
24. J. Gudmundsson, C. Levcopoulos, A fast approximation algorithm for TSP with neighborhoods. Nord. J. Comput. **6**(4), 469–488 (1999)
25. P. Kamousi, S. Suri, Euclidean traveling salesman tours through stochastic neighborhoods, in *Proceedings of 24th International Symposium on Algorithms and Computation (ISAAC).* Lecture Notes in Computer Science, vol 8283 (Springer, 2013), pp. 644–654
26. K. Kedem, R. Livne, J. Pach, M. Sharir, On the union of Jordan regions and collision-free translational motion amidst polygonal obstacles. Discret. Comput. Geom. **1**(1), 59–70 (1986)
27. C. Levcopoulos, Heuristics for minimum decompositions of polygons. Ph.D. thesis, Linköping Studies in Science and Technology, No. 74 (1987)
28. C. Levcopoulos, A. Lingas, Bounds on the length of convex partitions of polygons, in *Proceedings of 4th Conference on Foundations of Software Technology and Theoretical Computer Science*. Lecture Notes in Computer Science, vol 181 (Springer, 1984), pp. 279–295
29. C. Mata, J.S.B. Mitchell, Approximation algorithms for geometric tour and network design problems, in *Proceedings of 11th ACM Symposium on Computational Geometry (SOCG)* (ACM, 1995), pp. 360–369

30. J. Matoušek, M. Sharir, E. Welzl, A subexponential bound for linear programming. Algorithmica **16**(4), 498–516 (1996)
31. J.S.B. Mitchell, Guillotine subdivisions approximate polygonal subdivisions: a simple polynomial-time approximation scheme for geometric TSP, k-MST, and related problems. SIAM J. Comput. **28**(4), 1298–1309 (1999)
32. J.S.B. Mitchell, Geometric shortest paths and network optimization, in *Handbook of Computational Geometry*, ed. by J.-R. Sack, J. Urrutia (Elsevier, Amsterdam, 2000), pp. 633–701
33. J.S.B. Mitchell, Shortest paths and networks, in *Handbook of Discrete and Computational Geometry,* 3rd edn., ed. by J.E. Goodman, J. O'Rourke, C.D. Tóth (Chapman & Hall/CRC, Boca Raton, 2017, to appear)
34. J.S.B. Mitchell, A PTAS for TSP with neighborhoods among fat regions in the plane, in *Proceedings of 18th ACM-SIAM Symposium on Discrete Algorithms (SODA)* (SIAM, 2007), pp. 11–18
35. J.S.B. Mitchell, A constant-factor approximation algorithm for TSP with pairwise-disjoint connected neighborhoods in the plane, in *Proceedings of 26th Symposium on Computational Geometry (SOCG)* (ACM, 2010), pp. 183–191
36. A. Dumitrescu, J.S.B. Mitchell, Approximation algorithms for TSP with neighborhoods in the plane, CoRR abs/1703.01640 (2017). https://arxiv.org/abs/1703.01640
37. J.S.B. Mitchell, Updated version of "A constant-factor approximation algorithm for TSP with pairwise-disjoint connected neighborhoods in the plane," manuscript, http://www.ams.sunysb.edu/~jsbm/papers/tspn-socg10-updated.pdf. Accessed in Feb 2016
38. N.H. Mustafa, S. Ray, Improved results on geometric hitting set problems. Discret. Comput. Geom. **44**(4), 883–895 (2010)
39. J. Pach, P.K. Agarwal, *Combinatorial Geometry* (Wiley, New York, 1995)
40. C.H. Papadimitriou, Euclidean TSP is NP-complete. Theor. Comput. Sci. **4**(3), 237–244 (1977)
41. S.B. Rao, W.D. Smith, Approximating geometrical graphs via "spanners" and "banyans," in *Proceedings of 30th ACM Symposium on Theory of Computing (STOC)* (ACM, 1998), pp. 540–550
42. S. Safra, O. Schwartz, On the complexity of approximating TSP with neighborhoods and related problems. Comput. Complex. **14**(4), 281–307 (2005)
43. P. Slavik, The errand scheduling problem, CSE technical report 97-02, University of Buffalo, Buffalo (1997)
44. S. Spirkl, The guillotine subdivision approach for TSP with neighborhoods revisited (2014, preprint). arXiv:1312.0378v2
45. K.R. Varadarajan, Epsilon nets and union complexity, in *Proceedings of 25th Symposium on Computational Geometry (SOCG)* (ACM, 2009), pp. 11–16

Transport-Entropy Inequalities and Curvature in Discrete-Space Markov Chains

Ronen Eldan, James R. Lee, and Joseph Lehec

Abstract Let $G = (\Omega, E)$ be a graph and let d be the graph distance. Consider a discrete-time Markov chain $\{Z_t\}$ on Ω whose kernel p satisfies $p(x, y) > 0 \Rightarrow \{x, y\} \in E$ for every $x, y \in \Omega$. In words, transitions only occur between neighboring points of the graph. Suppose further that (Ω, p, d) has coarse Ricci curvature at least $1/\alpha$ in the sense of Ollivier: For all $x, y \in \Omega$, it holds that

$$W_1(Z_1 \mid \{Z_0 = x\}, Z_1 \mid \{Z_0 = y\}) \le \left(1 - \frac{1}{\alpha}\right) d(x, y),$$

where W_1 denotes the Wasserstein 1-distance.

In this note, we derive a transport-entropy inequality: For any measure μ on Ω, it holds that

$$W_1(\mu, \pi) \le \sqrt{\frac{2\alpha}{2 - 1/\alpha} D(\mu \parallel \pi)}\,,$$

where π denotes the stationary measure of $\{Z_t\}$ and $D(\cdot \parallel \cdot)$ is the relative entropy.

Peres and Tetali have conjectured a stronger consequence of coarse Ricci curvature, that a modified log-Sobolev inequality (MLSI) should hold, in analogy with the setting of Markov diffusions. We discuss how our approach suggests a natural attack on the MLSI conjecture.

1 Introduction

In geometric analysis on manifolds, it is by now well-established that the Ricci curvature of the underlying manifold has profound consequences for functional inequalities and the rate of convergence of Markov semigroups toward equilibrium. One can consult, in particular the books [1] and [17]. Indeed, in the setting of

R. Eldan • J.R. Lee (✉) • J. Lehec
Paul G. Allen Center, 352350 Seattle, WA, USA
e-mail: jrl@cs.washington.edu

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_16

diffusions (see [1, §1.11]), there is an elegant theory around the Bakry–Emery *curvature-dimension* condition.

Roughly speaking, in the setting of diffusion on a continuous space, when there is an appropriate "integration by parts" formula (that connects the Dirichlet form to the Laplacian), a positive curvature condition implies powerful functional inequalities. Most pertinent to the present discussion, positive curvature yields transport-entropy and logarithmic Sobolev inequalities.

For discrete state spaces, the situation appears substantially more challenging. There are numerous attempts at generalizing lower bounds on the Ricci curvature to discrete metric measure spaces. At a broad level, these approaches suffer from one of two drawbacks: Either the notion of "positive curvature" is difficult to verify for concrete spaces, or the "expected" functional analytic consequences do not follow readily.

In the present note, we consider the notion of *coarse Ricci curvature* due to Ollivier [14]. It constitutes an approach of the latter type: There is a large body of finite-state Markov chains that have positive curvature in Ollivier's sense, but for many of them we do not yet know if strong functional-analytic consequences hold. This study is made more fascinating by the straightforward connection between coarse Ricci curvature on graphs and the notion of *path coupling* arising in the study of rapid mixing of Markov chains [3]. This is a powerful method to establish fast convergence to the stationary measure; see, for example, [10, Ch. 14].

In particular, if there were an analogy to the diffusion setting that allowed coarse Ricci curvature lower bounds to yield logarithmic Sobolev inequalities (or variants thereof), it would even imply new mixing time bounds for well-studied chains arising from statistical physics, combinatorics, and theoretical computer science. A conjecture of Peres and Tetali asserts that a *modified log-Sobolev inequality* (MLSI) should always hold in this setting. Roughly speaking, this means that the underlying Markov chain has exponential convergence to equilibrium in the relative entropy distance.

Our aim is to give some preliminary results in this direction and to suggest a new approach to establishing MLSI. In particular, we prove a W_1 transport-entropy inequality. By results of Bobkov and Götze [2], this is equivalent to a sub-Gaussian concentration estimate for Lipschitz functions. Sammer has shown that such an inequality follows formally from MLSI [16], thus one can see verification as evidence in support of the Peres-Tetali conjecture. Our result also addresses PROBLEM J in Ollivier's survey [15].

1.1 Coarse Ricci Curvature and Transport-Entropy Inequalities

Let Ω be a countable state space, and let $p : \Omega \times \Omega \to [0, 1]$ denote a transition kernel. For $x \in \Omega$, we will use the notation $p(x, \cdot)$ to denote the function $y \mapsto p(x, y)$.

For a probability measure π on Ω and $f : \Omega \to \mathbb{R}_+$, we define the entropy of f by

$$\mathrm{Ent}_\pi(f) = \mathbb{E}_\pi \left[f \log \left(\frac{f}{\mathbb{E}_\pi[f]} \right) \right].$$

We also equip Ω with a metric d. If μ and ν are two probability measures on Ω, we denote by $W_1(\mu,\nu)$ the transportation cost (or Wasserstein 1-distance) between μ and ν, with the cost function given by the distance d. Namely,

$$W_1(\mu,\nu) = \inf\{\mathbb{E}[d(X,Y)]\}$$

where the infimum is taken on all couplings (X,Y) of (μ,ν). Recall the Monge–Kantorovitch duality formula for W_1 (see, for instance, [17, Case 5.16]):

$$W_1(\mu,\nu) = \sup\left\{ \int_\Omega f\, d\mu - \int_\Omega f\, d\nu \right\}, \tag{1}$$

where the supremum is taken over 1–Lipschitz functions f. We consider the following notion of curvature introduced by Ollivier [14].

Definition 1.1 The *coarse Ricci curvature* of (Ω, p, d) is the largest $\kappa \in [-\infty, 1]$ such that the inequality

$$W_1(p(x,\cdot), p(y,\cdot)) \le (1-\kappa)\, d(x,y)$$

holds true for every $x, y \in \Omega$.

In the sequel we will be interested in positive Ricci curvature. Under this condition the map $\mu \mapsto \mu p$ is a contraction for W_1. As a result, it has a unique fixed point and μp^n converges to this fixed point as $n \to \infty$. In other words the Markov kernel p has a unique stationary measure and is ergodic. The main purpose of this note is to show that positive curvature yields a transport-entropy inequality, or equivalently a Gaussian concentration inequality for the stationary measure.

Definition 1.2 We say that a probability measure μ on Ω satisfies the *Gaussian concentration property* with constant C if the inequality

$$\int_\Omega \exp(f)\, d\mu \le \exp\left(\int_\Omega f\, d\mu + C\, \|f\|_{\mathrm{Lip}}^2 \right)$$

holds true for every Lipschitz function f.

Now we spell out the dual formulation of the Gaussian concentration property in terms of transport inequality. Recall first the definition of the relative entropy (or Kullback divergence): for two measures μ, ν on $(\Omega, \mathcal{B})$,

$$D(\nu \,\|\, \mu) = \mathrm{Ent}_\mu[\tfrac{d\nu}{d\mu}] = \int_\Omega \log\left(\frac{d\nu}{d\mu} \right) d\nu$$

if ν is absolutely continuous with respect to μ and $D(\nu \parallel \mu) = +\infty$ otherwise. As usual, if X and Y are random variables with laws ν and μ, we will take $D(X \parallel Y)$ to be synonymous with $D(\nu \parallel \mu)$.

Definition 1.3 We say that μ satisfies (T_1) with constant C if for every probability measure ν on Ω we have

$$W_1(\mu, \nu)^2 \leq C \cdot D(\nu \parallel \mu)\,. \tag{T_1}$$

As observed by Bobkov and Götze [2], the inequality (T_1) and the Gaussian concentration property are equivalent.

Lemma 1.4 *A probability measure μ satisfies the Gaussian concentration property with constant C if and only if it satisfies (T_1) with constant $4C$.*

This is a relatively straightforward consequence of the Monge–Kantorovitch duality (1); we refer to [2] for details.

Theorem 1.5 *Assume that (Ω, p, d) has positive coarse Ricci curvature $1/\alpha$ and that the one–step transitions all satisfy (T_1) with the same constant C: Suppose that for every $x \in \Omega$ and for every probability measure ν we have*

$$W_1(\nu, p(x,\cdot))^2 \leq C \cdot D(\nu \parallel p(x,\cdot))\ . \tag{2}$$

Then the stationary measure π satisfies (T_1) with constant $\frac{C\alpha}{2-1/\alpha}$.

Remark 1.6 Observe that Theorem 1.5 does not assume reversibility.

The hypothesis (2) might seem unnatural at first sight but it is automatically satisfied for the random walk on a graph when d is the graph distance. Indeed, recall Pinsker's inequality: For every probability measures μ, ν we have

$$\mathrm{TV}(\mu, \nu) \leq \sqrt{\frac{1}{2} D(\mu \parallel \nu)},$$

where TV denotes the total variation distance. This yields the following lemma.

Lemma 1.7 *Let μ be a probability measure on a metric space (M, d) and assume that the support of μ has finite diameter Δ. Then μ satisfies (T_1) with constant $\Delta^2/2$.*

Proof Let ν be absolutely continuous with respect to μ. Then both μ and ν are supported on a set of diameter Δ. This implies that

$$W_1(\mu, \nu) \leq \Delta \cdot \mathrm{TV}(\mu, \nu).$$

Combining this with Pinsker's inequality we get $W_1(\mu, \nu)^2 \leq \frac{\Delta^2}{2} D(\nu \parallel \mu)$, which is the desired result. □

Random walks on graphs A particular case of special interest will be random walks on finite graphs. Let $G = (V, E)$ be a connected, undirected graph, possibly with self-loops. Given non-negative conductances $c : E \to \mathbb{R}_+$ on the edges, we recall the Markov chain $\{X_t\}$ defined by

$$\Pr[X_{t+1} = y \mid X_t = x] = \frac{c(\{x, y\})}{\sum_{z \in V} c(\{x, z\})} .$$

We refer to any such chain as *a random walk on the graph G*. If it holds that $c(\{x, x\}) \geq \frac{1}{2} \sum_{Z \in V} c(\{x, z\})$ for all $x \in V$, we say that the corresponding random walk is *lazy.* We will equip G with its graph distance d.

In this setting, the transitions of the walk are supported on a set of diameter 2. So combining the preceding lemma with Theorem 1.5, one arrives at the following.

Corollary 1.8 *If a random walk on a graph has positive coarse Ricci curvature* $\frac{1}{\alpha}$ *(with respect to the graph distance), then the stationary measure* π *satisfies*

$$W_1(\mu, \pi)^2 \leq \frac{2\alpha}{2 - 1/\alpha} D(\mu \parallel \pi) ,$$

for every probability measure μ.

Remark 1.9 One should note that in this context we have

$$d(x, y) \leq W_1 \left(p(x, \cdot), p(y, \cdot)\right) + 2, \quad \forall x, y \in \Omega ,$$

just because after one step the walk is at distance 1 at most from its starting point. As a result, having coarse Ricci curvature $1/\alpha$ implies that the diameter Δ of the graph is at most 2α. So by the previous lemma, *every* measure on the graph satisfies T_1 with constant $2\alpha^2$. The point of Corollary 1.8 is that for the stationary measure π the constant is order α rather than α^2.

We now present two proofs of Theorem 1.5. The first proof is rather short and based on the duality formula (1). The second argument provides an explicit coupling based on an entropy-minimal drift process. In Sect. 3, we discuss logarithmic Sobolev inequalities. In particular, we present a conjecture about the structure of the entropy-minimal drift that is equivalent to the Peres–Tetali MLSI conjecture.

After the first version of this note was released we were notified that Theorem 1.5 was proved by Djellout, Guillin and Wu in [4, Proposition 2.10]. Note that this article actually precedes Ollivier's work. The proof given there corresponds to our first proof, by duality. Our second proof is more original but does share some similarities with the argument given by K. Marton in [11, Proposition 1]. Also, after hearing about our work, Fathi and Shu [5] used their transport-information framework to provide yet another proof.

2 The W_1 Transport-Entropy Inequality

We now present two proofs of Theorem 1.5. Recall the relevant data (Ω, p, d). Define the process $\{B_t\}$ to be the discrete-time random walk on Ω corresponding to the transition kernel p. For $x \in \Omega$, we will use $B_t(x)$ to denote the random variable $B_t \mid \{B_0 = x\}$. For $t \geq 0$, we make the definition

$$P_t[f](x) = \mathbb{E}[f(B_t(x))] \,.$$

2.1 Proof by Duality

Let $f : \Omega \to \mathbb{R}$ be a Lipschitz function. Using the hypothesis (2) and Lemma 1.4 we get

$$P_1[\exp(f)](x) \leq \exp\left(P_1[f](x) + \frac{C}{4}\|f\|_{\mathrm{Lip}}^2\right),$$

for all $x \in \Omega$. Applying this inequality repeatedly we obtain

$$P_n[\exp(f)](x) \leq \exp\left(P_n[f](x) + \frac{C}{4}\sum_{k=0}^{n-1}\|P_k f\|_{\mathrm{Lip}}^2\right), \tag{3}$$

for every integer n and all $x \in \Omega$. Now we use the curvature hypothesis. Note that the Monge–Kantorovitch duality (1) yields easily

$$\begin{aligned} 1 - \tfrac{1}{\alpha} &= \sup_{x \neq y}\left\{\frac{W_1(p(x,\cdot), p(y,\cdot))}{d(x,y)}\right\} \\ &= \sup_{x \neq y, g}\left\{\frac{P_1[g](x) - P_1[g](y)}{\|g\|_{\mathrm{Lip}} d(x,y)}\right\} \\ &= \sup_{g}\left\{\frac{\|P_1[g]\|_{\mathrm{Lip}}}{\|g\|_{\mathrm{Lip}}}\right\}. \end{aligned}$$

Therefore $\|P_1[g]\|_{\mathrm{Lip}} \leq (1 - 1/\alpha)\|g\|_{\mathrm{Lip}}$ for every Lipschitz function g and thus

$$\|P_n[f]\|_{\mathrm{Lip}} \leq (1 - 1/\alpha)^n \|f\|_{\mathrm{Lip}},$$

for every integer n. Inequality (3) then yields

$$P_n\left[\exp(f)\right](x) \leq \exp\left(P_n[f](x) + \frac{C\alpha}{4(2 - 1/\alpha)}\|f\|_{\mathrm{Lip}}^2\right).$$

Letting $n \to \infty$ yields

$$\int_\Omega \exp(f)\, d\pi \leq \exp\left(\int_\Omega f\, d\pi + \frac{C\alpha}{4\,(2-1/\alpha)}\|f\|_{\mathrm{Lip}}^2\right).$$

The stationary measure π thus satisfies Gaussian concentration with constant $\frac{C\alpha}{4\,(2-1/\alpha)}$. Another application of the duality, Lemma 1.4, yields the desired outcome, proving Theorem 1.5.

2.2 An Explicit Coupling

As promised, we now present a second proof of Theorem 1.5 based on an explicit coupling. The proof does not rely on duality, and our hope is that the method presented will be useful for establishing MLSI; see Sect. 3.

The first step of the proof follows a similar idea to the one used in [11, Proposition 1]. Given the random walk $\{B_t\}$ and another process $\{X_t\}$ (not necessarily Markovian), there is a natural coupling between the two processes that takes advantage of the curvature condition and gives a bound on the distance between the processes at time T in terms of the relative entropy. This step is summarized in the following result.

Proposition 2.1 *Assume that (Ω, p, d) satisfies the conditions of Theorem 1.5. Fix a time T and a point $x_0 \in M$. Let $\{B_0 = x_0, B_1, \dots, B_T\}$ be the corresponding discrete time random walk starting from x_0 and let $\{X_0 = x_0, X_1, \dots, X_T\}$ be an arbitrary random process on Ω starting from x_0. Then, there exists a coupling between the processes (X_t) and (B_t) such that*

$$E[d(X_T, B_T)] \leq \sqrt{\frac{C\alpha}{2-1/\alpha} D(\{X_0, X_1, \dots, X_T\} \,\|\, \{B_0, B_1, \dots, B_T\})}.$$

In view of the above proposition, proving a transportation-entropy inequality for (Ω, p, d) is reduced to the following: given a measure ν on Ω, we are looking for a process $\{X_t\}$ which satisfies: (i) $X_T \sim \nu$ and (ii) the relative entropy between $(X_0, \dots, X_T)$ and $(B_0, \dots, B_T)$ is as small as possible.

To achieve the above, our key idea is the construction of a process X_t which is entropy minimal in the sense that it satisfies

$$X_T \sim \nu \text{ and } D(\{X_0, X_1, \dots, X_T\} \,\|\, \{B_0, B_1, \dots, B_T\}) = D(X_T \,\|\, B_T)\,. \tag{4}$$

This process can be thought of as the Doob transform of the random walk with a given target law. In the setting of Brownian motion on $\mathbb{R}^n$ equipped with the Gaussian measure, the corresponding process appears in work of Föllmer [6, 7]. See [8] for applications to functional inequalities, and the work of Léonard [9] for a somewhat different perspective on the connection to optimal transportation.

2.2.1 Proof of Proposition 2.1: Construction of the Coupling

Given $t \in \{1, \dots, T\}$ and $x_1, \dots, x_{t-1} \in M$, let $\nu(t, x_0, \dots, x_{t-1}, \cdot)$ be the conditional law of X_t given $X_0 = x_0, \dots, X_{t-1} = x_{t-1}$. Now we construct the coupling of X and B as follows. Set $X_0 = B_0 = x_0$ and given $(X_1, B_1), \dots, (X_{t-1}, B_{t-1})$ set (X_t, B_t) to be a coupling of $\nu(t, X_0, \dots, X_{t-1}, \cdot)$ and $p(B_{t-1}, \cdot)$ which is optimal for W_1. Then by construction the marginals of this process coincide with the original processes $\{X_t\}$ and $\{B_t\}$.

The next lemma follows from the coarse Ricci curvature property and the definition of our coupling.

Lemma 2.2 *For every $t \in \{1, \dots, T\}$,*

$$\mathbb{E}_{t-1}\left[d(X_t, B_t)\right] \le \sqrt{C \cdot D(\nu(t, X_0, \dots, X_{t-1}, \cdot) \parallel p(X_{t-1}, \cdot))} + \left(1 - \frac{1}{\alpha}\right) d(X_{t-1}, B_{t-1})$$

where $\mathbb{E}_{t-1}[\cdot]$ stands for the conditional expectation given $(X_0, B_0), \dots, (X_{t-1}, B_{t-1})$.

Proof By definition of the coupling, the triangle inequality for W_1, the one-step transport inequality (2) and the curvature condition

$$\begin{aligned}
&\mathbb{E}_{t-1}\left[d(X_t, B_t)\right] \\
&\quad = W_1\left(\nu(t, X_0, \dots, X_{t-1}, \cdot), p(B_{t-1}, \cdot)\right) \\
&\quad \le W_1\left(\nu(t, X_0, \dots, X_{t-1}, \cdot), p(X_{t-1}, \cdot)\right) + W_1\left(p(X_{t-1}, \cdot), p(B_{t-1}, \cdot)\right) \\
&\quad \le \sqrt{C \cdot D(\nu(t, X_0, \dots, X_{t-1}, \cdot) \parallel p(X_{t-1}, \cdot))} + \left(1 - \frac{1}{\alpha}\right) d(X_{t-1}, B_{t-1}) . \qquad \square
\end{aligned}$$

Remark that the chain rule for relative entropy asserts that

$$\sum_{t=1}^{T} \mathbb{E}[D(\nu(t, X_0, \dots, X_{t-1}, \cdot) \parallel p(X_{t-1}, \cdot))] = D(\{X_0, X_1, \dots, X_T\} \parallel \{B_0, B_1, \dots, B_T\}) . \tag{5}$$

Using the preceding lemma inductively and then Cauchy-Schwarz yields

$$\begin{aligned}
\mathbb{E}[d(X_T, B_T)] &\le \sum_{t=1}^{T} \left(1 - \frac{1}{\alpha}\right)^{T-t} \mathbb{E}\left[\sqrt{C \cdot D(\nu(t, X_0, \dots, X_{t-1}, \cdot) \parallel p(X_{t-1}, \cdot))}\right] \\
&\le \sqrt{\sum_{t=1}^{T} \left(1 - \frac{1}{\alpha}\right)^{2(T-t)}} \sqrt{\sum_{t=1}^{T} C \cdot \mathbb{E}\left[D(\nu(t, X_0, \dots, X_{t-1}, \cdot) \parallel p(X_{t-1}, \cdot))\right]} \\
&\overset{(5)}{\le} \sqrt{\frac{\alpha}{2 - 1/\alpha}} \sqrt{C \cdot D(\{X_0, X_1, \dots, X_T\} \parallel \{B_0, B_1, \dots, B_T\})},
\end{aligned}$$

completing the proof of Proposition 2.1.

2.2.2 The Entropy-Optimal Drift Process

Our goal in this section is to construct a process $x_0 = X_0, X_1, \dots, X_T$ satisfying equation (4). Suppose that we are given a measure ν on Ω along with an initial point $x_0 \in \Omega$ and a time $T \geq 1$. We define the *Föllmer drift process* associated to (ν, x_0, T) as the stochastic process $\{X_t\}_{t=0}^T$ defined as follows.

Let μ_T be the law of $B_T(x_0)$ and denote by f the density of ν with respect to μ_T. Note that f is well-defined as long as the support of μ_T is Ω. Now let $\{X_t\}_{t=0}^T$ be the non homogeneous Markov chain on Ω whose transition probabilities at time t are given by

$$q_t(x,y) := \mathbb{P}(X_t = y) \mid X_{t-1} = x) = \frac{P_{T-t}f(y)}{P_{T-t+1}f(x)}\, p(x,y). \tag{6}$$

We will take care in what follows to ensure the denominator does not vanish. Note that (q_t) is indeed a transition matrix as

$$\sum_{y\in\Omega} P_{T-t}f(y)\, p(x,y) = P_{T-t+1}f(x). \tag{7}$$

We now state a key property of the drift.

Lemma 2.3 *If* $p^T(x_0, x) > 0$ *for all* $x \in \operatorname{supp}(\mu)$*, then* $\{X_t\}$ *is well-defined. Furthermore, for every* $x_1, \dots c, x_T \in \Omega$ *we have*

$$\mathbb{P}\left((X_1, \dots, X_T) = (x_1, \dots, x_T)\right) = \mathbb{P}\left((B_1, \dots, B_T) = (x_1, \dots, x_T)\right) f(x_T). \tag{8}$$

In particular X_T *has law* $d\nu = f\, d\mu_T$.

Proof By definition of the process (X_t) we have

$$\begin{aligned}
\mathbb{P}\left((X_1, \dots, X_T) {=} (x_1, \dots, x_T)\right) &= \prod_{t=1}^{T} \mathbb{P}\left(X_t = x_t \mid (X_1, \dots, X_{t-1}) = (x_1, \dots, x_{t-1})\right) \\
&= \prod_{t=1}^{T} \frac{P_{T-t}f(x_t)}{P_{T-t+1}f(x_{t-1}} p(x_{t-1}, x_t) \\
&= \frac{f(x_T)}{P_T f(x_0)} \left(\prod_{t=1}^{T} p(x_{t-1}, x_t) \right),
\end{aligned}$$

which is the result. □

In words, the preceding lemma asserts that the law of the process $\{X_t\}$ has density $f(x_T)$ with respect to the law of the process $\{B_t\}$. As a result we have in particular

$$D(\{X_0, X_1, \dots, X_T\} \parallel \{B_0, B_1, \dots, B_T\}) = \mathbb{E}[\log f(X_T)] = D(\nu \parallel \mu_T)\,,$$

since X_T has law μ. Note that for any other process $\{Y_t\}$ such that $Y_0 = x_0$ and Y_T has law ν, one always has the inequality

$$D(\{Y_0, Y_1, \dots, Y_T\} \,\|\, \{B_0, B_1, \dots, B_T\}) \geq D(Y_T \,\|\, B_T) = D(\nu \,\|\, \mu_T)\,. \tag{9}$$

Besides $\{X_t\}$ is the unique random process for which this inequality is tight. Uniqueness follows from strict convexity of the relative entropy.

We summarize this section in the following lemma.

Lemma 2.4 *Let (Ω, p) be a Markov chain. Fix $x_0 \in \Omega$ and let $x_0 = B_0, B_1, \dots$ be the associated random walk. Let ν be a measure on Ω and let $T > 0$ be such that for any $y \in \Omega$ one has that $\mathbb{P}(B_T = y) > 0$. Then there exists a process $x_0 = X_0, X_1, \dots, X_T$ such that:*

- *$\{X_0, \dots, X_T\}$ is a (time inhomogeneous) Markov chain.*
- *X_T is distributed with the law ν.*
- *The process satisfies Eq. (4), namely*

$$D(X_T \,\|\, B_T) = D(\{X_0, X_1, \dots, X_T\} \,\|\, \{B_0, B_1, \dots, B_T\})\,.$$

2.3 Finishing up the Proof

Fix an arbitrary $x_0 \in \Omega$ and consider some $T \geq \mathrm{diam}(\Omega, d)$. Let $\{X_t\}$ be the Föllmer drift process associated to the initial data (ν, x_0, T). Then combining Proposition 2.1 and Eq. (4), we have

$$W_1(\nu, \mu_T) = \mathbb{E}[d(X_T, B_T)] \leq \sqrt{\frac{C\alpha}{2 - 1/\alpha} D(\nu \,\|\, \mu_T)}. \tag{10}$$

Now let $T \to \infty$ so that $\mu_T \to \pi$, yielding the desired claim.

3 The Peres–Tetali Conjecture and Log-Sobolev Inequalities

Recall that $p : \Omega \times \Omega \to [0, 1]$ is a transition kernel on the finite state space Ω with a unique stationary measure π. Let $L^2(\Omega, \pi)$ denote the space of real-valued functions $f : \Omega \to \mathbb{R}$ equipped with the inner product $\langle f, g\rangle = \mathbb{E}_\pi[fg]$. From now on we assume that the measure π is reversible, which amounts to saying that the operator $f \mapsto pf$ is self-adjoint in $L^2(\Omega, \pi)$.

We define the associated Dirichlet form

$$\mathcal{E}(f, g) = \langle f, (p - I)g\rangle = \tfrac{1}{2} \sum_{x,y\in\Omega} \pi(x)p(x, y)(f(x) - f(y))(g(x) - g(y))\,.$$

Recall the definition of the entropy of a function $f : \Omega \to \mathbb{R}_+$:

$$\mathrm{Ent}_\pi(f) = \mathbb{E}_\pi\left[f \log\left(\frac{f}{\mathbb{E}_\pi f}\right)\right].$$

Now define the quantities

$$\rho = \inf_{f:\Omega\to\mathbb{R}_+} \frac{\mathcal{E}(\sqrt{f}, \sqrt{f})}{\mathrm{Ent}_\pi(f)}$$

$$\rho_0 = \inf_{f:\Omega\to\mathbb{R}_+} \frac{\mathcal{E}(f, \log f)}{\mathrm{Ent}_\pi(f)}.$$

These numbers are called, respectively, the *log-Sobolev* and *modified log-Sobolev* constants of the chain (Ω, p). We refer to [13] for a detailed discussion of such inequalities on discrete-space Markov chains and their relation to mixing times.

One can understand both numbers as measuring the rate of convergence to equilibrium in appropriate senses. The modified log-Sobolev constant, in particular, can be equivalently characterized as the largest value ρ_0 such that

$$\mathrm{Ent}_\pi(H_t f) \le e^{-\rho_0 t}\mathrm{Ent}_\pi(f) \tag{11}$$

for all $f : \Omega \to \mathbb{R}_+$ and $t > 0$ (see [13, Prop. 1.7]). Here, $H_t : L^2(\Omega, \pi) \to L^2(\Omega, \pi)$ is the heat-flow operator associated to the *continuous-time* random walk, i.e., $H_t = e^{-t(I-P)}$, where P is the operator defined by $Pf(x) = \sum_{y\in\Omega} p(x, y)f(y)$.

The log-Sobolev constant ρ controls the hypercontractivity of the semigroup (H_t), which in turn yields a stronger notion of convergence to equilibrium; again see [13] for a precise statement. Interestingly, in the setting of diffusions, there is no essential distinction between the two notions; one should consider the following calculation only in a formal sense:

$$\mathcal{E}(f, \log f) = \int \nabla f \nabla \log f = \int \frac{|\nabla f|^2}{f} = 4\int \left|\nabla\sqrt{f}\right|^2 = 4\,\mathcal{E}\left(\sqrt{f}, \sqrt{f}\right).$$

However, in the discrete-space setting, the tools of differential calculus are not present. Indeed, one has the bound $\rho \le 2\rho_0$ [13, Prop 1.10], but there is no uniform bound on ρ_0 in terms of ρ.

3.1 MLSI and Curvature

We are now in position to state an important conjecture linking curvature and the modified log-Sobolev constant; it asserts that on spaces with positive coarse Ricci curvature, the random walk should converge to equilibrium exponentially fast in the relative entropy distance.

Conjecture 3.1 (Peres–Tetali, unpublished) *Suppose* (Ω, p) *corresponds to lazy random walk on a finite graph and d is the graph distance. If* (Ω, p, d) *has coarse Ricci curvature* $\kappa > 0$*, then the modified log-Sobolev constant satisfies*

$$\rho_0 \geq C\kappa\,. \tag{12}$$

where $C > 0$ *is a universal constant.*

A primary reason for our interest in Corollary 1.8 is that, by results of Sammer [16], Conjecture 3.1 implies Corollary 1.8. We suspect that a stronger conclusion should hold in many cases; under stronger assumptions, it should be that one can obtain a lower bound on the (non-modified) log-Sobolev constant $\rho \geq C\kappa$. See, for instance, the beautiful approach of Marton [12] that establishes a log-Sobolev inequality for product spaces assuming somewhat strong contraction properties of the Gibbs sampler.

However, we recall that this cannot hold under just the assumptions of Conjecture 3.1. Indeed, if $G = (V, E)$ is the complete graph on n vertices, it is easy to see that the coarse Ricci curvature κ of the lazy random walk is $1/2$. On the other hand, one can check that the log-Sobolev constant ρ decays asymptotically like $\frac{1}{\log n}$ (use the test function $f = \delta_x$ for some fixed $x \in V$).

3.2 An Entropic Interpolation Formulation of MLSI

We now suggest an approach to Conjecture 3.1 using an entropy-optimal drift process. While we chose to work with discrete-time chains in Sect. 2.2.2, working in continuous-time will allow us more precision in exploring Conjecture 3.1. We will use the notation introduced at the beginning of this section.

A continuous-time drift process Suppose we have some initial data (f, x_0, T) where $x_0 \in \Omega$ and $f : \Omega \to \mathbb{R}_+$ satisfies $\mathbb{E}_\pi[f] = 1$. Let $\{B_t : t \in [0, \infty)\}$ denote the continuous-time random walk with jump rates p on the discrete state space Ω starting from x_0. We let μ_T be the law of B_T and let ν be the probability measure defined by

$$d\nu = \frac{f}{H_T f(x_0)}\, d\mu_T,$$

where (H_t) is the semigroup associated to the jump rates $p(x, y)$. Note that ν is indeed a probability measure as $\int f\, d\mu_T = H_T f(x_0)$ by definition of μ_T.

We now define the continuous-time Föllmer drift process associated to the data (x_0, T, f) as the (time inhomogeneous) Markov chain $\{X_t,\ t \leq T\}$ starting from x_0

and having transition rates at time t given by

$$q_t(x,y) = p(x,y)\frac{H_{T-t}f(y)}{H_{T-t}f(x)}\,. \tag{13}$$

Informally this means that the conditional probability that the process $\{X_t\}$ jumps from x to y between time t and $t + dt$ given the past is $q_t(x,y)dt$. This should be thought as the continuous-time analogue of the discrete Föllmer process defined by (6). We claim that again the law of the process $\{X_t,\ t \le T\}$ has density $f(x_T)/H_Tf(x_0)$ with respect to the law of $\{B_t,\ t \le T\}$. Let us give a brief justification of this claim. Define a new probability measure $\mathbb{Q}$ by setting

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \frac{f(B_T)}{H_Tf(x_0)}. \tag{14}$$

We want to prove that the law of B under $\mathbb{Q}$ coincides with the law of X under $\mathbb{P}$. Let $(\mathcal{F}_t)$ be the natural filtration of the process (B_t), let $t \in [0,T)$ and let $y \in M$. We then have the following computation:

$$\begin{aligned}\mathbb{Q}(B_{t+\Delta t} = y \mid \mathcal{F}_t) &= \frac{\mathbb{E}^{\mathbb{P}}[f(B_T)\,\mathbf{1}_{\{B_{t+\Delta t}=y\}} \mid \mathcal{F}_t]}{\mathbb{E}^{\mathbb{P}}[f(B_T) \mid \mathcal{F}_t]} + o(\Delta t)\\ &= \frac{H_{T-t}f(y)}{H_{T-t}f(B_t)}\,\mathbb{P}(B_{t+\Delta(t)} = y \mid \mathcal{F}_t) + o(\Delta t)\\ &= \frac{H_{T-t}f(y)}{H_{T-t}f(B_t)}\,p(B_t,y)\,\Delta t + o(\Delta t).\end{aligned}$$

This shows that under $\mathbb{Q}$, the process $\{B_t,\ t \le T\}$ is Markovian (non homogeneous) with jump rates at time t given by (13). Hence the claim.

This implies in particular that X_T has law ν. This also yields the following formula for the relative entropy of $\{X_t\}$:

$$D(\{X_t,\ t \le T\} \parallel \{B_t,\ t \le T\}) = \mathbb{E}\left[\log\frac{f(X_T)}{H_Tf(x_0)}\right] = D(\nu \parallel \mu_T)\,. \tag{15}$$

The process $\{X_t\}$ starts from x_0 and has law ν at time T. Because X_T has law ν and B_T has law μ_T, the two processes must evolve differently. One can think of the process $\{X_t\}$ as "spending information" in order to achieve the discrepancy between X_T and B_T. The amount of information spent must at least account for the difference in laws at the endpoint, i.e.,

$$D(\{X_t,\ t \le T\} \parallel \{B_T,\ t \le T\}) \ge D(X_T \parallel B_T)\,.$$

As pointed out in Sect. 2.2.2, the content of (15) is that $\{X_t\}$ spends exactly this minimum amount.

For $0 \leq s \leq s'$, we use the notations $B_{[s,s']} = \{B_t : t \in [s,s']\}$ and $X_{[s,s']} = \{X_t : t \in [s,s']\}$ for the corresponding trajectories. From the definition of $\mathbb{Q}$ we easily get

$$\frac{d\mathbb{Q}}{d\mathbb{P}}\bigg|_{\mathcal{F}_t} = \mathbb{E}\left[\frac{f(B_T)}{H_T f(x_0)} \mid \mathcal{F}_t\right] = \frac{H_{T-t}f(B_t)}{H_T f(x_0)}. \tag{16}$$

As a result

$$D\big(X_{[0,t]} \,\|\, B_{[0,t]}\big) = \mathbb{E}\left[\log \frac{H_{T-t}f(X_t)}{H_T f(x_0)}\right] \tag{17}$$

for all $t \leq T$. Let us now define the rate of information spent at time t:

$$I_t = \frac{d}{dt} D\big(X_{[0,t]} \,\|\, B_{[0,t]}\big) \ .$$

Intuitively, the entropy-optimal process $\{X_t\}$ will spend progressively more information as t approaches T. Information spent earlier in the process is less valuable (as the future is still uncertain). Let us observe that a formal version of this statement for random walks on finite graphs is equivalent to Conjecture 3.1.

Conjecture 3.2 *Suppose (Ω, p) corresponds to a lazy random walk on a finite graph and d is the graph distance, and that (Ω, p, d) has coarse Ricci curvature $1/\alpha$. Given $f : \Omega \to \mathbb{R}_+$ with $\mathbb{E}_\pi[f] = 1$ and $x_0 \in \Omega$, for all sufficiently large times T, it holds that if $\{X_t : t \in [0,T]\}$ is the associated continuous-time Föllmer drift process process with initial data (f, x_0, T), then*

$$D(X_T \,\|\, B_T) \leq C\alpha I_T \ , \tag{18}$$

where $C > 0$ is a universal constant.

As $T \to \infty$, we have $P_T f(x_0) \to \mathbb{E}_\pi f = 1$ and thus

$$D(X_T \,\|\, B_T) \to \mathrm{Ent}_\pi(f)$$

Moreover, we claim that $I_T \to \mathcal{E}(f, \log f)$ as $T \to \infty$. Together, these show that Conjectures 3.1 and 3.2 are equivalent.

To verify the latter claim, note that from (17) and (16) we have

$$\begin{aligned} D\big(X_{[0,t]} \,\|\, B_{[0,t]}\big) &= \mathbb{E}\left[\log\left(\frac{H_{T-t}f(X_t)}{H_T f(x_0)}\right)\right] \\ &= \mathbb{E}\left[\log\left(\frac{H_{T-t}f(B_t)}{H_T f(x_0)}\right)\frac{H_{T-t}f(B_t)}{H_T f(x_0)}\right] \\ &= \frac{1}{H_T f(x_0)} H_t\left(H_{T-t}f \log H_{T-t}f\right)(x_0) - \log H_T f(x_0). \end{aligned}$$

Differentiating at $t = T$ yields

$$I_T = \frac{1}{H_T f(x_0)} H_T \left(\Delta(f \log f) - (\Delta f)(\log f + 1)\right)(x_0).$$

where $\Delta = I - p$ denotes the generator of the semigroup (H_t). Recall that $\delta_{x_0} H_T$ converges weakly to π, and that by stationarity $\mathbb{E}_\pi \Delta g = 0$ for every function g. Thus

$$\lim_{T \to \infty} I_T = -\mathbb{E}_\pi[(\Delta f) \log f].$$

The latter equals $\mathcal{E}(f, \log f)$ by reversibility, hence the claim.

References

1. D. Bakry, I. Gentil, M. Ledoux, *Analysis and Geometry of Markov Diffusion Operators*. Volume 348 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences] (Springer, Cham, 2014)
2. S.G. Bobkov, F. Götze, Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. J. Funct. Anal. **163**(1), 1–28 (1999)
3. R. Bubley, M.E. Dyer, Path coupling: a technique for proving rapid mixing in markov chains, in *38th Annual Symposium on Foundations of Computer Science, FOCS'97*, Miami Beach, 19–22 Oct 1997, pp. 223–231
4. H. Djellout, A. Guillin, L. Wu, Transportation cost-information inequalities and applications to random dynamical systems and diffusions. Ann. Probab. **32**(3B), 2702–2732 (2004)
5. M. Fathi, Y. Shu, Curvature and transport inequalities for markov chains in discrete spaces (2015, preprint). arXiv:1509.07160
6. H. Föllmer, An entropy approach to the time reversal of diffusion processes, in *Stochastic Differential Systems*, Marseille-Luminy, 1984. Volume 69 of Lecture Notes in Control and Information Sciences (Springer, Berlin, 1985), pp. 156–163
7. H. Föllmer, Time reversal on Wiener space, in *Stochastic Processes—Mathematics and Physics*, Bielefeld, 1984. Volume 1158 of Lecture Notes in Mathematics (Springer, Berlin, 1986), pp. 119–129
8. J. Lehec, Representation formula for the entropy and functional inequalities. Ann. Inst. Henri Poincaré Probab. Stat. **49**(3), 885–899 (2013)
9. C. Léonard, A survey of the Schrödinger problem and some of its connections with optimal transport. Discret. Contin. Dyn. Syst. **34**(4), 1533–1574 (2014)
10. D.A. Levin, Y. Peres, E.L. Wilmer, *Markov Chains and Mixing Times* (American Mathematical Society, Providence, 2009). With a chapter by J.G. Propp and D.B. Wilson
11. K. Marton, Bounding $\overline{d}$-distance by informational divergence: a method to prove measure concentration. Ann. Probab. **24**(2), 857–866 (1996)
12. K. Marton, Logarithmic Sobolev inequalities in discrete product spaces: a proof by a transportation cost distance (2015, Preprint). arXiv:1507.02803
13. R. Montenegro, P. Tetali, Mathematical aspects of mixing times in Markov chains. Found. Trends Theor. Comput. Sci. **1**(3), x+121 (2006)
14. Y. Ollivier, Ricci curvature of Markov chains on metric spaces. J. Funct. Anal. **256**(3), 810–864 (2009)

15. Y. Ollivier, A survey of Ricci curvature for metric spaces and Markov chains, in *Probabilistic Approach to Geometry*. Volume 57 of Advanced Studies in Pure Mathematics (The Mathematical Society of Japan, Tokyo, 2010), pp. 343–381
16. M.D. Sammer, *Aspects of Mass Transportation in Discrete Concentration Inequalities* (ProQuest LLC, Ann Arbor, 2005). Thesis (Ph.D.), Georgia Institute of Technology
17. C. Villani, *Optimal Transport*, Old and new. Volume 338 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences] (Springer, Berlin, 2009).

Bounding Helly Numbers via Betti Numbers

Xavier Goaoc, Pavel Paták, Zuzana Patáková, Martin Tancer, and Uli Wagner

Dedicated to the memory of Jiří Matoušek, wonderful teacher, mentor, collaborator, and friend.

Abstract We show that very weak topological assumptions are enough to ensure the existence of a Helly-type theorem. More precisely, we show that for any non-negative integers b and d there exists an integer $h(b, d)$ such that the following holds. If $\mathcal{F}$ is a finite family of subsets of $\mathbb{R}^d$ such that $\tilde{\beta}_i\left(\bigcap \mathcal{G}\right) \leq b$ for any $\mathcal{G} \subsetneq \mathcal{F}$ and every $0 \leq i \leq \lceil d/2 \rceil - 1$ then $\mathcal{F}$ has Helly number at most $h(b, d)$. Here $\tilde{\beta}_i$ denotes the reduced $\mathbb{Z}_2$-Betti numbers (with singular homology). These topological conditions are sharp: not controlling any of these $\lceil d/2 \rceil$ first Betti numbers allow for families with unbounded Helly number.

Our proofs combine homological non-embeddability results with a Ramsey-based approach to build, given an arbitrary simplicial complex K, some well-behaved chain map $C_*(K) \to C_*(\mathbb{R}^d)$.

An extended abstract of this work was presented at the *31st International Symposium on Computational Geometry* [21].

PP, ZP and MT were partially supported by the Charles University Grant GAUK 421511. ZP was partially supported by the Charles University Grant SVV-2014-260103. ZP and MT were partially supported by the ERC Advanced Grant No. 267165 and by the project CE-ITI (GACR P202/12/G061) of the Czech Science Foundation. UW was partially supported by the Swiss National Science Foundation (grants SNSF-200020-138230 and SNSF-PP00P2-138948). Part of this work was done when XG was affiliated with INRIA Nancy Grand-Est and when MT was affiliated with Institutionen för matematik, Kungliga Tekniska Högskolan, then IST Austria.

X. Goaoc
Université Paris-Est Marne-la-Vallée, Marne-la-Vallée, France

P. Paták
Department of Algebra, Charles University, Prague, Czech Republic

Z. Patáková • M. Tancer (✉)
Department of Applied Mathematics, Charles University, Prague, Czech Republic
e-mail: tancer@kam.mff.cuni.cz

U. Wagner
IST Austria, Klosterneuburg, Austria

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_17

2010 Mathematics Subject Classification: 52A35 (Helly-type theorems and geometric transversal theory) and 55U10 (Simplicial sets and complexes)

1 Introduction

Helly's classical theorem [27] states that a finite family of convex subsets of $\mathbb{R}^d$ must have a point in common if any $d + 1$ of the sets have a point in common. Together with Radon's and Caratheodory's theorems, two other "very finite properties" of convexity, Helly's theorem is a pillar of combinatorial geometry. Along with its variants (e.g. colorful or fractional), it underlies many fundamental results in discrete geometry, from the centerpoint theorem [44] to the existence of weak ε-nets [2] or the (p, q)-theorem [1].

In the contrapositive, Helly's theorem asserts that any finite family of convex subsets of $\mathbb{R}^d$ with empty intersection contains a sub-family of size at most $d + 1$ that already has empty intersection. This inspired the definition of the *Helly number* of a family $\mathcal{F}$ of arbitrary sets. If $\mathcal{F}$ has empty intersection then its Helly number is defined as the size of the largest sub-family $\mathcal{G} \subseteq \mathcal{F}$ with the following properties: $\mathcal{G}$ has empty intersection and any proper sub-family of $\mathcal{G}$ has nonempty intersection; if $\mathcal{F}$ has nonempty intersection then its Helly number is, by convention, 1. With this terminology, Helly's theorem simply states that any finite family of convex sets in $\mathbb{R}^d$ has Helly number at most $d + 1$.

Helly already realized that bounds on Helly numbers independent of the cardinality of the family are not a privilege of convexity: his *topological* theorem [28] asserts that a finite family of open subsets of $\mathbb{R}^d$ has Helly number at most $d + 1$ if the intersection of any sub-family of at most d members of the family is either empty or a *homology cell*.[1] Such *uniform* bounds are often referred to as *Helly-type theorems*. In discrete geometry, Helly-type theorems were found in a variety of contexts, from simple geometric assumptions (*eg.* homothets of a planar convex curve [53]) to more complicated implicit conditions (sets of line intersecting prescribed geometric shapes [10, 23, 56], sets of norms making a given subset of $\mathbb{R}^d$ equilateral [43, Theorem 5], etc.) and several surveys [16, 54, 61] were devoted to this abundant literature. These Helly numbers give rise to similar finiteness properties in other areas, for instance in variants of Whitney's extension problem [48] or the combinatorics of generators of certain groups [18].

[1] By definition, a homology cell is a topological space X all of whose (reduced, singular, integer coefficient) homology groups are trivial, as is the case if $X = \mathbb{R}^d$ or X is a single point. Here and in what follows, we refer the reader to standard textbooks like [26, 42] for further topological background and various topological notions that we leave undefined.

Many Helly numbers are established via ad hoc arguments, and decades sometimes go by before a conjectured bound is effectively proven, as illustrated by Tverberg's proof [56] of a conjecture of Grünbaum [24]. This is true not only for the quantitative question (*what is the best bound?*) but also for the existential question (*is the Helly number uniformly bounded?*); in this example, establishing a first bound [31] was already a matter of decades. Substantial effort was devoted to identify general conditions ensuring bounded Helly numbers, and *topological conditions*, as opposed to more geometric ones like convexity, received particular attention. The general picture that emerges is that requiring that intersections have *trivial* low-dimensional homotopy [35] or have *trivial* high-dimensional homology [11] is sufficient (see below for a more comprehensive account).

1.1 Problem Statement and Results

In this paper, we focus on the existential question and give the following new homological sufficient condition for bounding Helly numbers. Throughout the paper, we consider homology with coefficients[2] in $\mathbb{Z}_2$, and denote by $\tilde{\beta}_i(X)$ the *i*th reduced Betti number (over $\mathbb{Z}_2$) of a space X. Furthermore, we use the notation $\bigcap \mathcal{F} := \bigcap_{U \in \mathcal{F}} U$ as a shorthand for the intersection of a family of sets.

Theorem 1 *For any non-negative integers b and d there exists an integer $h(b, d)$ such that the following holds. If $\mathcal{F}$ is a finite family of subsets of $\mathbb{R}^d$ such that $\tilde{\beta}_i\left(\bigcap \mathcal{G}\right) \leq b$ for any $\mathcal{G} \subsetneq \mathcal{F}$ and every $0 \leq i \leq \lceil d/2 \rceil - 1$ then $\mathcal{F}$ has Helly number at most $h(b, d)$.*

Our proof, which we sketch in Sect. 1.4, hinges on a general principle, which we learned from Matoušek [35] but which already underlies the classical proof of Helly's theorem from Radon's lemma, to derive Helly-type theorems from results of non-embeddability of certain simplicial complexes. The novelty of our approach is to examine these non-embeddability arguments from a homological point of view. This turns out to be a surprisingly effective idea, as homological analogues of embeddings appear to be much richer and easier to build than their homotopic counterparts. More precisely, our proof of Theorem 1 builds on two contributions of independent interest:

- We reformulate some non-embeddability results in homological terms. We obtain a homological analogue of the Van Kampen–Flores Theorem (Corollary 13)

[2] The choice of $\mathbb{Z}_2$ as the ring of coefficient ring has two reasons. On the one hand, we work with the van Kampen obstruction to prove certain non-embeddability results, and the obstruction is naturally defined either for integer coefficients or over $\mathbb{Z}_2$ (it is a torsion element of order two). On the other hand, the Ramsey arguments used in our proof require working over a fixed finite ring of coefficients to ensure a finite number of color classes (cf. Claim 1).

and, as a side-product, a homological version of Radon's lemma (Lemma 15). This is part of a systematic effort to translate various homotopy technique to a more tractable homology setting. It builds on, and extends, previous work on homological minors [58].

- By working with homology rather than homotopy, we can generalize a technique of Matoušek [35] that uses Ramsey's theorem to find embedded structures. In this step, roughly speaking, we construct some auxiliary (chain) map, with certain homological constraints, inductively by increasing the dimension of the preimage complex while decreasing the size of it. This approach turned out to be also useful in a rather different setting, regarding the (non-)embeddability of skeleta of complexes into manifolds [20].

Our method also proves:

- A bound of $d+1$ on the Helly number of any family $\mathcal{F}$ of subsets of $\mathbb{R}^d$ such that $\tilde{\beta}_i\left(\bigcap \mathcal{G}\right) = 0$ for all $\mathcal{G} \subsetneq \mathcal{F}$ and all $i \leq d$ (see Corollary 24), which generalizes Helly's topological theorem as the sets of $\mathcal{F}$ are, for instance, not assumed to be open. (In the original proof, this assumption is crucial and used to ensure that the union of the sets must have trivial homology in dimensions larger than d; this may fail if the sets are not open.)
- A bound of $d + 2$ on the Helly number of any family $\mathcal{F}$ of subsets of $\mathbb{R}^d$ such that $\tilde{\beta}_i\left(\bigcap \mathcal{G}\right) = 0$ for all $\mathcal{G} \subsetneq \mathcal{F}$ but only for $i \leq \lceil d/2 \rceil - 1$ (see Corollary 23).

In both cases the bounds are tight.

Quantitatively, the bound on $h(b,d)$ that we obtain in the general case is very large as it follows from successive applications of Ramsey's theorems. The conditions of Theorem 1 relax the conditions of a Helly-type theorem of Amenta [4] (see the discussion below) for which a lower bound of $b(d + 1)$ is known [33]; a stronger lower bound is possible for $h(b,d)$ (see Example 2) but we consider narrowing this gap further to be outside the scope of the present paper. Qualitatively, Theorem 1 is sharp in the sense that all (reduced) Betti numbers $\tilde{\beta}_i$ with $0 \leq i \leq \lceil d/2 \rceil - 1$ need to be bounded to obtain a bounded Helly number (see Example 3).

Example 2 First, we observe that for every $d \geq 2$ there is a geometric simplicial complex Γ_d with $d+2$ vertices, embedded in $\mathbb{R}^d$, such that every nonempty induced subcomplex L of Γ_d is connected and satisfies $\tilde{\beta}_i(L) = 0$ for $i \neq d-1$ and $\tilde{\beta}_{d-1}(L) \leq 1$.

Indeed, we can take Γ_d to be the stellar subdivision of the d-simplex (i.e., the cone over the boundary of the d-simplex): Among the vertices of Γ_d, $d+1$ of them, say $v_1, \ldots, v_{d+1}$, form a d-simplex, and the last one, say w, is situated in the barycenter of that simplex. The maximal simplices of Γ_d contain w and d of the vertices v_i. Given an induced subcomplex L, either L misses one of the v-vertices, and then L is a k-simplex for some $k \leq d$; or L contains all the v_i, in which case either $L = \Gamma_d$ or L is the boundary of the simplex spanned by the vertices v_i.

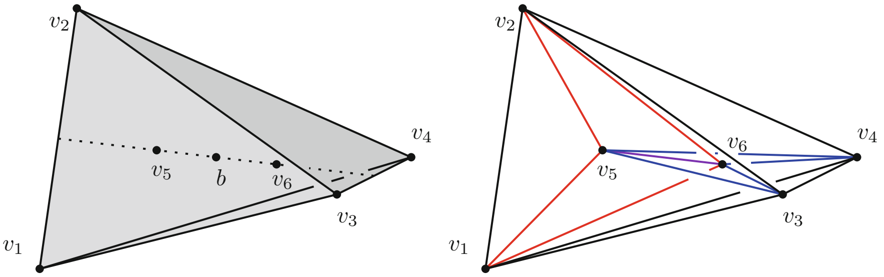


Fig. 1 The simplex Δ (*left*) and the 1-skeleton of Γ'_3 (*right*)

Now, let $\Gamma_{b,d}$ be a complex that consists of b disjoint copies of Γ_d, embedded in $\mathbb{R}^d$. For a vertex v of $\Gamma_{b,d}$, let U_v be the union of all simplices of $\Gamma_{b,d}$ not containing v (i.e., U_v is the geometric realization of the induced subcomplex of $\Gamma_{b,d}$ on all vertices but v). We define $\mathcal{F}$ to be the collection of all subcomplexes F_v, where v ranges over all vertices of $\Gamma_{b,d}$. Thus, by construction, $\mathcal{F}$ contains $b(d+2)$ sets, $\bigcap \mathcal{F} = \emptyset$, and for any nonempty proper subsystem $\mathcal{G} \subset \mathcal{F}$, the intersection $\bigcap \mathcal{G}$ is nonempty, and by the properties of Γ_d, the reduced Betti numbers of $\bigcap \mathcal{G}$ are bounded by b.[3]

Example 3 Let us fix some k with $0 \leq k \leq \lceil d/2 \rceil - 1$. For n arbitrarily large, consider a geometric realization in $\mathbb{R}^d$ of the k-skeleton of the $(n-1)$-dimensional simplex (see [36, Section 1.6]); more specifically, let $V = \{v_1, \ldots, v_n\}$ be a set of points in general position in $\mathbb{R}^d$ and consider all geometric simplices $\sigma_A := \mathrm{conv}(A)$ spanned by subsets $A \subseteq V$ of cardinality $|A| \leq k+1$.

Similarly as in the previous example, let U_j be the union of all the simplices not containing the vertex v_j, for $1 \leq j \leq n$. We set $\mathcal{F} = \{U_1, \ldots, U_n\}$. Then, $\bigcap \mathcal{F} = \emptyset$, and for any proper sub-family $\mathcal{G} \subsetneq \mathcal{F}$, the intersection $\bigcap \mathcal{G}$ is either $\mathbb{R}^d$ (if $\mathcal{G} = \emptyset$) or (homeomorphic to) the k-dimensional skeleton of a $(n-1-|\mathcal{G}|)$-dimensional simplex. Thus, the Helly number of $\mathcal{F}$ equals n. Moreover, the k-skeleton $\Delta^{(k)}_{m-1}$ of an $(m-1)$-dimensional simplex has reduced Betti numbers $\tilde{\beta}_i = 0$ for $i \neq k$ and $\tilde{\beta}_k = \binom{m-1}{k+1}$. Thus, we can indeed obtain arbitrarily large Helly number as soon as at least one $\tilde{\beta}_k$ is unbounded.

[3] We remark that this construction can be further improved (at the cost of simplicity). For example, for $d = 3$, it is possible to find a geometric simplicial complex Γ'_3 with six vertices (instead of five) with properties analogous to Γ_3: Consider a simplex $\Delta \subseteq \mathbb{R}^3$ with vertices v_1, v_2, v_3 and v_4. Let b the barycenter of this simplex and we set v_5 to be the barycenter of the triangle $v_1 v_2 b$ and v_6 to be the barycenter of $v_3 v_4 b$. Finally, we set Γ'_3 to be the subdivision of Δ with vertices $v_1, \ldots, v_6$ and with maximal simplices 1245, 1235, 3416, 3426, 5613, 5614, 5623, and 5624 where the label $ABCD$ stands for $\mathrm{conv}\{v_A, v_B, v_C, v_D\}$. One can check that this indeed yields a simplicial complex with the required properties. See the 1-skeleton of Γ'_3 in Fig. 1. We believe that an analogous example can be also constructed for $d \geq 4$.

1.2 Relation to Previous Work

The search for topological conditions that ensure bounded Helly numbers started with Helly's topological theorem [28] (see also [12] for a modern version of the proof) and organized along several directions related to classical questions in topology. Theorem 1 unifies topological conditions originating from two different approaches:

- Helly-type theorem can be derived from non-embeddability results, in the spirit of the classical proof of Helly's theorem from Radon's lemma. Using this approach, Matoušek [35] showed that it is sufficient to control the *low-dimensional homotopy* of intersections of sub-families to ensure bounded Helly numbers: for any non-negative integers b and d there exists a constant $c(b, d)$ such that any finite family of subsets of $\mathbb{R}^d$ in which every sub-family intersects in at most b connected components, each $(\lceil d/2 \rceil - 1)$ *-connected*, has Helly number at most $c(b, d)$. (We recall that a topological space X is k-connected, for some integer $k \geq 0$, if every continuous map $S^i \to X$ from the i-dimensional sphere to X, $0 \leq i \leq k$, can be extended to a map $D^{i+1} \to X$ from the $(i+1)$-dimensional disk to X.) By Hurewicz' Theorem and the Universal Coefficient Theorem [26, Theorem 4.37 and Corollary 3A.6], a k-connected space X satisfies $\tilde{\beta}_i(X) = 0$ for all $i \leq k$. Thus, our condition indeed relaxes Matoušek's, in two ways: by using $\mathbb{Z}_2$-homology instead of the homotopy-theoretic assumptions of k-connectedness,[4] and by allowing an arbitrary fixed bound b instead of $b = 0$.
- Helly's topological theorem can be easily derived from classical results in algebraic topology relating the homology/homotopy of the nerve of a family to that of its union: Leray's *acyclic cover theorem* [9, Sections III.4.13, VI.4 and VI.13] for homology, and Borsuk's *Nerve theorem* [7, 8] for homotopy (in that case one considers finite open *good covers*[5]). More general Helly numbers were obtained via this approach by Dugundji [15], Amenta [4],[6] Kalai and Meshulam [30], and[7] Colin de Verdière et al. [11]. The outcome is that if a family of subsets of $\mathbb{R}^d$ is such that any sub-family intersects in at most b connected components, each a homology cell (over $\mathbb{Q}$), then it has Helly number at most $b(d+1)$. This therefore relaxes Helly's original assumption by allowing intersections of sub-families to have $\tilde{\beta}_0$'s bounded by an arbitrary fixed

[4] We also remark that our condition can be verified algorithmically since Betti numbers are easily computable, at least for sufficiently nice spaces that can be represented by finite simplicial complexes, say. By contrast, it is algorithmically undecidable whether a given 2-dimensional simplicial complex is 1-connected, see, e.g., the survey [50].

[5] An open good cover is a finite family of open subsets of $\mathbb{R}^d$ such that the intersection of any sub-family is either empty or is contractible (and hence, in particular, a homology cell).

[6] The role of nerves is implicit in Amenta's proof but becomes apparent when compared to an earlier work of Wegner [60] that uses similar ideas.

[7] The result of Colin de Verdière et al. [11] holds in any paracompact topological space; Theorem 1 only subsumes the $\mathbb{R}^d$ case.

bound b instead of $b = 0$. Theorem 1 makes the same relaxation for the $\tilde{\beta}_1$'s, $\tilde{\beta}_2$'s, ... $\tilde{\beta}_{\lceil d/2\rceil-1}$'s and drops *all* assumptions on higher-dimensionnal homology, including the requirement that sets be open (which is used to control the ($> d$)-dimensional homology of intersections).

Let us highlight two Helly-type results that stand out in this line of research as *not* subsumed (qualitatively) by Theorem 1. On the one hand, Eckhoff and Nischke [17] gave a purely combinatorial argument that derives the theorems of Amenta [4] and Kalai and Meshulam [30] from Helly's convex and topological theorems. On the other hand, Montejano [40] relaxed Helly's original assumption on the intersection of sub-families of size $k \le d+1$ from being a homology cell into having trivial $d-k$ homology (so only one Betti number needs to be controlled for each intersection, but it must be zero). These results neither contain nor are contained in Theorem 1.

We remark that another non-topological structural condition, known to ensure bounded Helly numbers, also falls under the umbrella of Theorem 1. As observed by Motzkin [41, Theorem 7] (see also Deza and Frankl [14]), any family of real algebraic subvarieties of $\mathbb{R}^d$ defined by polynomials of degree at most k has Helly number bounded by a function of d and k (more precisely, by the dimension of the vector subspace of $\mathbb{R}[x_1, x_2, \ldots, x_d]$ spanned by these polynomials); since the Betti numbers of an algebraic variety in $\mathbb{R}^n$ can be bounded in terms of the degree of the polynomials that define it [39, 55], this also follows from Theorem 1. We give some other examples in Sect. 1.3, where we easily derive from Theorem 1 generalizations of various existing Helly-type theorems.

Note that Theorem 1 is similar, in spirit, to some of the general relations between the growth of Betti numbers and *fractional* Helly theorems conjectured by Kalai and Meshulam [29, Conjectures 6 and 7]. Kalai and Meshulam, in their conjectures, allow a polynomial growth of the Betti numbers in $|\bigcap \mathcal{G}|$. As the following example shows, Theorem 1 is also sharp in the sense that even a linear growth of Betti numbers, already in $\mathbb{R}^1$, may yield unbounded Helly numbers. In particular, the conjectures of Kalai and Meshulam cannot be strengthened to include Theorem 1.

Example 4 Consider a positive integer n and open intervals $I_i := (i - 1.1; i + 0.1)$ for $i \in [n]$. Let $X_i := [0, n] \setminus I_i$. The intersection of all X_i is empty but the intersection of any proper subfamily is nonempty. In addition, the intersection of k such X_i can be obtained from $[0, n]$ by removing at most k open intervals, thus the reduced Betti numbers of such an intersection are bounded by k.

1.3 Further Consequences

We conclude this introduction with a few implications of our main result.

New geometric Helly-type theorems The main strength of our result is that very weak topological assumptions on families of sets are enough to guarantee a bounded Helly number. This can be used to identify new Helly-type theorems, for instance by easily detecting generalizations of known results, as we now illustrate on two Helly-type theorems of Swanepoel.

A first example is given by a Helly-type theorem for hollow boxes [52], which generalizes (qualitatively) as follows:

Corollary 5 *For all integers $s, d \geq 1$, there exists an integer $h'(s,d)$ such that the following holds. Let S be a set of s nonzero vectors in $\mathbb{R}^d$, and let $\mathcal{F} = \{U_1, U_2, \ldots, U_n\}$ where each U_i is a polyhedral subcomplex of some polytope P_i in $\mathbb{R}^d$ which can be obtained as an intersection of half-spaces with normal vectors in S. Then $\mathcal{F}$ has Helly number at most $h'(s,d)$.*

Swanepoel's result corresponds to the case $S = \{\pm e_1, \pm e_2, \ldots, \pm e_d\}$ where $e_1, \ldots, e_d$ form a basis of $\mathbb{R}^d$.

Proof of Corollary 5 We verify the assumptions of Theorem 1, i.e., we consider a subfamily $\mathcal{G} = \{U_i : i \in I\} \subseteq \mathcal{F}$ and we check that $\tilde{\beta}_i(\bigcap \mathcal{G})$ is bounded by a function of s and d for any $i \geq 0$ (to apply Theorem 1, it would be sufficient to consider $i \leq \lceil d/2 \rceil - 1$, but in the present setting, there is no difference in reasoning for other values of i).

Let $\mathcal{P} = \mathcal{P}(S)$ be the set of all polytopes which can be obtained as an intersection of half-spaces with normal vectors in S. Let $P_i \in \mathcal{P}$ be a polytope such that U_i is a polyhedral subcomplex of P_i.

Let us consider the polytope $P = \bigcap_{i \in I} P_i$. From the definition of $\mathcal{P}$ we immediately deduce that $P \in \mathcal{P}$. Moreover, the intersection $U := \bigcap \mathcal{G}$ is a polyhedral subcomplex of P. (The faces U are of form $\bigcap_{i \in I} \sigma_i$ where σ_i is a face of U_i; see [46, Exercise 2.8(5) + hint].)

Since $P \in \mathcal{P}$ we deduce that it has at most $2s$ facets. By the dual version of the upper bound theorem [62, Theorem 8.23], the number of faces of P is bounded by a function of s and d. Consequently, $\tilde{\beta}_i(U)$ is bounded by a function of s and d, since U is a subcomplex of P. □

A second example concerns a Helly-type theorem for families of translates and homothets of a convex curve [53], which are special cases of families of *pseudo-circles*. More generally, a family of *pseudo-spheres* is defined as a set $\mathcal{F} = \{U_1, U_2, \ldots, U_n\}$ of subsets of $\mathbb{R}^d$ such that or any $\mathcal{G} \subseteq \mathcal{F}$, the intersection $\cap(\mathcal{G})$ is homeomorphic to a k-dimensional sphere for some $k \in \{0, 1, \ldots, d-1\}$ or to a single point. The case $b = 1$ of Theorem 1 immediately implies the following:

Corollary 6 *For any integer d there exists an integer $h(d)$ such that the Helly number of any finite family of pseudo-spheres in $\mathbb{R}^d$ is at most $h(d)$.*

We note that the special case of Euclidean spheres falls under the umbrella of intersections of real algebraic varieties of bounded degree, for which the Helly number is bounded as observed by Motzkin and others, as discussed above [14, 34]. For the more general setting pseudo-spheres, however, the above result is new, to the best of our knowledge. An optimal bound $h(d) = d + 1$ as soon as the family contains at least $d + 3$ pseudo-spheres was obtained by Sosnovec [51], after discussing the contents of Corollary 6 with us.

Generalized linear programming Theorem 1 also has consequences in the direction of optimization problems. Various optimization problems can be formulated as the minimization of some function $f : \mathbb{R}^d \to \mathbb{R}$ over some intersection $\bigcap_{i=1}^n C_i$ of subsets $C_1, C_2, \ldots, C_n$ of $\mathbb{R}^d$. If, for $t \in \mathbb{R}$, we let $L_t = f^{-1}((-\infty, t])$ and $\mathcal{F}_t = \{C_1, C_2, \ldots, C_n, L_t\}$ then

$$\min_{x \in \bigcap_{i=1}^n C_i} f(x) = \min \left\{ t \in \mathbb{R} : \bigcap \mathcal{F}_t \neq \emptyset \right\}.$$

If the Helly number of the families $\mathcal{F}_t$ can be bounded *uniformly* in t by some constant h then there exists a subset of $h - 1$ constraints $C_{i_1}, C_{i_2}, \ldots, C_{i_{h-1}}$ that suffice to define the minimum of f:

$$\min_{x \in \bigcap_{i=1}^n C_i} f(x) = \min_{x \in \bigcap_{j=1}^{h-1} C_{i_j}} f(x).$$

A consequence of this observation, noted by Amenta [3], is that the minimum of f over $C_1 \cap C_2 \cap \ldots \cap C_n$ can[8] be computed in randomized $O(n)$ time by *generalized linear programming* [47] (see de Loera et al. [13] for other uses of this idea). Together with Theorem 1, this implies that an optimization problem of the above form can be solved in randomized linear time if it has the property that every intersection of some subset of the constraints with a level set of the function has bounded "topological complexity" (measured in terms of the sum of the first $\lceil d/2 \rceil$ Betti numbers). Let us emphasize that this linear-time bound holds in a real-RAM model of computation, where any constant-size subproblems can be solved in $O(1)$-time; it therefore concerns the *combinatorial difficulty* of the problem and says nothing about its *numerical difficulty*.

1.4 Proof Outline

Let us briefly sketch the proof of Theorem 1.

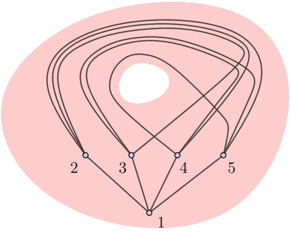
Consider the simplified setting where we have subsets $A_1, A_2, \ldots, A_5$ of $\mathbb{R}^2$ such that any four have non-empty intersection and any three have path-connected intersection. Draw K_5, the complete graph on 5 vertices, *inside* the union of the five sets by picking points $p_i \in \cap_{j \neq i} A_j$ and connecting any two points p_u, p_v inside the intersection $\cap_{j \neq u,v} A_j$. The (stronger form of the) non-planarity of K_5 ensures that two edges that share no vertex must cross, and the intersection point witnesses that $\cap_{i=1}^5 A_i$ is non-empty (cf. Fig. 3). This idea, more systematically, ensures that any family of planar sets with path-connected intersections has Helly number at most 4.

[8] This requires f and $C_1, C_2, \ldots, C_n$ to be generic in the sense that the number of minima of f over $\cap_{i \in I} C_i$ is bounded uniformly for $I \subseteq \{1, 2, \ldots, n\}$.

Now consider subsets $A_1, A_2, \ldots, A_n$ of $\mathbb{R}^2$ such that the intersection of any proper subfamily is nonempty and has at most b path-connected components. We can again pick $p_i \in \cap_{j \neq i} A_j$. Two points p_u, p_v may end up in different connected components of $\cap_{j \neq u,v} A_j$, but among any $b+1$ points $p_{i_1}, p_{i_2}, \ldots, p_{i_{b+1}}$, two can be connected inside $\cap_{j \neq i_1, i_2, \ldots, i_{b+1}} A_j$. We can thus still draw a large graph inside the union, but each edge misses an extra set of A_i's. A Ramsey-type argument ensures that for n large enough, we can find a copy of K_5 where each edge misses distinct extra sets, and therefore that $\cap_{i=1}^n A_i$ is non-empty.

These arguments generalize to higher dimension: once we can draw $p_u p_v$, $p_v p_w$ and $p_u p_w$ inside the intersection of some family of subsets, we can fill the triangle in that intersection if it is 1-connected (in homotopy). More systematically, given a family of subsets of $\mathbb{R}^{2k}$ whose proper intersections are k-connected (in homotopy), we can draw $\Delta_{2k+2}^{(k)}$ inside their union and find, via the Van Kampen-Flores theorem, that the complete intersection is non-empty (and similarly in odd dimensions). This is, in short, Matoušek's theorem [35].

We extend Matoušek's approach to allow intersections to have bounded but non-trivial homotopy in dimension 1 or more. The main difficulty is that we may not be able to fill any elementary cycle: as illustrated on the figure below, for n arbitrarily large, K_n can be drawn in an annulus so that no triangle can be filled. There still exist cycles that can be filled, for instance 2435; they are simply not boundaries of triangles. Such cycles are more easily found by working with the additive structure of $\mathbb{Z}_2$-homology: the sum of any two homologous cycles is a boundary (and therefore "fillable"), and many pairs of homologous cycles exist because a bounded Betti number ensures a constant number of homology classes.



The key idea is, then, to look for sufficiently large sets of vertices where, as in the example above, every triangle has the same $\mathbb{Z}_2$-homology, and to map the *barycentric subdivision* of a triangle to these vertices (as described in Fig. 7); the resulting sum of evenly many homologous simplices must be a boundary. These large sets of vertices with homologous triangles exist as soon as the Betti number is bounded: indeed, one can simply apply Ramsey's theorem to the 3-uniform hypergraph on the vertices where every triangle is "colored" by its homology class. This idea generalizes to arbitrary dimension.

Because of the switch to homology, we do not build a map of $\Delta_{2k+2}^{(k)}$ into the target space $\mathbb{R}^d$ ($d = 2k$ or $d = 2k - 1$) but only a chain map from the simplicial chain complex of $\Delta_{2k+2}^{(k)}$ into the singular chain complex of $\mathbb{R}^d$. Hence, we can no longer rely on the classical non-embeddability results and have to develop homological analogs.

We set up our homological machinery in Sect. 2 (homological almost-embeddings, homological Van Kampen-Flores Theorem, and homological Radon lemma). We then spell out, in Sect. 3, variations of the technique that derives Helly-type theorems from non-embeddability. We finally introduce our refinement of this technique and the proof of Theorem 1 in Sect. 4.

1.5 Notation

We assume that the reader is familiar with basic topological notions and facts concerning simplicial complexes and singular and simplicial homology, as described in textbooks like [26, 42]. As remarked above, throughout this paper we will work with homology with $\mathbb{Z}_2$-coefficients unless explicitly stated otherwise. Moreover, while we will consider singular homology groups for topological spaces in general, for simplicial complexes we will work with simplicial homology groups. In particular, if X is a topological space then $C_*(X)$ will denote the singular chain complex of X, while if K is a simplicial complex, then $C_*(K)$ will denote the simplicial chain complex of K (both with $\mathbb{Z}_2$-coefficients).

We use the following notation. Let K be a (finite, abstract) simplicial complex. The *underlying topological space* of K is denoted by $|K|$. Moreover, we denote by $K^{(i)}$ the *i-dimensional skeleton* of K, i.e., the set of simplices of K of dimension at most i; in particular $K^{(0)}$ is the set of vertices of K. For an integer $n \geq 0$, let Δ_n denote the n-dimensional simplex.

2 Homological Almost-Embeddings

In this section, we define *homological almost-embedding*, an analogue of topological embeddings on the level of chain maps, and show that certain simplicial complexes do not admit homological almost-embeddings in $\mathbb{R}^d$, in analogy to classical non-embeddability results due to Van Kampen and Flores. In fact, when this comes at no additional cost we phrase the auxiliary results in a slightly more general setting, replacing $\mathbb{R}^d$ by a general topological space $\mathbf{R}$. Readers that focus on the proof of Theorem 1 can safely replace every occurrence of **R** with $\mathbb{R}^d$.

2.1 Non-embeddable Complexes

We recall that an *embedding* of a finite simplicial complex K into $\mathbb{R}^d$ is simply an injective continuous map $|K| \to \mathbb{R}^d$. The fact that the complete graph on five vertices cannot be embedded in the plane has the following generalization.

Proposition 7 (Van Kampen [57], Flores [19]) *For $k \geq 0$, the complex $\Delta^{(k)}_{2k+2}$, the k-dimensional skeleton of the $(2k+2)$-dimensional simplex, cannot be embedded in $\mathbb{R}^{2k}$.*

A basic tool for proving the non-embeddability of a simplicial complex is the so-called *Van Kampen obstruction*. To be more precise, we emphasize that in keeping with our general convention regarding coefficients, we work with the $\mathbb{Z}_2$-coefficient version[9] of the Van Kampen obstruction, which will be reviewed in some detail in Sect. 2.3 below. Here, for the benefit of readers who are willing to accept certain topological facts as given, we simply collect those statements necessary to motivate the definition of homological almost-embeddings and to follow the logic of the proof of Theorem 1.

Given a simplicial complex K, one can define, for each $d \geq 0$, a certain cohomology class $\mathfrak{o}^d(K)$ that resides in the cohomology group $H^d(\overline{K})$ of a certain auxiliary complex $\overline{K}$ (the quotient of the combinatorial deleted product by the natural $\mathbb{Z}_2$-action, see below); see the paragraph on obstructions following Lemma 19 for a more proper definition of $\mathfrak{o}^d(K)$. This cohomology class $\mathfrak{o}^d(K)$ is called the Van Kampen obstruction to embeddability into $\mathbb{R}^d$ because of the following fact:

Proposition 8 *Suppose that K is a finite simplicial complex with $\mathfrak{o}^d(K) \neq 0$. Then K is not embeddable into $\mathbb{R}^d$. In fact, a slightly stronger conclusion holds: there is no* almost-embedding *$f \colon |K| \to \mathbb{R}^d$, i.e., no continuous map such that the images of disjoint simplices of K are disjoint.*

Another basic fact is the following result (for a short proof see, for instance, [37, Example 3.5]).

Proposition 9 ([19, 57]) *For every $k \geq 0$, $\mathfrak{o}^{2k}\left(\Delta^{(k)}_{2k+2}\right) \neq 0$.*

As a consequence, one obtains Proposition 7, and in fact the slightly stronger statement that $\Delta^{(k)}_{2k+2}$ does not admit an almost-embedding into $\mathbb{R}^{2k}$.

[9] There is also a version of the Van Kampen obstruction with integer coefficients, which in general yields more precise information regarding embeddability than the $\mathbb{Z}_2$-version, but we will not need this here. We refer to [37] for further background.

2.2 Homological Almost-Embeddings and a Van Kampen–Flores Result

For the proof of Theorem 1, we wish to replace homotopy-theoretic notions (like k-connectedness) by homological assumptions (bounded Betti numbers). The simple but useful observation that allows us to do this is that in the standard proof of Proposition 8, which is based on (co)homological arguments, maps can be replaced by suitable chain maps at every step.[10] The appropriate analogue of an almost-embedding is the following.

Definition 10 Let $\mathbf{R}$ be a (nonempty) topological space, K be a simplicial complex, and consider a chain map[11] $\gamma\colon C_*(K) \to C_*(\mathbf{R})$ from the simplicial chains in K to singular chains in $\mathbf{R}$.

(i) The chain map γ is called *nontrivial*[12] if the image of every vertex of K is a finite set of points in $\mathbf{R}$ (a 0-chain) of *odd* cardinality.

(ii) The chain map γ is called a *homological almost-embedding* of a simplicial complex K in $\mathbf{R}$ if it is nontrivial and if, additionally, the following holds: whenever σ and τ are disjoint simplices of K, their image chains $\gamma(\sigma)$ and $\gamma(\tau)$ have disjoint supports, where the support of a chain is the union of (the images of) the singular simplices with nonzero coefficient in that chain.

Remark 11 Suppose that $f\colon |K| \to \mathbb{R}^d$ is a continuous map.

(i) The induced chain map[13] $f_\sharp\colon C_*(K) \to C_*(\mathbb{R}^d)$ is nontrivial.

(ii) If f is an almost-embedding then the induced chain map is a homological almost-embedding.

Moreover, note that without the requirement of being nontrivial, we could simply take the constant zero chain map, for which the second requirement is trivially satisfied.

We have the following analogue of Proposition 8 for homological almost-embeddings.

[10] This observation was already used in [58] to study the (non-)embeddability of certain simplicial complexes. What we call a *homological almost-embedding* in the present paper corresponds to the notion of a *homological minor* used in [58].

[11] We recall that a chain map $\gamma\colon C_* \to D_*$ between chain complexes is simply a sequence of homomorphisms $\gamma_n\colon C_n \to D_n$ that commute with the respective boundary operators, $\gamma_{n-1} \circ \partial_C = \partial_D \circ \gamma_n$.

[12] If we consider augmented chain complexes with chain groups also in dimension -1, then being nontrivial is equivalent to requiring that the generator of $C_{-1}(K) \cong \mathbb{Z}_2$ (this generator corresponds to the empty simplex in K) is mapped to the generator of $C_{-1}(\mathbf{R}) \cong \mathbb{Z}_2$.

[13] The induced chain map is defined as follows: We assume that we have fixed a total ordering of the vertices of K. For a p-simplex σ of K, the ordering of the vertices induces a homeomorphism $h_\sigma\colon |\Delta_p| \to |\sigma| \subseteq |K|$. The image $f_\sharp(\sigma)$ is defined as the singular p-simplex $f \circ h_\sigma$.

Proposition 12 *Suppose that K is a finite simplicial complex with $\mathfrak{o}^d(K) \neq 0$. Then K does not admit a homological almost-embedding in $\mathbb{R}^d$.*

As a corollary, we get the following result, which underlies our proof of Theorem 1.

Corollary 13 *For any $k \geq 0$, the k-skeleton $\Delta_{2k+2}^{(k)}$ of the $(2k+2)$-dimensional simplex has no homological almost-embedding in $\mathbb{R}^{2k}$.*

We conclude this subsection by two facts that are not needed for the proof of the main result but are useful for the presentation of our method in Sect. 3.

If the ambient dimension $d = 2k + 1$ is odd, we can immediately see that $\Delta_{2k+4}^{(k+1)}$ has no homological almost-embedding in $\mathbb{R}^{2k+1}$ since it has no homological almost-embedding in $\mathbb{R}^{2k+2}$; this result can be slightly improved:

Corollary 14 *For any $d \geq 0$, the $\lceil d/2 \rceil$-skeleton $\Delta_{d+2}^{(\lceil d/2 \rceil)}$ of the $(d+2)$-dimensional simplex has no homological almost-embedding in $\mathbb{R}^d$.*

Proof The statement for even d is already covered by the case $k = d/2$ of Corollary 13, so assume that d is odd and write $d = 2k + 1$. If K is a finite simplicial complex with $\mathfrak{o}^d(K) \neq 0$ and if CK is the cone over K then $\mathfrak{o}^{d+1}(CK) \neq 0$ (for a proof, see, for instance, [6, Lemma 8]). Since we know that $\mathfrak{o}^{2k}(\Delta_{2k+2}^{(k)}) \neq 0$ it follows that $\mathfrak{o}^{2k+1}(C\Delta_{2k+2}^{(k)}) \neq 0$. Consequently, $\mathfrak{o}^{2k+1}(\Delta_{2k+3}^{(k+1)}) \neq 0$ since $C\Delta_{2k+2}^{(k)}$ is a subcomplex of $\Delta_{2k+3}^{(k+1)}$ and there exists an equivariant map from the deleted product of the subcomplex to the deleted product of the complex. Proposition 12 then implies that $\Delta_{2k+3}^{(k+1)}$ admits no homological almost-embedding in $\mathbb{R}^{2k+1}$. □

The next fact is the following analogue of Radon's lemma, proved in the next subsection along the proof of Proposition 12.

Lemma 15 (Homological Radon's lemma) *For any $d \geq 0$, $\mathfrak{o}^d(\partial\Delta_{d+1}) \neq 0$. Consequently, the boundary of $(d+1)$-simplex $\partial\Delta_{d+1}$ admits no homological almost-embedding in $\mathbb{R}^d$.*

2.3 Deleted Products and Obstructions

Here, we review the standard proof of Proposition 8 and explain how to adapt it to prove Proposition 12, which will follow from Lemma 19 and Lemma 20(b) below. The reader unfamiliar with cohomology and willing to accept Proposition 12 can safely proceed to Sect. 3.

$\mathbb{Z}_2$-spaces and equivariant maps We begin by recalling some basic notions of equivariant topology: An *action* of the group $\mathbb{Z}_2$ on a space X is given by an automorphism $\nu: X \to X$ such that $\nu \circ \nu = 1_X$; the action is *free* if ν does not have any fixed points. If X is a simplicial complex (or a cell complex), then the action is called simplicial (or cellular) if it is given by a simplicial (or cellular) map. A space with a given (free) $\mathbb{Z}_2$-action is also called a (free) $\mathbb{Z}_2$-space.

A map $f: X \to Y$ between $\mathbb{Z}_2$-spaces (X, ν) and (Y, μ) is called *equivariant* if it commutes with the respective $\mathbb{Z}_2$-actions, i.e., $f \circ \nu = \mu \circ f$. Two equivariant maps $f_0, f_1: X \to Y$ are *equivariantly homotopic* if there exists a homotopy $F: X \times [0, 1] \to Y$ such that all intermediate maps $f_t := F(\cdot, t)$, $0 \leq t \leq 1$, are equivariant.

A $\mathbb{Z}_2$-action ν on a space X also yields a $\mathbb{Z}_2$-action on the chain complex $C_*(X)$, given by the induced chain map $\nu_\sharp: C_*(X) \to C_*(X)$ (if ν is simplicial or cellular, respectively, then this remains true if we consider the simplicial or cellular chain complex of X instead of the singular chain complex), and if $f: X \to Y$ is an equivariant map between $\mathbb{Z}_2$-spaces then the induced chain map is also equivariant (i.e., it commutes with the $\mathbb{Z}_2$-actions on the chain complexes).

Spheres Important examples of free $\mathbb{Z}_2$-spaces are the standard spheres $\mathbb{S}^d$, $d \geq 0$, with the action given by antipodality, $x \mapsto -x$. There are natural inclusion maps $\mathbb{S}^{d-1} \hookrightarrow \mathbb{S}^d$, which are equivariant. Antipodality also gives a free $\mathbb{Z}_2$-action on the union $\mathbb{S}^\infty = \bigcup_{d \geq 0} \mathbb{S}^d$, the infinite-dimensional sphere. Moreover, one can show that $\mathbb{S}^\infty$ is contractible, and from this it is not hard to deduce that $\mathbb{S}^\infty$ is a universal $\mathbb{Z}_2$-space, in the following sense (see [38] or also [32, Prop. 8.16 and Thm. 8.17] for a more detailed textbook treatment).

Proposition 16 *If X is any cell complex with a free cellular $\mathbb{Z}_2$-action, then there exists an equivariant map $f: X \to \mathbb{S}^\infty$. Moreover, any two equivariant maps $f_0, f_1: X \to \mathbb{S}^\infty$ are equivariantly homotopic.*

Any equivariant map $f: X \to \mathbb{S}^\infty$ induces a nontrivial equivariant chain map $f_\sharp: C_*(X) \to C_*(\mathbb{S}^\infty)$. A simple fact that will be crucial in what follows is that Proposition 16 has an analogue on the level of chain maps.

We first recall the relevant notion of homotopy between chain maps: Let $C_*(X)$ and $C_*(Y)$ be (singular or simplicial, say) chain complexes, and let $\varphi, \psi: C_*(X) \to C_*(Y)$ be chain maps. A *chain homotopy* η between φ and ψ is a family of homomorphisms $\eta_j: C_j(X) \to C_{j+1}(Y)$ such that

$$\varphi_j - \psi_j = \partial^Y_{j+1} \circ \eta_j + \eta_{j-1} \circ \partial^X_j$$

for all j.[14] If X and Y are $\mathbb{Z}_2$-spaces then a chain homotopy is called equivariant if it commutes with the (chain maps induced by) the $\mathbb{Z}_2$-actions.[15]

Lemma 17 *If X is a cell complex with a free cellular $\mathbb{Z}_2$-action then any two nontrivial equivariant chain maps $\varphi, \psi: C_*(X) \to C_*(\mathbb{S}^\infty)$ are equivariantly chain homotopic.*[16]

[14] Here, we use subscripts and superscripts on the boundary operators to emphasize which dimension and which chain complex they belong to; often, these indices are dropped and one simply writes $\varphi - \psi = \partial\eta + \eta\partial$.

[15] We also recall that if $f, g\ X \to Y$ are (equivariantly) homotopic then the induced chain maps are (equivariantly) chain homotopic. Moreover, chain homotopic maps induce *identical* maps in homology and cohomology.

[16] We stress that we work with the cellular chain complex for X.

Proof of Lemma 17 Let the $\mathbb{Z}_2$-action on X be given by the automorphism $\nu\colon X \to X$. For each dimension $i \geq 0$, the action partitions the i-dimensional cells of X (the basis elements of $C_i(X)$) into pairs $\sigma, \nu(\sigma)$. For each such pair, we arbitrarily pick one of the cells and call it the representative of the pair.

We define the desired equivariant chain homotopy η between φ and ψ by induction on the dimension, using the fact that all reduced homology groups of $\mathbb{S}^\infty$ are zero. (This just mimics the argument for the existence of an equivariant homotopy, which uses the contractibility of $\mathbb{S}^\infty$.)

We start the induction in dimension at $j = -1$ (and for convenience, we also use the convention that all chain groups, chain maps, and η_i are understood to be zero in dimensions $i < -1$). Since we assume that both φ and ψ are nontrivial, we have that $\varphi_{-1}, \psi_{-1}\colon C_{-1}(X) \to C_{-1}(\mathbb{S}^\infty)$ are identical, and we set $\eta_{-1}\colon C_{-1}(X) \to C_0(\mathbb{S}^\infty)$ to be zero.

Next, assume inductively that equivariant homomorphisms $\eta_i\colon C_i(X) \to C_i(\mathbb{S}^\infty)$ have already been defined for $i < j$ and satisfy

$$\varphi_i - \psi_i = \eta_{i-1} \circ \partial + \partial \circ \eta_i \tag{1}$$

for all $i < j$ (note that initially, this holds true for $j = 0$).

Suppose that σ is a j-dimensional cell of X representing a pair $\sigma, \nu(\sigma)$. Then $\partial\sigma \in C_{j-1}(X)$, and so $\eta_{j-1}(\partial\sigma) \in C_j(\mathbb{S}^\infty)$ is already defined. We are looking for a suitable chain $c \in C_{j+1}(\mathbb{S}^\infty)$ which we can take to be $\eta_j(\sigma)$ in order to satisfy the chain homotopy relation (1) also for $i = j$, such a chain c has to satisfy $\partial c = b$, where

$$b := \varphi_j(\sigma) - \psi_j(\sigma) - \eta_{j-1}(\partial(\sigma)).$$

To see that we can find such a c, we compute

$$\begin{aligned}\partial b &= \partial\varphi_j(\sigma) - \partial\psi_j(\sigma) - \partial\eta_{j-1}(\partial(\sigma))\\ &= \varphi_{j-1}(\partial\sigma) - \psi_{j-1}(\partial\sigma) - \Big(\varphi_{j-1}(\partial\sigma) - \psi_{j-1}(\partial\sigma) - \eta_{j-2}(\partial\partial\sigma)\Big) = 0\end{aligned}$$

Thus, b is a cycle, and since $H_j(\mathbb{S}^\infty) = 0$, b is also a boundary. Pick an arbitrary chain $c \in C_{j+1}(\mathbb{S}^\infty)$ with $\partial c = b$ and set $\eta_j(\sigma) := c$ and $\eta_j(\nu(\sigma)) := \nu_\sharp(c)$. We do this for all representative j-cells σ and then extend η_j by linearity. By definition, η_j is equivariant and (1) is now satisfied also for $i = j$. This completes the induction step and hence the proof. □

Deleted products and Gauss maps Let K be a finite simplicial complex. Then the Cartesian product $K \times K$ is a cell complex whose cells are the Cartesian products of pairs of simplices of K. The (*combinatorial*) *deleted product* $\widetilde{K}$ of K is defined as the polyhedral subcomplex of $K \times K$ whose cells are the products of vertex-disjoint pairs of simplices of K, i.e., $\widetilde{K} := \{\sigma \times \tau : \sigma, \tau \in K, \sigma \cap \tau = \emptyset\}$. The deleted product is equipped with a natural free $\mathbb{Z}_2$-action that simply exchanges coordinates,

$(x, y) \mapsto (y, x)$. Note that this action is cellular since each cell $\sigma \times \tau$ is mapped to $\tau \times \sigma$.

Lemma 18 *If $f: |K| \hookrightarrow \mathbb{R}^d$ is an embedding (or, more generally, an almost-embedding) then*[17] *there exists an equivariant map $\tilde{f}: \widetilde{K} \to S^{d-1}$.*

Proof Define $\tilde{f}(x, y) := \frac{f(x)-f(y)}{\|f(x)-f(y)\|}$. This map, called the *Gauss map*, is clearly equivariant. □

For the proof of Proposition 12, we use the following analogue of Lemma 18.

Lemma 19 *Let K be a finite simplicial complex. If $\gamma: C_*(K) \to C_*(\mathbb{R}^d)$ is a homological almost-embedding then there is a nontrivial equivariant chain map (called the* Gauss chain map*) $\tilde{\gamma}: C_*(\widetilde{K}) \to C_*(\mathbb{S}^{d-1})$.*
The proof of this lemma is not difficult but a bit technical, so we postpone it until the end of this section.

Obstructions Here, we recall a standard method for proving the non-existence of equivariant maps between $\mathbb{Z}_2$-spaces. The arguments are formulated in the language of cohomology, and, as we will see, what they actually establish is the non-existence of nontrivial equivariant chain maps.

Let K be a finite simplicial complex and let $\widetilde{K}$ be its (combinatorial) deleted product. By Proposition 16, there exists an equivariant map $G_K: \widetilde{K} \to \mathbb{S}^\infty$, which is unique up to equivariant homotopy. By factoring out the action of $\mathbb{Z}_2$, this induces a map $\overline{G}_K: \overline{K} \to \mathbb{RP}^\infty$ between the quotient spaces $\overline{K} = \widetilde{K}/\mathbb{Z}_2$ and $\mathbb{RP}^\infty = \mathbb{S}^\infty/\mathbb{Z}_2$ (the infinite-dimensional real projective space), and the homotopy class of the map $\overline{G}_K$ depends only[18] on K. Passing to cohomology, there is a uniquely defined induced homomorphism

$$\overline{G}_K^*: H^*(\mathbb{RP}^\infty) \to H^*(\overline{K}).$$

It is known that $H^d(\mathbb{RP}^\infty) \cong \mathbb{Z}_2$ for every $d \geq 0$. Letting ξ^d denote the unique generator of $H^d(\mathbb{RP}^\infty)$, there is a uniquely defined cohomology class

$$\mathfrak{o}^d(K) := \overline{G}_K^*(\xi^d),$$

called the *van Kampen obstruction* (with $\mathbb{Z}_2$-coefficients) to embedding K into $\mathbb{R}^d$. For more details and background regarding the van Kampen obstruction, we refer the reader to [37].

[17] We remark that a classical result due to Haefliger and Weber [25, 59] asserts that if $\dim K \leq (2d-3)/3$ (the so-called *metastable range*) then the existence of an equivariant map from $\widetilde{K}$ to $\mathbb{S}^{d-1}$ is also *sufficient* for the existence of an embedding $K \hookrightarrow \mathbb{R}^d$ (outside the metastable range, this fails); see [49] for further background.

[18] We stress that this does not mean that there is only one homotopy class of continuous maps $\overline{K} \to \mathbb{RP}^\infty$; indeed, there exist such maps that do not come from equivariant maps $\widetilde{K} \to \mathbb{S}^\infty$, for instance the constant map that maps all of $\overline{K}$ to a single point.

The basic fact about the van Kampen obstruction (and the reason for its name) is that K does not embed (not even almost-embed) into $\mathbb{R}^d$ if $\mathfrak{o}^d(K) \neq 0$ (Proposition 8). This follows from Lemma 18 and Part (a) of the following lemma:

Lemma 20 *Let K be a simplicial complex and suppose that $\mathfrak{o}^d(K) \neq 0$.*

(a) *Then there is no equivariant map $\widetilde{K} \to \mathbb{S}^{d-1}$.*
(b) *In fact, there is no nontrivial equivariant chain map $C_*(\widetilde{K}) \to C_*(\mathbb{S}^{d-1})$.*

Together with Lemma 19, Part (b) of the lemma also implies Proposition 12, as desired. The simple observation underlying the proof of Lemma 20 is the following

Observation 21 *Suppose $\varphi\colon C_*(\widetilde{K}) \to C_*(\mathbb{S}^\infty)$ is a nontrivial equivariant chain map (not necessarily induced by a continuous map). By factoring out the action of $\mathbb{Z}_2$, φ induces a chain map $\overline{\varphi}\colon C_*(\overline{K}) \to C_*(\mathbb{RP}^\infty)$. The induced homomorphism in cohomology*

$$\overline{\varphi}^*\colon H^*(\mathbb{RP}^\infty) \to H^*(\overline{K})$$

is equal *to the homomorphism $\overline{G}_K^*$ used in the definition of the Van Kampen obstruction, hence in particular*

$$\mathfrak{o}^d(K) = \overline{\varphi}^*(\xi^d).$$

Proof By Lemma 17, φ is equivariantly chain homotopic to the nontrivial equivariant chain map $(G_K)_\sharp$ induced by the map G_K. Thus, after factoring out the $\mathbb{Z}_2$-action, the chain maps $\overline{\varphi}$ and $(\overline{G}_K)_\sharp$ from $C_*(\overline{K})$ to $C_*(\mathbb{RP}^\infty)$ are chain homotopic, and so induce identical homomorphisms in cohomology. □

Proof of Lemma 20 If there exists an equivariant map $f\colon\widetilde{K} \to \mathbb{S}^{d-1}$, then the induced chain map $f_\sharp\colon C_*(\widetilde{K}) \to C_*(\mathbb{S}^{d-1})$ is equivariant and nontrivial, so (b) implies (a), and it suffices to prove the former.

Next, suppose for a contradiction that $\psi\colon C_*(\widetilde{K}) \to C_*(\mathbb{S}^{d-1})$ is a nontrivial equivariant chain map. Let $i : \mathbb{S}^{d-1} \to \mathbb{S}^\infty$ denote the inclusion map, and let $i_\sharp\colon C_*(\mathbb{S}^{d-1}) \to C_*(\mathbb{S}^\infty)$ denote the induced equivariant, nontrivial chain map. Then the composition $\varphi = (i_\sharp \circ \psi)\colon C_*(\widetilde{K}) \to C_*(\mathbb{S}^\infty)$ is also nontrivial and equivariant, and so, by the preceding observation, for the induced homomorphism in cohomology, we get

$$\mathfrak{o}^d(K) = \overline{(i_\sharp \circ \psi)}^*(\xi^d) = \overline{\psi}^*\left(\overline{i}^*(\xi^d)\right).$$

However, $\overline{i}^*(\xi^d) \in H^d(\mathbb{RP}^{d-1}) = 0$ (for reasons of dimension), hence $\mathfrak{o}^d(K) = 0$, contradicting our assumption. □

Remark 22 The same kind of reasoning also yields the well-known *Borsuk–Ulam Theorem*, which asserts that there is no equivariant map $\mathbb{S}^d \to \mathbb{S}^{d-1}$, using the fact that the inclusion $\overline{i}\colon\mathbb{RP}^d \to \mathbb{RP}^\infty$ (induced by the equivariant inclusion $i\colon\mathbb{S}^d \to$

$\mathbb{S}^\infty$) has the property that $\bar{i}^*(\xi^d)$, the pullback of the generator $\xi^d \in H^d(\mathbb{RP}^\infty)$, is *nonzero*.[19] In fact, once again one gets a homological version of the Borsuk–Ulam theorem for free: there is no nontrivial equivariant chain map $C_*(\mathbb{S}^d) \to C_*(\mathbb{S}^{d-1})$.

Proof of Lemma 15 It is not hard to see that the deleted product $\widetilde{\partial\Delta_{d+1}} = \widetilde{\Delta_{d+1}}$ of the boundary of $(d+1)$-simplex is combinatorially isomorphic to the boundary of a certain convex polytope and hence homeomorphic to $\mathbb{S}^d$(respecting the antipodality action), see [36, Exercise 5.4.3]. Thus, the assertion $\mathfrak{o}^d(\partial\Delta_{d+1}) \neq 0$ follows immediately from the preceding remark (the homological proof of the Borsuk–Ulam theorem). Together with Proposition 12, this implies that there is no homological almost-embedding of $\partial\Delta_{d+1}$ in $\mathbb{R}^d$. □

The proof of Proposition 12 is complete, except for the following:

Proof of Lemma 19 Once again, we essentially mimic the definition of the Gauss map on the level of chains. There is one minor technical difficulty due to the fact that the cells of $\widetilde{K}$ are products of simplices, whereas the singular homology of spaces is based on maps whose domains are simplices, not products of simplices (this is the same issue that arises in the proof of Künneth-type formulas in homology).

Assume that $\gamma\colon C_*(K) \to C_*(\mathbb{R}^d)$ is a homological almost-embedding. The desired nontrivial equivariant chain map $\widetilde{\gamma}\colon C_*(\widetilde{K}) \to C_*(\mathbb{S}^{d-1})$ will be defined as the composition of three intermediate nontrivial equivariant chain maps

$$\begin{tikzcd} C_*(\widetilde{K}) \arrow[r, "\alpha"] \arrow[rrr, bend right, "\widetilde{\gamma}=p_\sharp\circ\beta\circ\alpha"'] & D_* \arrow[r, "\beta"] & C_*(\widetilde{\mathbb{R}^d}) \arrow[r, "p_\sharp"] & C_*(\mathbb{S}^{d-1}). \end{tikzcd}$$

These maps and intermediate chain complexes will be defined presently.

We define D_* as a chain subcomplex of the tensor product $C_*(\mathbb{R}^d) \otimes C_*(\mathbb{R}^d)$. The tensor product chain complex has a basis consisting of all elements of the form $s \otimes t$, where s and t range over the singular simplices of $\mathbb{R}^d$, and we take D_* as the subcomplex spanned by all $s \otimes t$ for which s and t have disjoint supports (note that D_* is indeed a chain subcomplex, i.e., closed under the boundary operator, since if s and t have disjoint supports, then so do any pair of simplices that appear in the boundary of s and of t, respectively). The chain complex $C_*(\widetilde{K})$ has a canonical basis consisting of cells $\sigma \times \tau$, and the chain map α is defined on these basis elements by "tensoring" γ with itself, i.e.,

$$\alpha(\sigma \times \tau) := \gamma(\sigma) \otimes \gamma(\tau).$$

Since γ is nontrivial, so is α, the disjointness properties of γ ensure that the image of α does indeed lie in D_*, and α is clearly $\mathbb{Z}_2$-equivariant.

[19] In fact, it is known that $H^*(\mathbb{RP}^\infty)$ is isomorphic to the polynomial ring $\mathbb{Z}_2[\xi]$, that $H^*(\mathbb{RP}^d) \cong \mathbb{Z}_2[\xi]/(\xi^{d+1})$, and that $\bar{i}^*$ is just the quotient map.

Next, consider the Cartesian product $\mathbb{R}^d \times \mathbb{R}^d$ with the natural $\mathbb{Z}_2$-action given by flipping coordinates. This action is not free since it has a nonempty set of fixed points, namely the "diagonal" $\Delta = \{(x,x) : x \in \mathbb{R}^d\}$. However, the action on $\mathbb{R}^d \times \mathbb{R}^d$ restricts to a free action on the subspace $\widetilde{\mathbb{R}^d} := (\mathbb{R}^d \times \mathbb{R}^d) \setminus \Delta$ obtained by removing the diagonal (this subspace is sometimes called the topological deleted product of $\mathbb{R}^d$). Moreover, there exists an equivariant map $p\colon \widetilde{\mathbb{R}^d} \to \mathbb{S}^{d-1}$ defined as follows: we identify $\mathbb{S}^{d-1}$ with the unit sphere in the orthogonal complement $\Delta^\perp = \{(w,-w) \in \mathbb{R}^d : w \in \mathbb{R}^d\}$ and take $p\colon \widetilde{\mathbb{R}^d} \to \mathbb{S}^{d-1}$ to be the orthogonal projection onto $\Delta^\perp$ (which sends (x,y) to $\frac{1}{2}(x-y, y-x)$), followed by renormalizing,

$$p(x,y) := \frac{\frac{1}{2}(x-y, y-x)}{\|\frac{1}{2}(x-y, y-x)\|} \in \mathbb{S}^{d-1} \subset \Delta^\perp.$$

The map p is equivariant and so the induced chain map $p_\sharp$ is equivariant and nontrivial.

It remains to define $\beta\colon D_* \to C_*(\widetilde{\mathbb{R}^d})$. For this, we use a standard chain map

$$\mathrm{EML}\colon C_*(\mathbb{R}^d) \otimes C_*(\mathbb{R}^d) \to C_*(\mathbb{R}^d \times \mathbb{R}^d),$$

sometimes called the Eilenberg–Mac Lane chain map, and then take β to be the restriction to D_*.

Given a basis element $s \otimes t$ of $C_*(\mathbb{R}^d) \otimes C_*(\mathbb{R}^t)$, where $s\colon \Delta_p \to \mathbb{R}^d$ and $t\colon \Delta_q \to \mathbb{R}^d$ are singular simplices, we can view $s \otimes t$ as the map $s \otimes t\colon \Delta_p \times \Delta_q \to \mathbb{R}^d \times \mathbb{R}^d$ with $(x,y) \mapsto (s(x), t(y))$. This is almost like a singular simplex in $\mathbb{R}^d \times \mathbb{R}^d$, except that the domain is not a simplex but a prism (product of simplices). The Eilenberg–Mac Lane chain map is defined by prescribing a systematic and coherent way of triangulating products of simplices $\Delta_p \times \Delta_q$ that is consistent with taking boundaries; then $\mathrm{EML}(s \otimes t) \in C_{p+q}(\mathbb{R}^d \times \mathbb{R}^d)$ is defined as the singular chain whose summands are the restrictions of the map $\sigma \otimes \tau\colon \Delta_p \times \Delta_q$ to the $(p+q)$-simplices that appear in the triangulation of $\Delta_p \times \Delta_q$. We refer to [22] for explicit formulas for the chain map EML. What is important for us is that the chain map EML is equivariant and nontrivial. Both properties follow more or less directly from the construction of the triangulation of the prisms $\Delta_p \times \Delta_q$, which can be explained as follows: Implicitly, we assume that the vertex sets $\{0, 1, \ldots, p\}$ and $\{0, 1, \ldots, q\}$ are totally ordered in the standard way. The vertex set of $\Delta_p \times \Delta_q$ is the grid $\{0, 1, \ldots, p\} \times \{0, 1, \ldots, q\}$, on which we consider the coordinatewise partial order defined by $(x,y) \leq (x',y')$ if $x \leq x'$ and $y \leq y'$. Then the simplices of the triangulation are all totally ordered subsets of this partial order. Thus, if $\sigma = \{(x_0, y_0), (x_1, y_1), \ldots, (x_r, y_r)\}$ is a simplex that appears in the triangulation of $\Delta_p \times \Delta_q$ then the simplex $\sigma = \{(y_0, x_0), (y_1, x_1), \ldots, (y_r, x_r)\}$ obtained by flipping all coordinates appears in the triangulation of $\Delta_q \times \Delta_p$; see Fig. 2. This implies equivariance of EML (and it is nontrivial since it maps a single vertex to a single vertex). □

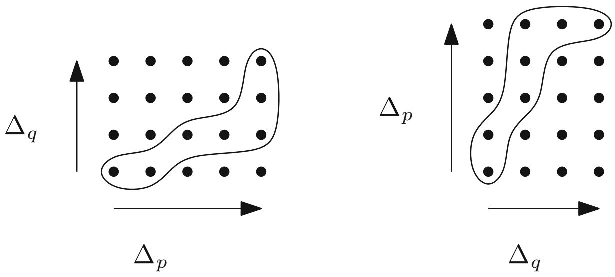


Fig. 2 A simplex in a triangulation of $\Delta_p \times \Delta_q$ and its twin in $\Delta_q \times \Delta_p$

3 Helly-Type Theorems from Non-embeddability

We now detail the technique outlined in Sect. 1.4 and illustrate it on a few examples before formalizing its ingredients.

Notation Given a set X we let 2^X and $\binom{X}{k}$ denote, respectively, the set of all subsets of X (including the empty set) and the set of all k-element subsets of X. If $f : X \to Y$ is an arbitrary map between sets then we abuse the notation by writing $f(S)$ for $\{f(s) \mid s \in S\}$ for any $S \subseteq X$; that is, we implicitly extend f to a map from 2^X to 2^Y whenever convenient.

3.1 Homotopic Assumptions

Let $\mathcal{F} = \{U_1, U_2, \ldots, U_n\}$ denote a family of subsets of $\mathbb{R}^d$. We assume that $\mathcal{F}$ has empty intersection and that any proper subfamily of $\mathcal{F}$ has nonempty intersection. Our goal is to show how various conditions on the topology of the intersections of the subfamilies of $\mathcal{F}$ imply bounds on the cardinality of $\mathcal{F}$. For any (possibly empty) proper subset I of $[n] = \{1, 2, \ldots, n\}$ we write $U_{\overline{I}}$ for $\bigcap_{i \in [n] \setminus I} U_i$. We also put $U_{\overline{[n]}} = \mathbb{R}^d$.

Path-connected intersections in the plane Consider the case where $d = 2$ and the intersections $\bigcap \mathcal{G}$ are path-connected for all subfamilies $\mathcal{G} \subsetneq \mathcal{F}$. Since every intersection of $n - 1$ members of $\mathcal{F}$ is nonempty, we can pick, for every $i \in [n]$, a point p_i in $U_{\overline{\{i\}}}$. Moreover, as every intersection of $n-2$ members of $\mathcal{F}$ is connected, we can connect any pair of points p_i and p_j by an arc $s_{i,j}$ inside $U_{\overline{\{i,j\}}}$. We thus obtain a drawing of the complete graph on $[n]$ in the plane in a way that the edge between i and j is contained in $U_{\overline{\{i,j\}}}$ (see Fig. 3). If $n \geq 5$ then the stronger form of non-planarity of K_5 implies that there exist two edges $\{i,j\}$ and $\{k,\ell\}$ with no vertex in common and whose images intersect (see Proposition 8 and Lemma 9). Since $U_{\overline{\{i,j\}}} \cap U_{\overline{\{k,\ell\}}} = \bigcap \mathcal{F} = \emptyset$, this cannot happen and $\mathcal{F}$ has cardinality at most 4.

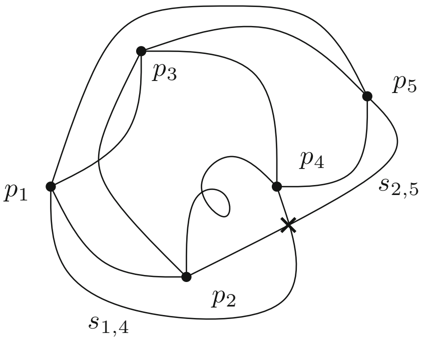


Fig. 3 Two edges (arcs) with no common vertices intersect (in this case $s_{1,4}$ and $s_{2,5}$). The point in the intersection then belongs to all sets in $\mathcal{F}$

$\lceil d/2 \rceil$-connected intersections in $\mathbb{R}^d$ The previous argument generalizes to higher dimension as follows. Assume that the intersections $\bigcap \mathcal{G}$ are $\lceil d/2 \rceil$-connected[20] for all subfamilies $\mathcal{G} \subsetneq \mathcal{F}$. Then we can build by induction a function f from the $\lceil d/2 \rceil$-skeleton of Δ_{n-1} to $\mathbb{R}^d$ in a way that for any simplex σ, the image $f(\sigma)$ is contained in $U_{\overline{\sigma}}$. The previous case shows how to build such a function from the 1-skeleton of Δ_{n-1}. Assume that a function f from the ℓ-skeleton of Δ_{n-1} is built. For every $(\ell + 1)$-simplex σ of Δ_{n-1}, for every facet τ of σ, we have $f(\tau) \subset U_{\overline{\tau}} \subseteq U_{\overline{\sigma}}$. Thus, the set

$$\bigcup_{\tau \text{ facet of } \sigma} f(\tau)$$

is the image of an ℓ-dimensional sphere contained in $U_{\overline{\sigma}}$, which has vanishing homotopy of dimension ℓ. We can extend f from this sphere to an $(\ell + 1)$-dimensional ball so that the image is still contained in $U_{\overline{\sigma}}$. This way we extend f to the $(\ell + 1)$-skeleton of Δ_{n-1}.

The Van Kampen-Flores theorem asserts that for any continuous function from $\Delta^{(k)}_{2k+2}$ to $\mathbb{R}^{2k}$ there exist two disjoint faces of $\Delta^{(k)}_{2k+2}$ whose images intersect (see Proposition 8 and Lemma 9). So, if $n \geq 2\lceil d/2 \rceil + 3$, then there exist two disjoint simplices σ and τ of $\Delta^{(\lceil d/2 \rceil)}_{2\lceil d/2 \rceil+2}$ such that $f(\sigma) \cap f(\tau)$ is nonempty. Since $f(\sigma) \cap f(\tau)$ is contained in $U_{\overline{\sigma}} \cap U_{\overline{\tau}} = \bigcap \mathcal{F} = \emptyset$, this is a contradiction and $\mathcal{F}$ has cardinality at most $2\lceil d/2 \rceil + 2$.

By a more careful inspection of odd dimensions, the bound $2\lceil d/2 \rceil + 2$ can be improved to $d + 2$. We skip this in the homotopic setting, but we will do so in the homological setting (which is stronger anyway); see Corollary 23 below.

Contractible intersections Of course, the previous argument works with other non-embeddability results. For instance, if the intersections $\bigcap \mathcal{G}$ are contractible for all subfamilies then the induction yields a map f from the d-skeleton of Δ_{n-1}

[20] Recall that a set is k-connected if it is connected and has vanishing homotopy in dimension 1 to k.

to $\mathbb{R}^d$ with the property that for any simplex σ, the image $f(\sigma)$ is contained in $U_{\overline{\sigma}}$. The topological Radon theorem [5] (see also [36, Theorem 5.1.2]) states that for any continuous function from Δ_{d+1} to $\mathbb{R}^d$ there exist two disjoint faces of Δ_{d+1} whose images intersect. So, if $n \geq d+2$ we again obtain a contradiction (the existence of two disjoint simplices σ and τ such that $f(\sigma) \cap f(\tau) \neq \emptyset$ whereas $U_{\overline{\sigma}} \cap U_{\overline{\tau}} = \bigcap \mathcal{F} = \emptyset$), and the cardinality of $\mathcal{F}$ must be at most $d+1$.

3.2 From Homotopy to Homology

The previous reasoning can be transposed to homology as follows. Assume that for $i = 0, 1, \ldots, k-1$ and all subfamilies $\mathcal{G} \subsetneq \mathcal{F}$ we have $\tilde{\beta}_i(\bigcap \mathcal{G}) = 0$. We construct a nontrivial[21] chain map f from the simplicial chains of $\Delta_{n-1}^{(k)}$ to the singular chains of $\mathbb{R}^d$ by increasing dimension:

- For every $\{i\} \subset [n]$ we let $p_i \in U_{\overline{\{i\}}}$. This is possible since every intersection of $n-1$ members of $\mathcal{F}$ is nonempty. We then put $f(\{i\}) = p_i$ and extend it by linearity into a chain map from $\Delta_{n-1}^{(0)}$ to $\mathbb{R}^d$. Notice that f is nontrivial and that for any 0-simplex $\sigma \subseteq [n]$, the support of $f(\sigma)$ is contained in $U_{\overline{\sigma}}$.
- Now, assume, as an induction hypothesis, that there exists a nontrivial chain map f from the simplicial chains of $\Delta_{n-1}^{(\ell)}$ to the singular chains of $\mathbb{R}^d$ with the property that for any ($\leq \ell$)-simplex $\sigma \subseteq [n], \ell < k$, the support of $f(\sigma)$ is contained in $U_{\overline{\sigma}}$. Let σ be a $(\ell+1)$-simplex in $\Delta_{n-1}^{(\ell+1)}$. For every ℓ-dimensional face τ of σ, the support of $f(\tau)$ is contained in $U_{\overline{\tau}} \subseteq U_{\overline{\sigma}}$. It follows that the support of $f(\partial\sigma)$ is contained in $U_{\overline{\sigma}}$, which has trivial homology in dimension $\ell+1$. As a consequence, $f(\partial\sigma)$ is a boundary in $U_{\overline{\sigma}}$. We can therefore extend f to every simplex of dimension $\ell+1$ and then, by linearity, to a chain map from the simplicial chains of $\Delta_{n-1}^{(\ell+1)}$ to the singular chains of $\mathbb{R}^d$. This chain map remains nontrivial and, by construction, for any ($\leq \ell+1$)-simplex $\sigma \subseteq [n]$, the support of $f(\sigma)$ is contained in $U_{\overline{\sigma}}$.

If σ and τ are disjoint simplices of $\Delta_{n-1}^{(k)}$ then the intersection of the supports of $f(\sigma)$ and $f(\tau)$ is contained in $U_{\overline{\sigma}} \cap U_{\overline{\tau}} = \bigcap \mathcal{F} = \emptyset$ and these supports are disjoint. It follows that f is not only a nontrivial chain map, but also a homological almost-embedding in $\mathbb{R}^d$. We can then use obstructions to the existence of homological almost-embeddings to bound the cardinality of $\mathcal{F}$. Specifically, since we assumed that $\mathcal{F}$ has empty intersection and any proper subfamily of $\mathcal{F}$ has nonempty intersection, Corollary 14 implies:

Corollary 23 *Let $\mathcal{F}$ be a family of subsets of $\mathbb{R}^d$ such that $\tilde{\beta}_i(\bigcap \mathcal{G}) = 0$ for every $\mathcal{G} \subsetneq \mathcal{F}$ and $i = 0, 1, \ldots, \lceil d/2 \rceil - 1$. Then the Helly number of $\mathcal{F}$ is at most $d+2$.*

[21] See Definition 10.

The homological Radon's lemma (Lemma 15) yields (noting $\partial\Delta_{d+1} = \Delta_{d+1}^{(d)}$):

Corollary 24 *Let $\mathcal{F}$ be a family of subsets of $\mathbb{R}^d$ such that $\tilde{\beta}_i(\bigcap \mathcal{G}) = 0$ for every $\mathcal{G} \subsetneq \mathcal{F}$ and $i = 0, 1, \ldots, d-1$. Then the Helly number of $\mathcal{F}$ is at most $d+1$.*

Remark 25 The following modification of Example 3 shows that the two previous statements are sharp in various ways. First assume that for some values k, n there exists some embedding$f$ of $\Delta_{n-1}^{(k)}$ into $\mathbb{R}^d$. Let K_i be the simplicial complex obtained by deleting the ith vertex of $\Delta_{n-1}^{(k)}$ (as well as all simplices using that vertex) and put $U_i := f(K_i)$. The family $\mathcal{F} = \{U_1, \ldots, U_n\}$ has Helly number exactly n, since it has empty intersection and all its proper subfamilies have nonempty intersection. Moreover, for every $\mathcal{G} \subseteq \mathcal{F}$, $\bigcap \mathcal{G}$ is the image through f of the k-skeleton of a simplex on $|\mathcal{F} \setminus \mathcal{G}|$ vertices, and therefore $\tilde{\beta}_i(\bigcap \mathcal{G}) = 0$ for every $\mathcal{G} \subseteq \mathcal{F}$ and $i = 0, \ldots, k-1$. Now, such an embedding exists for:

$k = d$ and $n = d+1$, as the d-dimensional simplex easily embeds into $\mathbb{R}^d$. Consequently, the bound of $d+1$ is best possible under the assumptions of Corollary 24.

$k = d-1$ and $n = d+2$, as we can first embed the $(d-1)$-skeleton of the d-simplex linearly, then add an extra vertex at the barycentre of the vertices of that simplex and embed the remaining faces linearly. This implies that if we relax the condition of Corollary 24 by only controlling the first $d-2$ Betti numbers then the bound of $d+1$ becomes false. It also implies that the bound of $d+2$ is best possible under (a strengthening of) the assumptions of Corollary 23.

(Recall that, as explained in Example 3, the $\lceil d/2 \rceil - 1$ in the assumptions of Corollary 23 cannot be reduced without allowing unbounded Helly numbers.)

Constrained chain map Let us formalize the technique illustrated by the previous example. We focus on the homological setting, as this is what we use to prove Theorem 1, but this can be easily transposed to homotopy.

Considering a slightly more general situation, we let $\mathcal{F} = \{U_1, U_2, \ldots, U_n\}$ denote a family of subsets of some topological space $\mathbf{R}$. As before for any (possibly empty) proper subset I of $[n] = \{1, 2, \ldots, n\}$ we write $U_{\overline{I}}$ for $\bigcap_{i \in [n] \setminus I} U_i$ and we put $U_{\overline{[n]}} = \mathbf{R}$.

Let K be a simplicial complex and let $\gamma : C_*(K) \to C_*(\mathbf{R})$ be a chain map from the simplicial chains of K to the singular chains of $\mathbf{R}$. We say that γ is *constrained by* $(\mathcal{F}, \Phi)$ if:

(i) Φ is a map from K to $2^{[n]}$ such that $\Phi(\sigma \cap \tau) = \Phi(\sigma) \cap \Phi(\tau)$ for all $\sigma, \tau \in K$ and $\Phi(\emptyset) = \emptyset$.

(ii) For any simplex $\sigma \in K$, the support of $\gamma(\sigma)$ is contained in $U_{\overline{\Phi(\sigma)}}$.

See Fig. 4. We also say that a chain map γ from K is *constrained by* $\mathcal{F}$ if there exists a map Φ such that γ is constrained by $(\mathcal{F}, \Phi)$. In the above constructions, we simply set Φ to be the identity. As we already saw, constrained chain maps relate Helly

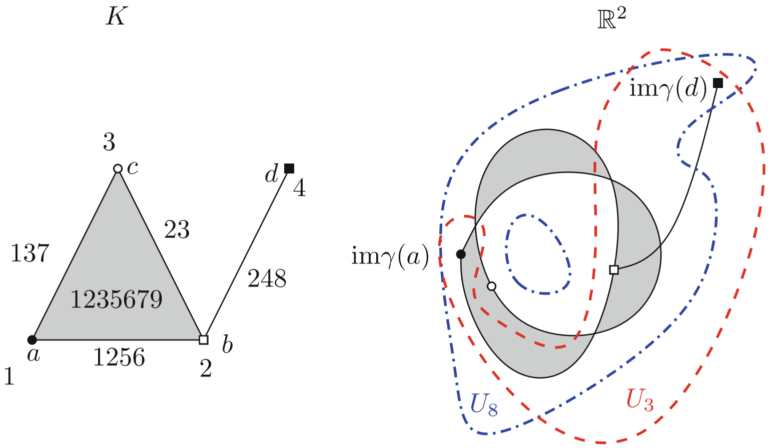


Fig. 4 An example of a constrained map $\gamma\colon K \to \mathbb{R}^2$. A label at a face σ of K denotes $\Phi(\sigma)$. Note, for example, that the support of $\gamma(\{a, b, c\})$ needn't be a triangle since we work with chain maps. Constrains by Φ mean that a set U_i must contain cover images of all faces without label i. It is demonstrated by U_3 and U_8 for example

numbers to homological almost-embeddings (see Definition 10) via the following observation:

Lemma 26 *Let $\gamma : C_*(K) \to C_*(\mathbf{R})$ be a nontrivial chain map constrained by $\mathcal{F}$. If $\bigcap \mathcal{F} = \emptyset$ then γ is a homological almost-embedding of K.*

Proof Let $\Phi : K \to 2^{[n]}$ be such that γ is constrained by $(\mathcal{F}, \Phi)$. Since γ is nontrivial, it remains to check that disjoint simplices are mapped to chains with disjoint support. Let σ and τ be two disjoint simplices of K. The supports of $\gamma(\sigma)$ and $\gamma(\tau)$ are contained, respectively, in $U_{\overline{\Phi(\sigma)}}$ and $U_{\overline{\Phi(\tau)}}$, and

$$U_{\overline{\Phi(\sigma)}} \cap U_{\overline{\Phi(\tau)}} = U_{\overline{\Phi(\sigma)\cap\Phi(\tau)}} = U_{\overline{\Phi(\sigma\cap\tau)}} = U_{\overline{\Phi(\emptyset)}} = U_{\overline{\emptyset}} = \bigcap \mathcal{F}.$$

Therefore, if $\bigcap \mathcal{F} = \emptyset$ then γ is a homological almost-embedding of K. □

3.3 Relaxing the Connectivity Assumption

In all the examples listed so far, the intersections $\bigcap \mathcal{G}$ must be connected. Matoušek [35] relaxed this condition into "having a bounded number of connected components", the assumptions then being on the topology of the components, by using Ramsey's theorem. The gist of our proof is to extend his idea to allow a bounded number of homology classes not only in the first dimension but in *any* dimension. Let us illustrate how Matoušek's idea works in two dimension:

Theorem 27 ([35, Theorem 2 with $d = 2$]) *For every positive integer b there is an integer $h(b)$ with the following property. If $\mathcal{F}$ is a finite family of subsets of $\mathbb{R}^2$ such that the intersection of any subfamily has at most b path-connected components, then the Helly number of $\mathcal{F}$ is at most $h(b)$.*

Let us fix b from above and assume that for any subfamily $\mathcal{G} \subsetneq \mathcal{F}$ the intersection $\bigcap \mathcal{G}$ consists of at most b path-connected components and that $\bigcap \mathcal{F} = \emptyset$. We start, as before, by picking for every $i \in [n]$, a point p_i in $U_{\overline{\{i\}}}$. This is possible as every intersection of $n-1$ members of $\mathcal{F}$ is nonempty. Now, if we consider some pair of indices $i, j \in [n]$, the points p_i and p_j are still in $U_{\overline{\{i,j\}}}$ but may lie in different connected components. It may thus not be possible to connect p_i to p_j *inside* $U_{\overline{\{i,j\}}}$. If we, however, consider $b+1$ indices $i_1, i_2, \ldots, i_{b+1}$ then all the points $p_{i_1}, p_{i_2}, \ldots, p_{i_{b+1}}$ are in $U_{\overline{\{i_1,i_2,\ldots,i_{b+1}\}}}$ which has at most b connected components, so at least one pair among of these points can be connected by a path inside $U_{\overline{\{i_1,i_2,\ldots,i_{b+1}\}}}$. Thus, while we may not get a drawing of the complete graph on n vertices we can still draw many edges.

To find many vertices among which every pair can be connected we will use the hypergraph version of the classical theorem of Ramsey:

Theorem 28 (Ramsey [45]) *For any x, y and z there is an integer $R_x(y, z)$ such that any x-uniform hypergraph on at least $R_x(y, z)$ vertices colored with at most y colors contains a subset of z vertices inducing a monochromatic sub-hypergraph.*

From the discussion above, for any $b+1$ indices $i_1 < i_2 < \ldots < i_{b+1}$ there exists a pair $\{k, \ell\} \in \binom{[b+1]}{2}$ such that p_{i_k} and p_{i_ℓ} can be connected inside $U_{\overline{\{i_1,i_2,\ldots,i_{b+1}\}}}$. Let us consider the $(b+1)$-uniform hypergraph on $[n]$ and color every set of indices $i_1 < i_2 < \ldots < i_{b+1}$ by one of the pairs in $\binom{[b+1]}{2}$ that can be connected inside $U_{\overline{\{i_1,i_2,\ldots,i_{b+1}\}}}$ (if more than one pair can be connected, we pick one arbitrarily). Let t be some integer to be fixed later. By Ramsey's theorem, if $n \geq R_{b+1}\left(\binom{b+1}{2}, t\right)$ then there exist a pair $\{k, \ell\} \in \binom{[b+1]}{2}$ and a subset $T \subseteq [n]$ of size t with the following property: for any $(b+1)$-element subset $S \subset T$, the points whose indices are the kth and ℓth indices of S can be connected inside $U_{\overline{S}}$.

Now, let us set $t = 5 + \binom{5}{2}(b-1) = 10b-5$. We claim that we can find five indices in T, denoted $i_1, i_2, \ldots, i_5$, and, for each pair $\{i_u, i_v\}$ among these five indices, some $(b+1)$-element subset $Q_{u,v} \subset T$ with the following properties:

(i) i_u and i_v are precisely in the kth and ℓth position in $Q_{u,v}$, and
(ii) for any $1 \leq u, v, u', v' \leq 5$, $\quad Q_{u,v} \cap Q_{u',v'} = \{i_u, i_v\} \cap \{i_{u'}, i_{v'}\}$.

We first conclude the argument, assuming that we can obtain such indices and sets. Observe that from the construction of T, the i_u's and the $Q_{u,v}$'s we have the following property: for any $u, v \in [5]$, we can connect p_{i_u} and p_{i_v} inside $U_{\overline{Q_{u,v}}}$. This gives a drawing of K_5 in the plane. Since K_5 is not planar, there exist two edges with no vertex in common, say $\{u, v\}$ and $\{u', v'\}$, that cross. This intersection point must lie in

$$U_{\overline{Q_{u,v}}} \cap U_{\overline{Q_{u',v'}}} = U_{\overline{Q_{u,v} \cap Q_{u',v'}}} = U_{\overline{\{i_u,i_v\} \cap \{i_{u'},i_{v'}\}}} = U_{\overline{\emptyset}} = \bigcap \mathcal{F} = \emptyset,$$

a contradiction. Hence the assumption that $n \geq R_{b+1}\left(\binom{b+1}{2}, t\right)$ is false and $\mathcal{F}$ has cardinality at most $R_{b+1}\left(\binom{b+1}{2}, 10b-5\right) - 1$, which is our $h(b)$.

The selection trick It remains to derive the existence of the i_u's and the $Q_{u,v}$'s. It is perhaps better to demonstrate the method by a simple example to develop some intuition before we formalize it.

Example Let us fix $b = 4$ and $\{k, \ell\} = \{2, 3\} \in \binom{[4+1]}{2}$. We first make a 'blueprint' for the construction inside the rational numbers. For any two indices $u, v \in [5]$ we form a totally ordered set $Q'_{u,v} \subseteq \mathbb{Q}$ of size $b + 1 = 5$ by adding three rational numbers (different from $1, \ldots, 5$) to the set $\{u, v\}$ in such a way that u appears at the second and v at the third position of $Q'_{u,v}$. For example, we can set $Q'_{1,4}$ to be $\{0.5; 1; 4; 4.7; 5.13\}$. Apart from this we require that we add a different set of rational numbers for each $\{u, v\}$. Thus $Q'_{u,v} \cap Q'_{u',v'} = \{u, v\} \cap \{u', v'\}$. Our blueprint now appears inside the set $T' := \bigcup_{1 \leq u < v \leq 5} Q'_{u,v}$; note that both this set T' and the set T in which we search for the sets $Q_{u,v}$ have 35 elements. To obtain the required indices i_u and sets $Q_{u,v}$ it remains to consider the unique strictly increasing bijection $\pi_0 \colon T' \to T$ and set $i_u := \pi_0(u)$ and $Q_{u,v} := \pi_0(Q'_{u,v})$.

The general case Let us now formalize the generalization of this trick that we will use to prove Theorem 1. Let Q be a subset of $[w]$. If $e_1 < e_2 < \ldots < e_w$ are the elements of a totally ordered set W then we call $\{e_i : i \in Q\}$ the *subset selected by Q in W*.

Lemma 29 *Let $1 \leq q \leq w$ be integers and let Q be a subset of $[w]$ of size q. Let Y and Z be two finite totally ordered sets and let $A_1, A_2, \ldots, A_r$ be q-element subsets of Y. If $|Z| \geq |Y| + r(w - q)$, then there exist an injection $\pi : Y \to Z$ and r subsets $W_1, W_2, \ldots, W_r \in \binom{Z}{w}$ such that for every $i \in [r]$, Q selects $\pi(A_i)$ in W_i. We can further require that $W_i \cap W_j = \pi(A_i \cap A_j)$ for any two $i, j \in [r]$, $i \neq j$.*

Proof Let π_0 denote the monotone bijection between Y and $[|Y|]$. For $i \in [r]$ we let D_i denote a set of $w - q$ rationals, disjoint from $[|Y|]$, such that Q selects $\pi_0(A_i)$ in $D_i \cup \pi_0(A_i)$. We further require that the D_i are pairwise disjoint, and put $Z' = [|Y|] \cup \left(\bigcup_{i \in [r]} D_i\right)$. Since $|Z| \geq |Y| + r(w-q) = |Z'|$ there exists a strictly increasing map $\nu : Z' \to Z$. We set $\pi := \nu \circ \pi_0$ and $W_i := \nu(D_i \cup \pi_0(A_i)) \in \binom{Z}{w}$. The desired condition is satisfied by this choice. See Fig. 5. □

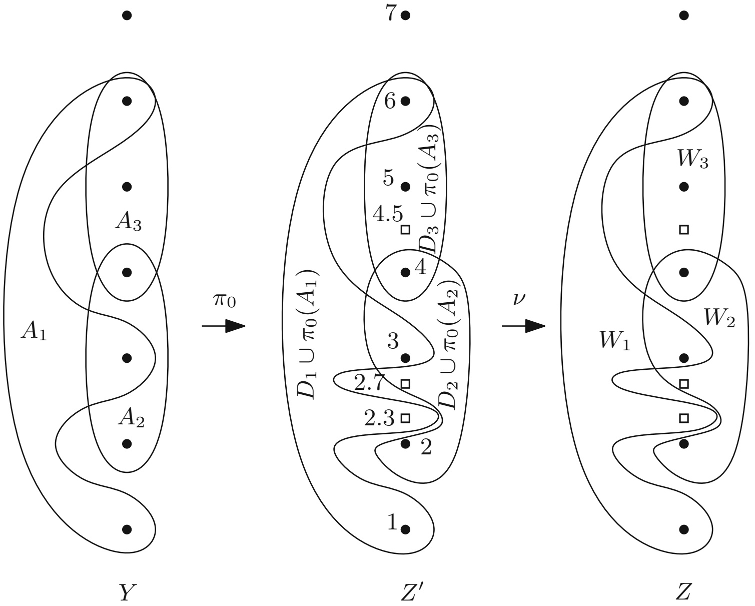


Fig. 5 Illustration for the proof of Lemma 29. We assume that $w = 4$ and $Q = \{1, 3, 4\}$

4 Constrained Chain Maps and Helly Number

We now generalize the technique presented in Sect. 3 to obtain Helly-type theorems from non-embeddability results. We will construct constrained chain maps for arbitrary complexes. As above, $\mathcal{F} = \{U_1, U_2, \ldots, U_n\}$ denotes a family of subsets of some topological space $\mathbf{R}$ and for $I \subseteq [n]$ we keep the notation $U_{\overline{I}}$ as used in the previous section (see the beginning of Sect. 3.1). Note that although so far we only used the *reduced* Betti numbers $\tilde{\beta}$, in this section it will be convenient to work with *standard* (non-reduced) Betti numbers β, starting with the following proposition.

Proposition 30 *For any finite simplicial complex K and non-negative integer b there exists a constant $h_K(b)$ such that the following holds. For any finite family $\mathcal{F}$ of at least $h_K(b)$ subsets of a topological space $\mathbf{R}$ such that $\bigcap \mathcal{G} \neq \emptyset$ and $\beta_i(\cap\mathcal{G}) \leq b$ for any $\mathcal{G} \subsetneq \mathcal{F}$ and any $0 \leq i < \dim K$, there exists a nontrivial chain map $\gamma : C_*(K) \to C_*(\mathbf{R})$ that is constrained by $\mathcal{F}$.*

The case $K = \Delta^{(k)}_{2k+2}$, with $k = \lceil d/2 \rceil$ and $\mathbf{R} = \mathbb{R}^d$, of Proposition 30 implies Theorem 1.

Proof of Theorem 1 Let b and d be fixed integers, let $k = \lceil d/2 \rceil$ and let $K = \Delta^{(k)}_{2k+2}$. Let $h_K(b+1)$ denote the constant from Proposition 30 (we plug in $b+1$ because we need to switch between reduced and non-reduced Betti numbers). Let $\mathcal{F}$ be a finite family of subsets of $\mathbb{R}^d$ such that $\tilde{\beta}_i(\bigcap \mathcal{G}) \leq b$ for any $\mathcal{G} \subsetneq \mathcal{F}$ and every

$0 \leq i \leq \dim K = \lceil d/2 \rceil - 1$, in particular $\beta_i(\bigcap \mathcal{G}) \leq b + 1$ for such $\mathcal{G}$. Let $\mathcal{F}^*$ denote an inclusion-minimal sub-family of $\mathcal{F}$ with empty intersection: $\bigcap \mathcal{F}^* = \emptyset$ and $\bigcap(\mathcal{F}^* \setminus \{U\}) \neq \emptyset$ for any $U \in \mathcal{F}^*$. If $\mathcal{F}^*$ has size at least $h_K(b + 1)$, it satisfies the assumptions of Proposition 30 and there exists a nontrivial chain map from K that is constrained by $\mathcal{F}^*$. Since $\mathcal{F}^*$ has empty intersection, this chain map is a homological almost-embedding by Lemma 26. However, no such homological almost-embedding exists by Corollary 13, so $\mathcal{F}^*$ must have size at most $h_K(b + 1) - 1$. As a consequence, the Helly number of $\mathcal{F}$ is bounded and the statement of Theorem 1 holds with $h(b, d) = h_K(b + 1) - 1$. □

The rest of this section is devoted to proving Proposition 30. We proceed by induction on the dimension of K, Sect. 4.1 settling the case of 0-dimensional complexes and Sect. 4.3 showing that if Proposition 30 holds for all simplicial complexes of dimension i then it also holds for all simplicial complexes of dimension $i + 1$. As the proof of the induction step is quite technical, as a warm-up, we provide the reader with a simplified argument for the induction step from $i = 0$ to $i = 1$ in Sect. 4.2. We let $V(K)$ and $v(K)$ denote, respectively, the set of vertices and the number of vertices of K.

4.1 Initialization (dim $K = 0$)

If K is a 0-dimensional simplicial complex then Proposition 30 holds with $h_K(b) = v(K)$. Indeed, consider a family $\mathcal{F}$ of at least $v(K)$ subsets of $\mathbf{R}$ such that all proper subfamilies have nonempty intersection. We enumerate the vertices of K as $\{v_1, v_2, \ldots, v_{v(K)}\}$ and define $\Phi(\{v_i\}) = \{i\}$; in plain English, Φ is a bijection between the set of vertices of K and $\{1, 2, \ldots, v(K)\}$. We first define γ on K by mapping every vertex $v \in K$ to a point $p(v) \in U_{\overline{\Phi(v)}}$, then extend it linearly into a chain map $\gamma : C_0(K) \to C_0(\mathbf{R})$. It is clear that γ is nontrivial and constrained by $(\mathcal{F}, \Phi)$, so Proposition 30 holds when dim $K = 0$.

4.2 Principle of the Induction Mechanism (dim $K = 1$)

As a warm-up, we now prove Proposition 30 for 1-dimensional simplicial complexes. While this merely amounts to reformulating Matoušek's proof for embeddings [35] in the language of chain maps, it still introduces several key ingredients of the induction while avoiding some of its complications. To avoid further technicalities, we use the non-reduced version of Betti numbers here.

Let K be a 1-dimensional simplicial complex with vertices $\{v_1, v_2, \ldots, v_{v(K)}\}$ and assume that $\mathcal{F}$ is a finite family of subsets of a topological space $\mathbf{R}$ such that for any $\mathcal{G} \subsetneq \mathcal{F}$, $\bigcap \mathcal{G} \neq \emptyset$ and $\beta_0(\cap \mathcal{G}) \leq b$. Let $s \in \mathbb{N}$ denote some parameter, to be fixed later. We assume that the cardinality of $\mathcal{F}$ is large enough (as a function of s) so that,

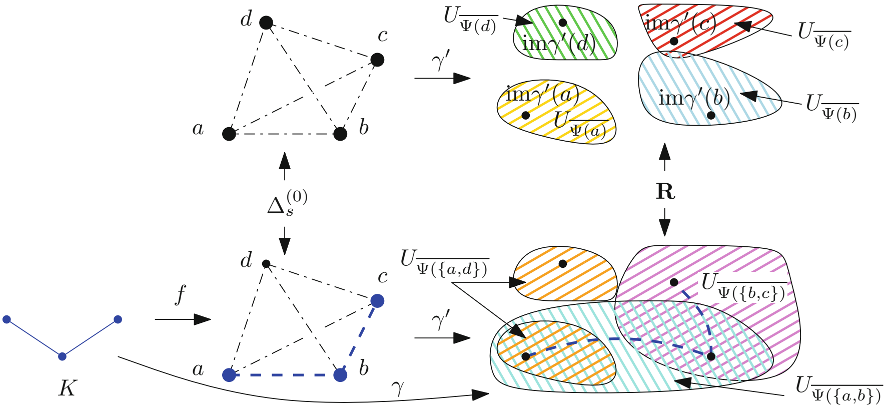


Fig. 6 Injecting $V(K)$ into $V(\Delta_s)$ by f in a way that the constrained chain map γ' from $V(\Delta_s)$ (*top*) can give rise to a constrained chain map from $V(K)$ (*bottom*); for the sake of illustration we use maps instead of chain maps. The situation considered here is simple, for instance $\gamma'(a+b)$ is a boundary in $U_{\overline{\Psi(\{a,b\})}}$ so $\gamma' \circ f_\sharp$ can be extended to the edge $\{f^{-1}(a), f^{-1}(b)\}$ of K. Note that if we wanted to use the edge ad, since $\gamma'(a+d)$ is not a boundary in $U_{\overline{\Psi(\{a,d\})}}$ we would need to add "dummy" elements to $\Psi(\{a,d\})$

as argued in Sect. 4.1, there exist a bijection $\Psi : \Delta_s^{(0)} \to [s+1]$ and a nontrivial chain map $\gamma' : C_*(\Delta_s^{(0)}) \to C_*(\mathbf{R})$ constrained by $(\mathcal{F}, \Psi)$. We extend Ψ to Δ_s by putting $\Psi(\sigma) = \cup_{v\in\sigma}\Psi(v)$ for any $\sigma \in \Delta_s$ and $\Psi(\emptyset) = \emptyset$. Remark that for any $\sigma, \tau \in \Delta_s$ we have $\Psi(\sigma \cap \tau) = \Psi(\sigma) \cap \Psi(\tau)$.

We now look for an injection f of $V(K)$ into $V(\Delta_s)$ such that the chain map $\gamma' \circ f_\sharp : C_*(K^{(0)}) \to C_*(\mathbf{R})$ can be extended into a chain map $\gamma : C_*(K) \to C_*(\mathbf{R})$ constrained by $\mathcal{F}$. Let $e = \{u, v\}$ be an edge in K. If we could arrange that $\gamma'(f(u) + f(v))$ is a boundary in $U_{\overline{\Psi(\{f(u),f(v)\})}}$ then we could simply define $\gamma(e)$ to be a chain in $U_{\overline{\Psi(\{f(u),f(v)\})}}$ bounded by $\gamma'(f(u)+f(v))$ (see Fig. 6). Unfortunately this is too much to ask for but we can still follow the Ramsey-based approach of Sect. 3.3: we add "dummy" vertices to $\{\Psi(\{f(u), f(v)\})\}$ to obtain a set W_e such that $\gamma'(f(u)+f(v))$ is a boundary in $U_{\overline{W_e}}$. If we use different dummy vertices for distinct edges then setting $\gamma(e)$ to be a chain in $U_{\overline{W_e}}$ bounded by $\gamma'(f(u)+f(v))$ still yields a chain map constrained by $\mathcal{F}$. We spell out the details in four steps.

Step 1. Any set S of 2^b+1 vertices of Δ_s contains two vertices $u_S, v_S \in S$ such that $\gamma'(u_S + v_S)$ is a boundary in $U_{\overline{\Psi(S)}}$.[22] Indeed, notice first that for any $u \in S$, the support of $\gamma'(u)$ is contained in $U_{\overline{\Psi(S)}}$. The assumption on $\mathcal{F}$ about bounded Betti

[22] We could require that γ' sends every vertex to a point in $U_{\overline{\Psi(S)}}$, i.e. is a chain map induced by a map, and simply argue that since $U_{\overline{\Psi(S)}}$ has at most b connected components, any $b+1$ vertices of Δ_s contains some pair that can be connected inside $U_{\overline{\Psi(S)}}$. This argument does not, however, work in higher dimension. Since Sect. 4.2 is meant as an illustration of the general case, we choose to follow the general argument.

numbers of intersections of subfamilies of $\mathcal{F}$ then ensures that there are at most 2^b distinct elements in $H_0(U_{\overline{\Psi(S)}})$, as $H_0(U_{\overline{\Psi(S)}}) \simeq \mathbb{Z}_2^m$ for some $m \leq b$. Thus, there are two vertices $u_S, v_S \in S$ such that $\gamma'(u_S)$ and $\gamma'(v_S)$ are in the same homology class in $H_0(U_{\overline{\Psi(S)}})$. Since we consider homology with coefficients over $\mathbb{Z}_2$, the sum of two chains that are in the same homology class is always a boundary. In particular, $\gamma'(u_S + v_S) = \gamma'(u_S) + \gamma'(v_S)$ is a boundary in $U_{\overline{\Psi(S)}}$.

Step 2. We use Ramsey's theorem (Theorem 28) to ensure a uniform "2-in-$(2^b + 1)$" selection. Let t be some parameter to be fixed in Step 3 and let H denote the $(2^b + 1)$-uniform hypergraph with vertex set $V(\Delta_s)$. For every hyperedge $S \in H$ there exists (by Step 1) a pair $Q_S \in \binom{[2^b+1]}{2}$ that selects a pair whose sum is mapped by γ' to a boundary in $U_{\overline{\Psi(S)}}$. We color H by assigning to every hyperedge S the "color" Q_S. Ramsey's theorem thus ensures that if $s \geq R_{2^b+1}\left(\binom{2^b+1}{2}, t\right)$ then there exist a set T of t vertices of Δ_s and a pair $Q^* \in \binom{[2^b+1]}{2}$ so that Q^* selects in *any* $S \in \binom{T}{2^b+1}$ a pair $\{u_S, v_S\}$ such that $\gamma'(u_S + v_S)$ is a boundary in $U_{\overline{\Psi(S)}}$.

Step 3. Now, let r be the number of edges of K and let $\sigma_1, \sigma_2, \ldots, \sigma_r$ denote the edges of K. We define

$$h_K(b) = R_{2^b+1}\left(\binom{2^b+1}{2}, r(2^b - 1) + v(K)\right) + 1$$

and assume that $s \geq h_K(b) - 1$. We set the parameter t introduced in Step 2 to $t = r(2^b - 1) + v(K)$. We can now apply Lemma 29 with $Y = V(K)$, $Z = T$, $q = 2$, $w = 2^b + 1$, and $A_i = \sigma_i$ for $i \in [r]$. As a consequence, there exist an injection $f : V(K) \to T$ and $W_1, W_2, \ldots, W_r$ in $\binom{T}{2^b+1}$ such that (i) for each i, Q^* selects $f(\sigma_i)$ in W_i, and (ii) $W_i \cap W_j = f(\sigma_i \cap \sigma_j)$ for $i, j \in [r], i \neq j$.

Step 4. We define Φ by

$$\begin{aligned} \Phi(\emptyset) &= \emptyset \\ \Phi(\{v_i\}) &= \Psi(f(v_i)) \text{ for } i = 1, 2, \ldots, v(K) \\ \Phi(\sigma_i) &= \Psi(W_i) \quad \text{ for } i = 1, 2, \ldots, r \end{aligned}$$

We define γ on the vertices of K by putting $\gamma(v) = \gamma'(f(v))$ for any $v \in V(K)$. Now remark that for any edge $\sigma_i = \{u, v\}$ of K, $\gamma'(f(u) + f(v))$ is a boundary in $U_{\overline{\Psi(W_i)}}$; this follows from the definition of T and the fact that Q^* selects $\{f(u), f(v)\}$ in W_i. We can therefore define $\gamma(\{u, v\})$ to be some (arbitrary) chain in $U_{\overline{\Psi(W_i)}}$ with boundary $\gamma'(f(u) + f(v))$. We then extend this map linearly into a chain map $\gamma : C_*(K) \to C_*(\mathbf{R})$.

To conclude the proof of Proposition 30 for 1-dimensional complexes it remains to check that the chain map γ and the function Φ defined in Step 4 have the desired properties.

Observation 31 *γ is a nontrivial chain map constrained by* $(\mathcal{F}, \Phi)$.

Proof First, it is clear from the definition that γ is a chain map. Moreover, the definition of γ' ensures that for every vertex $v \in K$ the support of $\gamma(v)$ is a finite set of points with odd cardinality. So γ is indeed a nontrivial chain map.

The map Φ is from K to $2^{[s+1]}$ and $\Phi(\emptyset)$ is by definition the empty set. The next property to check is that the identity $\Phi(\sigma \cap \tau) = \Phi(\sigma) \cap \Phi(\tau)$ holds for all $\sigma, \tau \in K$. When σ and τ are vertices this follows from the injectivity of Ψ and f. When σ and τ are edges this follows from the same identity for Ψ and the fact that Step 4 guaranteed that $W_i \cap W_j = f(\sigma_i \cap \sigma_j)$ for $i, j \in [r], i \neq j$. The remaining case is when $\sigma = \sigma_i$ is an edge and τ a vertex. Then, by construction, $\tau \in \sigma_i$ if and only if $f(\tau) \in W_i$, and

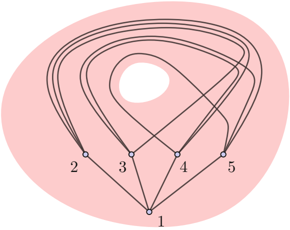
$$\begin{aligned}\Phi(\sigma_i) \cap \Phi(\tau) &= \Psi(W_i) \cap \Psi(f(\tau)) = \Psi(W_i \cap f(\tau)) \\ &= \left\{\begin{array}{ll} \Psi(\emptyset) & \text{if } f(\tau) \notin W_i \\ \Psi(f(\tau)) & \text{if } f(\tau) \in W_i \end{array}\right\} = \Phi(\sigma_i \cap \tau).\end{aligned}$$

It remains to check that for any simplex $\sigma \in K$, the support of $\gamma(\sigma)$ is contained in $U_{\overline{\Phi(\sigma)}}$. When $\sigma = \{v\}$ is a vertex then $\gamma(\sigma) = \gamma'(f(v))$. Since γ' is constrained by $(\mathcal{F}, \Psi)$, the support of $\gamma'(f(v))$ is contained in $U_{\overline{\Psi(f(v))}} = U_{\overline{\Phi(v)}}$, so the property holds. When $\sigma = \sigma_i$ is an edge, $\gamma(\sigma_i)$ is, by construction, a chain in $U_{\overline{\Psi(W_i)}} = U_{\overline{\Phi(\sigma_i)}}$ and the property also holds. □

4.3 The Induction

Let $k \geq 2$, let K be a simplicial complex of dimension k and assume that Proposition 30 holds for all simplicial complexes of dimension $k-1$ or less. Let $\mathcal{F}$ be a finite family of subsets of a topological space $\mathbf{R}$ such that for any $\mathcal{G} \subsetneq \mathcal{F}$ and any $0 \leq i \leq k-1$, $\bigcap \mathcal{G} \neq \emptyset$ and $\beta_i(\cap \mathcal{G}) \leq b$. Assuming that $\mathcal{F}$ contains sufficiently many sets, we want to construct a nontrivial chain map $\gamma : C_*(K) \to C_*(\mathbf{R})$ constrained by $\mathcal{F}$.

Preliminary example When going from $k = 0$ to $k = 1$, the first step (as described in Sect. 4.2) is to start with a constrained chain map $\gamma' : C_*(K^{(0)}) \to C_*(\mathbf{R})$ and observe that for some 1-simplices



$\{u, v\} \in K$ the chain $\gamma'(\partial\{u, v\})$ must already be a boundary. To see that this is not the case in general, consider the drawing of $\Delta_4^{(1)}$ in an annulus depicted in the figure above. Observe that for every triangle $\{i, j, k\} \in \Delta_4^{(2)}$ the image, in this drawing, of $\partial\{i, j, k\}$ is a cycle going around the hole of the annulus and is therefore not a boundary. So, if we start with a chain map γ' corresponding to that drawing, we will not be able to extend it by "filling" any triangle directly. This is not a peculiar example, and a similar construction can easily be done with arbitrarily many vertices. Observe, though, that the cycle going from 1 to 2, then 4, then 3 and then back to 1 *is* a boundary; in other words, if we replace, in the triangle $\partial\{1, 2, 3\}$, the edge from 2 to 3 by the concatenation of the edges from 2 to 4 and from 4 to 3, we build, using a chain map of $\Delta_4^{(1)}$ where no 2-face can be filled, a chain map of $\Delta_2^{(2)}$ where the 2-face can be filled. We systematize this observation using the barycentric subdivision of K.

Barycentric subdivision The idea behind the notion of *barycentric subdivision* is that the geometric realization of a simplicial complex K' can be subdivided by inserting a vertex at the barycentre of every face, resulting in a new, finer, simplicial complex, denoted $\operatorname{sd} K'$, that is still homeomorphic to K'. Formally, the vertices of $\operatorname{sd} K'$ consist of the faces of K', except for the empty face, and the faces of $\operatorname{sd} K'$ are the collections $\{\sigma_1, \dots, \sigma_\ell\}$ of faces of K' such that

$$\emptyset \neq \sigma_1 \subsetneq \sigma_2 \subsetneq \cdots \subsetneq \sigma_\ell.$$

In other words, the set of vertices of $\operatorname{sd} K'$ is $K' \setminus \{\emptyset\}$ and the faces of $\operatorname{sd} K'$ are the chains of $K' \setminus \{\emptyset\}$. For $\sigma \in K'$ we abuse the notation and let $\operatorname{sd} \sigma$ denote the subdivision of σ regarded as a subcomplex of $\operatorname{sd} K'$, that is,

$$\operatorname{sd} \sigma = \{\{\sigma_1, \dots, \sigma_\ell\} \subseteq K' : \emptyset \neq \sigma_1 \subsetneq \sigma_2 \subsetneq \cdots \subsetneq \sigma_\ell \subseteq \sigma\}.$$

We will mostly manipulate barycentric subdivisions through the $\operatorname{sd} \sigma$. For further reading on barycentric subdivisions we refer the reader, for example, to [36, Section 1.7].

Overview of the construction of $\boldsymbol{\gamma}$ Let $s \in \mathbb{N}$ be some parameter depending on K and to be determined later. To construct γ we will define three auxiliary chain maps

$$\begin{tikzcd} C_*\left(K^{(k-1)}\right) \arrow[r, "\alpha"] & C_*\left((\operatorname{sd} K)^{(k-1)}\right) \arrow[r, "\beta_\sharp"] & C_*\left(\Delta_s^{(k-1)}\right) \arrow[r, "\gamma'"] & C_*(\mathbf{R}) \end{tikzcd}$$

As before, γ' is a chain map from $C_*(\Delta_s^{(k-1)})$ constrained by $\mathcal{F}$ and is obtained by applying the induction hypothesis. Unlike in Sect. 4.2, we do not inject the vertices of K into those of Δ_s directly but proceed through $\operatorname{sd} K$, the barycentric subdivision of K. We "inject" $K^{(k-1)}$ into $\operatorname{sd} K^{(k-1)}$ by means of a chain map α (which will be the standard chain map corresponding to a subdivision). We then construct an injection

β of the vertices of $\operatorname{sd} K$ into the vertices of Δ_s which we extend linearly into a chain map $\beta_\sharp$. The key idea is the following:

> The boundary of any k-simplex σ of K is mapped, under α, to a sum of $k!$ boundaries of k-simplices of $\operatorname{sd} K$, all of which are mapped through $\beta_\sharp$ to chains with the same homology in some appropriate $U_{\overline{W_\sigma}}$.

Since $k!$ is even and we consider homology with coefficients in $\mathbb{Z}_2$, it follows that $\gamma' \circ \beta_\sharp \circ \alpha(\sigma)$ is a boundary in $U_{\overline{W_\sigma}}$. We therefore construct γ as an extension of $\gamma' \circ \beta_\sharp \circ \alpha$.

Definition of γ' Since $\Delta_s^{(k-1)}$ has dimension $k-1$, the induction hypothesis ensures that if the cardinality of $\mathcal{F}$ is large enough then there exists a nontrivial chain map $\gamma' : C_*(\Delta_s^{(k-1)}) \to C_*(\mathbf{R})$ constrained by $\mathcal{F}$. We denote by Ψ a map such that γ' is constrained by $(\mathcal{F}, \Psi)$. Remark that Ψ must be monotone over $\Delta_s^{(k-1)}$ as for any $\sigma \subseteq \tau \in \Delta_s^{(k-1)}$ we have $\Psi(\sigma) = \Psi(\sigma \cap \tau) = \Psi(\sigma) \cap \Psi(\tau) \subseteq \Psi(\tau)$. It follows that for any $\sigma \in \Delta_s^{(k-1)}$ we have

$$\Psi(\sigma) = \bigcup_{\tau \in \Delta_s^{(k-1)}, \tau \subseteq \sigma} \Psi(\tau)$$

We use this identity to extend Ψ to Δ_s, that is we define:

$$\forall A \subseteq V(\Delta_s), \quad \Psi(A) = \bigcup_{\tau \in \Delta_s^{(k-1)}, \tau \subseteq A} \Psi(\tau).$$

Remark that the extended map still commutes with the intersection:

Lemma 32 *For any $A, B \subseteq V(\Delta_s)$ we have $\Psi(A) \cap \Psi(B) = \Psi(A \cap B)$.*

Proof For any $A, B \subseteq V(\Delta_s)$ we have

$$\Psi(A) \cap \Psi(B) = \left(\bigcup_{\sigma \in \Delta_s^{(k-1)}, \sigma \subseteq A} \Psi(\sigma) \right) \cap \left(\bigcup_{\tau \in \Delta_s^{(k-1)}, \tau \subseteq B} \Psi(\tau) \right)$$

Distributing the union over the intersections we get

$$\Psi(A) \cap \Psi(B) = \bigcup_{\sigma, \tau \in \Delta_s^{(k-1)}, \sigma \subseteq A, \tau \subseteq B} \Psi(\sigma) \cap \Psi(\tau)$$

and as $\Psi(\sigma \cap \tau) = \Psi(\sigma) \cap \Psi(\tau)$ if σ, τ are simplices of $\Delta_s^{(k-1)}$, this rewrites as

$$\Psi(A) \cap \Psi(B) = \bigcup_{\sigma, \tau \in \Delta_s^{(k-1)}, \sigma \subseteq A, \tau \subseteq B} \Psi(\sigma \cap \tau).$$

Finally, observing that

$$\{\sigma \cap \tau : \sigma, \tau \in \Delta_s^{(k-1)}, \sigma \subseteq A, \tau \subseteq B\} = \{\vartheta : \vartheta \in \Delta_s^{(k-1)}, \vartheta \subseteq A \cap B\}$$

we get

$$\Psi(A) \cap \Psi(B) = \bigcup_{\vartheta \in \Delta_s^{(k-1)}, \vartheta \subseteq A \cap B} \Psi(\vartheta) = \Psi(A \cap B)$$

which proves the desired identity. □

Definition of $\boldsymbol{\alpha}$ Now we define a chain map $\alpha : C_* \left(K^{(k-1)}\right) \to C_* \left(\operatorname{sd} K^{(k-1)}\right)$ by first putting

$$\alpha : \sigma \in K^{(k-1)} \mapsto \sum_{\substack{\tau \in \operatorname{sd} \sigma \\ \dim \tau = \dim \sigma}} \tau,$$

and then extending that map linearly to $C_* \left(K^{(k-1)}\right)$. See Fig. 7. Remark that α behaves nicely with respect to the differential:

$$\alpha(\partial \sigma) = \sum_{\substack{\tau \in \operatorname{sd} \sigma \\ \dim \tau = \dim \sigma}} \partial \tau.$$

Note that the formula above makes sense and is valid even if σ is a k-simplex although we define α only up to dimension $k - 1$.

Definition of $\boldsymbol{\beta}$ We now construct the injection $\beta : V(\operatorname{sd} K) \to V(\Delta_s)$ and, for constraining purposes, an auxiliary function κ associating with every k-dimensional simplex of K some simplex of Δ_s. We want these functions to satisfy:

(P1) For any simplex $\sigma \in K$, $\kappa(\sigma) \cap \operatorname{Im} \beta = \beta(V(\operatorname{sd} \sigma))$.
(P2) For any k-simplices $\sigma, \tau \in K$, $\kappa(\sigma) \cap \kappa(\tau) = \beta(V(\operatorname{sd} \sigma)) \cap \beta(V(\operatorname{sd} \tau))$.

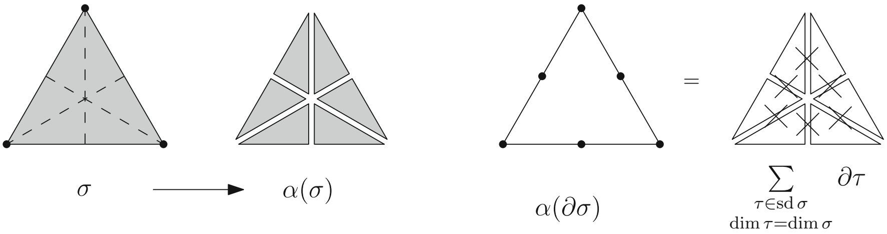

σ
$\alpha(\sigma)$
$\alpha(\partial\sigma)$
=
$\sum_{\substack{\tau \in \operatorname{sd} \sigma \\ \dim \tau = \dim \sigma}} \partial \tau$

Fig. 7 The map α applied to a simplex σ (*left*) and to $\partial \sigma$ (*right*). Significant parts of the boundaries $\partial \tau$ cancel out

(P3) For any k-simplex $\sigma \in K$, when τ ranges over all k-simplices of $\mathrm{sd}\,\sigma$, all chains $\gamma' \circ \beta_\sharp(\partial\tau)$ have support in $U_{\overline{\Psi(\kappa(\sigma))}}$ and are in the same homology class in $H_{k-1}(U_{\overline{\Psi(\kappa(\sigma))}})$.

The intuition behind these properties is that $\kappa(\sigma)$ should augment $\beta(V(\mathrm{sd}\,\sigma))$ by "dummy" vertices (P1) in a way that distinct simplices use disjoint sets of "dummy" vertices (P2). Property (P3), will allow building γ over k-simplices as explained in the preceding overview.

We start the construction of β and κ with a combinatorial lemma. Let $\ell = 2^{k+1} - 1$ stand for the number of vertices of the barycentric subdivision of a k-dimensional simplex, and set $m = R_{k+1}(2^b, \ell)$.

Claim 1 *For any integer t, if $s \geq R_m\left(\binom{m}{\ell}, t\right)$ then there exist a set T of t vertices of Δ_s and a set $Q^* \in \binom{[m]}{\ell}$ such that Q^* selects in any $M \in \binom{T}{m}$ a subset L_M with the following property: when σ ranges over all k-simplices of Δ_s with $\sigma \subseteq L_M$, all chains $\gamma'(\partial\sigma)$ are in the same homology class in $H_{k-1}\left(U_{\overline{\Psi(M)}}\right)$.*

Proof Let M be a subset of m vertices of Δ_s. Since γ' is constrained by $(\mathcal{F}, \Psi)$, for every k-simplex $\sigma \subseteq M$ the support of $\gamma'(\partial\sigma)$ is contained in $U_{\overline{\Psi(\partial\sigma)}} \subseteq U_{\overline{\Psi(\sigma)}} \subseteq U_{\overline{\Psi(M)}}$. We can therefore color the $(k+1)$-uniform hypergraph on M by assigning to every hyperedge σ the homology class of $\gamma'(\partial\sigma)$ in $U_{\overline{\Psi(M)}}$. Since $\beta_{k-1}\left(U_{\overline{\Psi(M)}}\right) \leq b$, there are at most 2^b colors in this coloring. As $m = R_{k+1}(2^b, \ell)$, Ramsey's Theorem implies that there exists a subset $L \subset M$ of ℓ vertices inducing a monochromatic hypergraph. We let Q_M denote an element of $\binom{[m]}{\ell}$ that selects such a subset L.

It remains to find a subset T of vertices of Δ_s so that all m-element subsets $M \subseteq T$ give rise to the same Q_M. This is done by another application of Ramsey's theorem to the m-uniform hypergraph on the vertices of Δ_s where each hyperedge M is colored by the ℓ-element subset Q_M. The subset T can have size t as soon as $s \geq R_m\left(\binom{m}{\ell}, t\right)$, which proves the statement. □

Now, back to the construction of β and κ. We first want a subset of $V(\Delta_s)$ with a "uniform ℓ-in-m selection" property of Claim 1 large enough so that we can inject $V(\mathrm{sd}\,K)$ using Lemma 29. We set:

$$t = v(\mathrm{sd}\,K) + r(m - \ell) \quad \text{and} \quad s^* = R_m\left(\binom{m}{\ell}, t\right),$$

and assume that $s \geq s^*$; since s^* only depends on b and K, this merely requires that $\mathcal{F}$ is large enough, again as a function of b and K, so that γ' still exists. We let T and Q^* denote the subset of $V(\Delta_s)$ and the element of $\binom{[m]}{\ell}$ whose existence follows from applying Claim 1. Let $\sigma_1, \sigma_2, \ldots, \sigma_r$ denote the k-dimensional simplices of K. We apply Lemma 29 with

$$Y = V(\mathrm{sd}\,K), \quad Z = T, \quad A_i = V(\mathrm{sd}\,\sigma_i), \quad q = \ell, \quad \text{and } w = m,$$

and obtain an injection $\pi : Y \to Z$ and $W_1, W_2, \dots, W_r \in \binom{Z}{m}$ such that (i) for every $i \leq r$, Q^* selects $\pi(A_i)$ in W_i, and (ii) for any $i \neq j \leq r$, $W_i \cap W_j = \pi(A_i \cap A_j)$. This injection π is our map β and we put $\kappa(\sigma_i) = W_i$. It is clear that Property (P1) holds, and since

$$\begin{aligned}\kappa(\sigma_i) \cap \kappa(\sigma_j) &= W_i \cap W_j = \pi(A_i \cap A_j) = \beta(V(\operatorname{sd}\sigma_i) \cap V(\operatorname{sd}\sigma_j)) \\ &= \beta(V(\operatorname{sd}\sigma_i)) \cap \beta(V(\operatorname{sd}\sigma_j)),\end{aligned}$$

Property (P2) also holds. The set Q^* selects $\pi(A_i)$ in W_i (Lemma 29) so Claim 1 ensures that when τ ranges over all k-simplices of Δ_s with $\tau \subseteq \pi(A_i)$, all chains $\gamma'(\partial\tau)$ have support in $U_{\overline{\Psi(W_i)}}$ and are in the same homology class in $H_{k-1}\left(U_{\overline{\Psi(W_i)}}\right)$. Substituting $\pi(A_i) = \beta(V(\operatorname{sd}\sigma_i))$ and $W_i = \kappa(\sigma_i)$, we see that (P3) holds.

Construction of γ Recall that we have the chain maps[23]:

$$\begin{tikzcd} C_*\left(K^{(k-1)}\right) \arrow[r, "\alpha"] & C_*\left((\operatorname{sd}K)^{(k-1)}\right) \arrow[r, "\beta_\sharp"] & C_*\left(\Delta_s^{(k-1)}\right) \arrow[r, "\gamma'"] & C_*(\mathbf{R}). \end{tikzcd}$$

We define $\gamma = \gamma' \circ \beta_\sharp \circ \alpha$ as a chain map from $C_*\left(K^{(k-1)}\right)$ to $C_*(\mathbf{R})$. Let σ be a k-dimensional simplex of K. From the definition of α we have

$$\gamma(\partial\sigma) = \sum_{\substack{\tau \in \operatorname{sd}\sigma \\ \dim\tau = \dim\sigma}} \gamma' \circ \beta_\sharp(\partial\tau).$$

By property (P3), all summands in the above chain have support in $U_{\overline{\Psi(\kappa(\sigma))}}$ and belong to the same homology class in $H_{k-1}\left(U_{\overline{\Psi(\kappa(\sigma))}}\right)$. There is an even number of summands, namely $k!$ and we are using homology over $\mathbb{Z}_2$, so $\gamma' \circ \beta_\sharp \circ \alpha(\partial\sigma)$ has support in $U_{\overline{\Psi(\kappa(\sigma))}}$ and is a boundary in $U_{\overline{\Psi(\kappa(\sigma))}}$. We can therefore extend γ into a chain map from $C_*(K)$ to $C_*(\mathbf{R})$ in a way that for any k-simplex σ of K, the support of $\gamma(\sigma)$ is contained in $U_{\overline{\Psi(\kappa(\sigma))}}$.

Properties of γ First we verify that γ is nontrivial. If v is a vertex of K then $\operatorname{sd}v$ consists of a single simplex, also a vertex. The chain $\alpha(v)$ is thus a single vertex of $\operatorname{sd}K$, and $\beta_\sharp \circ \alpha(v)$ is still a single vertex $\beta(\operatorname{sd}v)$. Since γ' is nontrivial, the support of $\gamma(v)$ is an odd number of points and therefore γ is also nontrivial. It remains to argue that γ is constrained by $(\mathcal{F}, \Phi)$ where:

$$\Phi : \begin{cases} K \to 2^{\mathcal{F}} \\ \sigma \mapsto \begin{cases} \Psi(\beta(V(\operatorname{sd}\sigma))) & \text{if } \dim\sigma \leq k-1 \\ \Psi(\kappa(\sigma)) & \text{if } \dim\sigma = k \end{cases} \end{cases}$$

[23] $\beta_\sharp$ is the chain map induced by β restricted to chains of dimension at most $(k-1)$.

It is clear that $\Phi(\emptyset) = \Psi(\emptyset) = \emptyset$ by definition of Ψ. Also, the construction of γ immediately ensures that for any $\sigma \in K$ the support of $\gamma(\sigma)$ is contained in $U_{\overline{\Phi(\sigma)}}$. To conclude the proof that γ is constrained by $(\mathcal{F}, \Phi)$ and therefore the induction it only remains to check that Φ commutes with the intersection:

Claim 2 *For any $\sigma, \tau \in K$, $\Phi(\sigma \cap \tau) = \Phi(\sigma) \cap \Phi(\tau)$.*

Proof The claim is obvious for $\sigma = \tau$, so from now on assume that this is not the case. First assume that σ and τ have dimension at most $k - 1$. Then,

$$\Phi(\sigma) \cap \Phi(\tau) = \Psi(\beta(V(\operatorname{sd}\sigma))) \cap \Psi(\beta(V(\operatorname{sd}\tau))) = \Psi(\beta(V(\operatorname{sd}\sigma)) \cap \beta(V(\operatorname{sd}\tau))),$$

the last equality following from Lemma 32. Since the map β on subsets of $V(\Delta_s)$ is induced by a map β on vertices of Δ_s we have $\beta(V(\operatorname{sd}\sigma)) \cap \beta(V(\operatorname{sd}\tau)) = \beta(V(\operatorname{sd}\sigma) \cap V(\operatorname{sd}\tau))$. Moreover, by the definition of the barycentric subdivision we have $V(\operatorname{sd}\sigma) \cap V(\operatorname{sd}\tau) = V(\operatorname{sd}(\sigma \cap \tau))$. Thus,

$$\Psi(\beta(V(\operatorname{sd}\sigma)) \cap \beta(V(\operatorname{sd}\tau))) = \Psi(\beta(V(\operatorname{sd}(\sigma \cap \tau)))) = \Phi(\sigma \cap \tau),$$

and the statement holds for simplices of dimension at most $k - 1$.

Now assume that σ and τ are both k-dimensional so that

$$\Phi(\sigma) \cap \Phi(\tau) = \Psi(\kappa(\sigma)) \cap \Psi(\kappa(\tau)) = \Psi(\kappa(\sigma) \cap \kappa(\tau)) = \Psi(\beta(V(\operatorname{sd}\sigma)) \cap \beta(V(\operatorname{sd}\tau))),$$

the last identity following from Property (P2) of the map κ. Again, from the definition of β and the barycentric subdivision we have

$$\beta(V(\operatorname{sd}\sigma)) \cap \beta(V(\operatorname{sd}\tau)) = \beta(V(\operatorname{sd}(\sigma \cap \tau))).$$

We thus obtain

$$\Phi(\sigma) \cap \Phi(\tau) = \Psi \circ \beta \circ V(\operatorname{sd}(\sigma \cap \tau)) = \Phi(\sigma \cap \tau),$$

the last identity following from the definition of Φ on simplices of dimension at most $k - 1$. The statement also holds for simplices of dimension k.

Finally assume that σ and τ are of dimension k and at most $k - 1$ respectively. Then, applying Lemma 32 we have:

$$\Phi(\sigma) \cap \Phi(\tau) = \Psi(\kappa(\sigma)) \cap \Psi(\beta(V(\operatorname{sd}\tau))) = \Psi(\kappa(\sigma) \cap \beta(V(\operatorname{sd}\tau))).$$

Note that $\beta(V(\operatorname{sd}\tau)) \subseteq \operatorname{Im}\beta$ and that, by property (P1), $\kappa(\sigma) \cap \operatorname{Im}\beta = \beta(V(\operatorname{sd}\sigma))$. We thus have

$$\kappa(\sigma) \cap \beta(V(\operatorname{sd}\tau)) = \beta(V(\operatorname{sd}\sigma)) \cap \beta(V(\operatorname{sd}\tau)) = \beta(V(\operatorname{sd}(\sigma \cap \tau))),$$

the last equality following, again, from the definition of barycentric subdivision. As $\sigma \cap \tau$ has dimension at most $k-1$ we have

$$\Phi(\sigma) \cap \Phi(\tau) = \Psi(\beta(V(\text{sd}(\sigma \cap \tau)))) = \Phi(\sigma \cap \tau)$$

and the statement holds for the last case. □

Acknowledgements We would like to express our immense gratitude to Jiří Matoušek, not only for raising the problem addressed in the present paper and valuable discussions about it, but, much more generally, for the privilege of having known him, as our teacher, mentor, collaborator, and friend. Through his tremendous depth and insight, and the generosity with which he shared them, he greatly influenced all of us.

We further thank Jürgen Eckhoff for helpful comments on a preliminary version of the paper, and Andreas Holmsen and Gil Kalai for providing us with useful references.

References

1. N. Alon, G. Kalai, Bounding the piercing number. Discrete Comput. Geom. **13**, 245–256 (1995)
2. N. Alon, I. Bárány, Z. Füredi, D.J. Kleitman. Point selections and weak *epsilon*-nets for convex hulls. Combin. Probab. Comput. **1**, 189–200 (1992)
3. N. Amenta, Helly-type theorems and generalized linear programming. Discrete Comput. Geom. **12**, 241–261 (1994)
4. N. Amenta, A short proof of an interesting Helly-type theorem. Discrete Comput. Geom. **15**, 423–427 (1996)
5. E.G. Bajmóczy, I. Bárány, On a common generalization of Borsuk's and Radon's theorem. Acta Math. Acad. Sci. Hungar. **34**(3–4), 347–350 (1979)
6. M. Bestvina, M. Kapovich, B. Kleiner, Van Kampen's embedding obstruction for discrete groups. Invent. Math. **150**(2), 219–235 (2002)
7. A. Björner, Nerves, fibers and homotopy groups. J. Combin. Theory Ser. A **102**(1), 88–93 (2003)
8. K. Borsuk, On the imbedding of systems of compacta in simplicial complexes. Fundamenta Mathematicae **35**, 217–234 (1948)
9. G.E. Bredon, *Sheaf Theory*. Volume 170 of Graduate Texts in Mathematics, 2nd edn. (Springer, New York, 1997)
10. O. Cheong, X. Goaoc, A. Holmsen, S. Petitjean, Hadwiger and Helly-type theorems for disjoint unit spheres. Discrete Comput. Geom. **1–3**, 194–212 (2008)
11. E. Colin de Verdiere, G. Ginot, X. Goaoc, Helly numbers of acyclic families. Adv. Mathe. **253**, 163–193 (2014)
12. H. Debrunner, Helly type theorems derived from basic singular homology. Am. Math. Mon. **77**, 375–380 (1970)
13. J. de Loera, S. Petrović, D. Stasi, Random sampling in computational algebra: Helly numbers and violator spaces. http://arxiv.org/abs/1503.08804
14. M. Deza, P. Frankl, A Helly type theorem for hypersurfaces. J. Combin. Theory Ser. A **45**, 27–30 (1987)
15. J. Dugundji, A duality property of nerves. Fundamenta Mathematicae **59**, 213–219 (1966)
16. J. Eckhoff, Helly, Radon and Carathéodory type theorems, in *Handbook of Convex Geometry*, ed. by P.M. Gruber, J.M. Wills (North Holland, Amsterdam/New York, 1993), pp. 389–448

17. J. Eckhoff, K.-P. Nischke, Morris's pigeonhole principle and the Helly theorem for unions of convex sets. Bull. Lond. Math. Soc. **41**, 577–588 (2009)
18. B. Farb, Group actions and Helly's theorem. Adv. Math. **222**, 1574–1588 (2009)
19. A.I. Flores, Über die Existenz n-dimensionaler Komplexe, die nicht in den $\mathbb{R}^{2n}$ topologisch einbettbar sind. Ergeb. Math. Kolloqu. **5**, 17–24 (1933)
20. X. Goaoc, I. Mabillard, P. Paták, Z. Patáková, M. Tancer, U. Wagner, On generalized Heawood Inequalities for manifolds: a Van Kampen-Flores-type nonembeddability result, in *31st International Symposium on Computational Geometry (SoCG 2015)*, Dagstuhl, ed. by L. Arge, J. Pach. Volume 34 of Leibniz International Proceedings in Informatics (LIPIcs). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, pp. 476–490
21. X. Goaoc, P. Paták, Z. Patáková, M. Tancer, U. Wagner, Bounding Helly numbers via Betti numbers, in *31st International Symposium on Computational Geometry (SoCG 2015)*, Dagstuhl, ed. by L. Arge, J. Pach. Volume 34 of Leibniz International Proceedings in Informatics (LIPIcs). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, pp. 507–521
22. R. González-Diaz, P. Real, Simplification techniques for maps in simplicial topology. J. Symb. Comput. **40**, 1208–1224 (2005)
23. J.E. Goodman, A. Holmsen, R. Pollack, K. Ranestad, F. Sottile, Cremona convexity, frame convexity, and a theorem of Santaló. Adv. Geom. **6**, 301–322 (2006)
24. B. Grünbaum, On common transversals. Archiv der Mathematik **9**, 465–469 (1958)
25. A. Haefliger, Plongements différentiables dans le domaine stable. Comment. Math. Helv. **37**, 155–176 (1962/1963)
26. A. Hatcher, *Algebraic Topology* (Cambridge University Press, Cambridge, 2002)
27. E. Helly, Über Mengen konvexer Körper mit gemeinschaftlichen Punkten. Jahresbericht Deutsch. Math. Verein. **32**, 175–176 (1923)
28. E. Helly, Über Systeme von abgeschlossenen Mengen mit gemeinschaftlichen Punkten. Monaths. Math. und Physik **37**, 281–302 (1930)
29. G. Kalai, Combinatorial expectations from commutative algebra, in *Combinatorial Commutative Algebra*, ed. by I. Peeva, V. Welker, vol. 1, no. 3 (Oberwolfach Reports, 2004), pp. 1729–1734
30. G. Kalai, R. Meshulam, Leray numbers of projections and a topological Helly-type theorem. J. Topol. **1**, 551–556 (2008)
31. M. Katchalski, A conjecture of Grünbaum on common transversals. Math. Scand. **59**(2), 192–198 (1986)
32. D. Kozlov, *Combinatorial Algebraic Topology*. Volume 21 of Algorithms and Computation in Mathematics (Springer, Berlin, 2008)
33. D.G. Larman, Helly type properties of unions of convex sets. Mathematika **15**, 53–59 (1968)
34. H. Maehara, Helly-type theorems for spheres. Discrete Comput. Geom. **4**(1), 279–285 (1989)
35. J. Matoušek, A Helly-type theorem for unions of convex sets. Discrete Comput. Geom. **18**, 1–12 (1997)
36. J. Matoušek, *Using the Borsuk–Ulam Theorem* (Springer, Berlin, 2003)
37. S.A. Melikhov, The van Kampen obstruction and its relatives. Proc. Steklov Inst. Math. **266**(1), 142–176 (2009)
38. J. Milnor, Construction of universal bundles, II. Ann. Math. **63**(3), 430–436 (1956)
39. J. Milnor, On the Betti numbers of real varieties. Proc. Am. Math. Soc. **15**, 275–280 (1963)
40. L. Montejano, A new topological Helly theorem and some transversal results. Discrete Comput. Geom. **52**(2), 390–398 (2014)
41. T.S. Motzkin, A proof of Hilbert's Nullstellensatz. Mathematische Zeitschrift **63**, 341–344 (1955)
42. J.R. Munkres, Elements of Algebraic Topology (Addison-Wesley, Menlo Park, 1984)
43. C.M. Petty, Equilateral sets in Minkowski spaces. Proc. Am. Math. Soc. **29**, 369–374 (1971)
44. R. Rado, A theorem on general measure. J. Lond. Math. Soc. **s1–21**(4), 291–300 (1946)
45. F.P. Ramsey, On a problem in formal logic. Proc. Lond. Math. Soc. **30**, 264–286 (1929)
46. C.P. Rourke, B.J. Sanderson, Introduction to Piecewise-Linear Topology (Springer, New York, 1972). Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 69

47. M. Sharir, E. Welzl, A combinatorial bound for linear programming and related problems, in *Proceedings of the 9th Symposium on Theoretical Aspects of Computer Science* (1992), pp. 569–579
48. P. Shvartsman, The Whitney extension problem and Lipschitz selections of set-valued mappings in jet-spaces. Trans. Am. Math. Soc. **360**, 5529–5550 (2008)
49. A.B. Skopenkov, Embedding and knotting of manifolds in Euclidean spaces, in *Surveys in Contemporary Mathematics*. Volume 347 of London Mathematical Society Lecture Note Series (Cambridge University Press, Cambridge, 2008), pp. 248–342
50. R.I. Soare, Computability theory and differential geometry. Bull. Symb. Log. **10**(4), 457–486 (2004)
51. J. Sosnovec, Draft of Bachelor's thesis (2015)
52. K.J. Swanepoel, Helly-type theorems for hollow axis-aligned boxes. Proc. Am. Math. Soc. **127**, 2155–2162 (1999)
53. K.J. Swanepoel, Helly-type theorems for homothets of planar convex curves. Proc. Am. Math. Soc. **131**, 921–932 (2003)
54. M. Tancer, Intersection patterns of convex sets via simplicial complexes: a survey, in *Thirty Essays on Geometric Graph Theory*, ed. by J. Pach (Springer, New York, 2013), pp. 521–540
55. R. Thom, Sur l'homologie des variétés algébriques réelles, in *Differential and Combinatorial Topology,* ed. by S.S. Cairns (Princeton University Press, Princeton, 1965), pp. 255–265
56. H. Tverberg, Proof of Grünbaum's conjecture on common transversals for translates. Discrete Comput. Geom. **4**, 191–203 (1989)
57. E.R. van Kampen, Komplexe in euklidischen Räumen. Abh. Math. Sem. Univ. Hamburg **9**, 72–78 (1932)
58. U. Wagner, Minors in random and expanding hypergraphs, in *Proceedings of the 27th Annual Symposium on Computational Geometry (SoCG)* (2011), pp. 351–360
59. C. Weber, Plongements de polyèdres dans le domaine métastable. Comment. Math. Helv. **42**, 1–27 (1967)
60. G. Wegner, d-collapsing and nerves of families of convex sets. Archiv der Mathematik **26**, 317–321 (1975)
61. R. Wenger, Helly-type theorems and geometric transversals, in *Handbook of Discrete & Computational Geometry*, 2nd edn., ed. by J.E. Goodman, J. O'Rourke (CRC Press LLC, Boca Raton, 2004), chapter 4, pp. 73–96
62. G.M. Ziegler, *Lectures on Polytopes*. Volume 152 of Graduate Texts in Mathematics (Springer, New York, 1995)

Ruled Surface Theory and Incidence Geometry

Larry Guth

Abstract We survey the applications of ruled surface theory in incidence geometry. We discuss some of the proofs and raise some open questions.

1 Introduction

In the last 5 years, there have been some interesting applications of ruled surface theory in incidence geometry, which started in the work that Nets Katz and I did on the Erdős distinct distance problem [5]. In this essay, we survey the role of ruled surface theory in incidence geometry.

Ruled surface theory is a subfield of algebraic geometry. A ruled surface is an algebraic variety that contains a line through every point. Ruled surface theory tries to classify ruled surfaces and to describe their structure. The incidence geometry questions that we study here are about finite sets of lines. A ruled surface can be roughly thought of as an algebraic family of lines. Some of the questions in the two fields are actually parallel, but they take place in two different settings – the discrete setting and the algebraic setting. We will discuss a connection between these two settings.

The applications of ruled surface theory are the most technical part of [5]. I wrote a book about polynomial methods in combinatorics, [4], including a chapter about applications of ruled surface theory. My goal in that chapter was to give a self-contained proof of the results from [5] and to make the technical details as clean as I could. In this essay, my goal is to give an overview – we will discuss some results, some of the main ideas in the proofs, and some open problems.

Here is an outline of the survey. In Sect. 2, we discuss the combinatorial results that have been proven using ruled surface theory. In Sect. 3, we sketch a proof of the simplest result in Sect. 2. In the course of this sketch, we try to explain some tools

L. Guth (✉)
MIT, 27 Banks Street #2, 02144 Somerville, MA, USA
e-mail: lguth@math.mit.edu

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_18

from ruled surface theory and how those tools help us to understand combinatorial problems. In Sect. 4, we discuss some open problems, exploring what other things we could hope to learn about incidence geometry by using the theory of ruled surfaces.

I would like to thank the anonymous referees for helpful suggestions.

2 Results and Open Questions

In [5], ruled surface theory is used to prove an estimate about the incidence geometry of lines in $\mathbb{R}^3$ (and this estimate eventually leads to estimates about the distinct distance problem). Recall that if $\mathfrak{L}$ is a set of lines, then a point x is an r-rich point of $\mathfrak{L}$ if x lies in at least r lines of $\mathfrak{L}$. We write $P_r(\mathfrak{L})$ for the set of r-rich points of $\mathfrak{L}$. The theorem says that a set of lines in $\mathbb{R}^3$ with many 2-rich points must have some special structure.

Before stating the theorem, we do a couple examples. Because any two lines intersect in at most one point, a set of L lines can have at most $\binom{L}{2}$ 2-rich points. A generic set of lines in the plane achieves this bound. So a set of L lines in $\mathbb{R}^3$ can have at most $\binom{L}{2}$ 2-rich points, and there is an example achieving this bound where all the lines lie in a plane. This suggests the following question: if a set of L lines in $\mathbb{R}^3$ has on the order of L^2 2-rich points, does it have to be the case that many of the lines lie in a plane? Interestingly, the answer is no. The counterexample is based on a degree 2 algebraic surface. Consider the surface defined by the equation

$$z - xy = 0.$$

This surface contains many lines. For any $a \in \mathbb{R}$, the surface contains the line parametrized by

$$t \mapsto (a, t, at).$$

Similarly, for any $b \in \mathbb{R}$, the surface contains the line parametrized by

$$t \mapsto (t, b, tb).$$

If we choose $L/2$ values of a and $L/2$ values of b, we get a set of L lines contained in our surface with $L^2/4$ 2-rich points. Any plane contains at most 2 of these lines. The polynomial $z - xy$ is not unique: there are many other degree 2 polynomials that work equally well.

But in some sense, this is the only counterexample. If a set of L lines in $\mathbb{R}^3$ has many 2-rich points, then it must be the case that many of the lines lie in either a plane or a degree 2 surface. Here is a precise version of this statement.

Theorem 2.1 (Guth and Katz [5]) *There is a constant K so that the following holds. Suppose that $\mathfrak{L}$ is a set of L lines in $\mathbb{R}^3$. Then either*

- $|P_2(\mathfrak{L})| \leq KL^{3/2}$ *or*
- *there is a plane or degree 2 algebraic surface that contains at least $L^{1/2}$ lines of $\mathfrak{L}$.*

By using this theorem repeatedly, we can prove a stronger estimate, which roughly says that if $|P_2(\mathfrak{L})|$ is much bigger than $L^{3/2}$, then almost all of the 2-rich points "come from" planes or degree 2 surfaces.

Corollary 2.2 *Suppose that $\mathfrak{L}$ is a set of L lines in $\mathbb{R}^3$. Then, there are disjoint subsets $\mathfrak{L}_i \subset \mathfrak{L}$ so that*

- *For each i, the lines of $\mathfrak{L}_i$ lie in a plane or a degree 2 surface.*
- $|P_2(\mathfrak{L}) \setminus \cup_i P_2(\mathfrak{L}_i)| \leq KL^{3/2}$.

Proof We prove the corollary by induction on the number of lines. If $|P_2(\mathfrak{L})| \leq KL^{3/2}$, then we are done. Otherwise, by Theorem 2.1, there is a subset $\mathfrak{L}_1 \subset \mathfrak{L}$, so that $|\mathfrak{L}_1| \geq L^{1/2}$ and all the lines of $\mathfrak{L}_1$ lie in a plane or degree 2 surface. We let $\mathfrak{L}' = \mathfrak{L} \setminus \mathfrak{L}_1$. By induction, we can assume the corollary holds for $\mathfrak{L}'$ – giving us disjoint subsets $\mathfrak{L}_i \subset \mathfrak{L}'$. Suppose that a point x is in $P_2(\mathfrak{L}) \setminus \cup_i P_2(\mathfrak{L}_i)$. Then either x lies in a line from $\mathfrak{L}_1$ and a line from $\mathfrak{L}'$, or else $x \in P_2(\mathfrak{L}') \setminus \cup_i P_2(\mathfrak{L}_i)$. The lines of $\mathfrak{L}_1$ all lie in a plane or regulus, and each line of $\mathfrak{L}'$ intersects this plane or regulus at most twice, so the number of points of the first type is at most $2L$. By induction, the number of points of the second type is at most $K|\mathfrak{L}'|^{3/2} \leq K(L - L^{1/2})^{3/2}$. In total, we see that

$$|P_2(\mathfrak{L}) \setminus \cup_i P_2(\mathfrak{L}_i)| \leq 2L + K(L - L^{1/2})^{3/2} \leq KL^{3/2},$$

closing the induction. (In the last step, we have to assume that K is sufficiently large, say $K \geq 100$.) □

Ruled surface theory plays a crucial role in the proof of Theorem 2.1. We will explain how in the next section. Before doing that, we survey generalizations of Theorem 2.1, discussing both known results and open problems.

The first question we explore is the choice of the field $\mathbb{R}$. Does the same result hold over other fields? This question was answered by Kollar in [11]. He first proved that Theorem 2.1 holds over any field of characteristic zero. Next he addressed fields of finite characteristic. As stated, Theorem 2.1 does not hold over finite fields. There is a counterexample over the field $\mathbb{F}_q$ when q is not prime – see [3] for a description of this example. Nevertheless, Kollar proved that Theorem 2.1 does hold over fields of finite characteristic if we add a condition on the number of lines.

Theorem 2.3 (Corollary 40 in [11]) *Suppose that k is any field. Suppose that $\mathfrak{L}$ is a set of L lines in k^3. If the characteristic of k is $p > 0$, then assume in addition that $L \leq p^2$. Then either*

- $|P_2(\mathfrak{L})| \leq KL^{3/2}$ *or*

- *there is a plane or degree 2 algebraic surface that contains at least $L^{1/2}$ lines of $\mathfrak{L}$.*

(In particular, this implies that Theorem 2.1 holds over prime finite fields $\mathbb{F}_p$. The reason is that $|P_2(\mathfrak{L})| \leq |\mathbb{F}_p^3| = p^3$. So if $L \geq p^2$, then $|P_2(\mathfrak{L})| \leq L^{3/2}$ trivially, and if $L \leq p^2$, then Theorem 2.3 applies.)

For context, we compare this situation with the Szemerédi–Trotter theorem, the most fundamental theorem in incidence geometry. The Szemerédi–Trotter theorem says that for a set of L lines in $\mathbb{R}^2$, the number of r-rich points is $\lesssim L^2r^{-3}+Lr^{-1}$. This theorem is also true over $\mathbb{C}^2$, but the proof is much harder – cf. [25] and [26]. The situation over finite fields is not understood and is a major open problem, cf. [1]. In contrast, Theorem 2.3 works equally well over any field. This makes the finite field case of Theorem 2.3 particularly interesting and useful. For instance, Rudnev [18] and Roche–Newton–Rudnev–Shkredov [17] have applied Theorem 2.3 to prove new bounds about the sum-product problem in finite fields. The preprint [19] discusses some other combinatorial problems that can be addressed using Theorem 2.3.

The second question we explore is the role of lines. What happens if we replace lines by circles? Or by other curves in $\mathbb{R}^3$? In [6], Josh Zahl and I proved a version of Theorem 2.1 for algebraic curves of controlled degree.

Theorem 2.4 ([6]) *For any d there are constants $C(d), c_1(d) > 0$ so that the following holds. Suppose that k is any field. Suppose that Γ is a set of L irreducible algebraic curves in k^3 of degree at most d. If the characteristic of k is $p > 0$, then assume in addition that $L \leq c_1(d)p^2$. Then either*

- $|P_2(\mathfrak{L})| \leq C(d)L^{3/2}$ *or*
- *there is an algebraic surface of degree at most $100d^2$ that contains at least $L^{1/2}$ curves of $\mathfrak{L}$.*

There are a couple reasons why I think it is natural to consider various algebraic curves instead of just straight lines. One reason is that the proof is closely based on algebraic geometry. Once we have a good understanding of the ideas involved, they apply naturally to all algebraic curves. A second reason is that this more general result will probably have more applications. For instance, we recall a little about the distinct distance problem in the plane. In [2] Elekes and Sharir suggested an interesting new approach to the problem, connecting distinct distances in the plane to problems about the incidence geometry of some degree 2 algebraic curves in $\mathbb{R}^3$. In [5], there is a clever change of coordinates so that these degree 2 curves become lines, and then Theorem 2.1 applies to bound the number of 2-rich points. It appears to me that this clever change of coordinates was rather fortuitous. I believe that most problems about algebraic curves cannot be reduced to the straight line case by a change of coordinates, and I think that when results along the lines of Theorem 2.4 arise in applications, the curves involved will only sometimes be straight lines. Theorem 2.4 applies to the problem about degree 2 curves from [2], and I think it will probably have more applications in the future.

The third question that we discuss is what happens in higher dimensions. The situation in higher dimensions is not yet understood. The following conjecture seems natural to me. (Similar questions were raised in [21] and [27]).

Conjecture 2.5 *Let k be any field. Suppose that Γ is a set of L irreducible algebraic curves in k^n, of degree at most d. If the characteristic of k is $p > 0$, then also assume that $L \le p^{n-1}$. Then either*

- $|P_2(\Gamma)| \le C(d,n)L^{\frac{n}{n-1}}$ *or*
- *There is a dimension $2 \le m \le n-1$, and an algebraic variety Z of dimension m and degree at most $D(d,n)$ so that Z contains at least $L^{\frac{m-1}{n-1}}$ curves of Γ.*

There is some significant progress on this conjecture in four dimensions. In [22], Sharir and Solomon prove estimates for r-rich points of a set of lines in $\mathbb{R}^4$. These estimates only apply for fairly large r, not $r = 2$, so they don't literally address this conjecture, but they establish sharp bounds in a similar spirit for larger values of r. In [7], Josh Zahl and I prove a slightly weaker estimate of this form for algebraic curves in $\mathbb{R}^4$. So far nothing close to this conjecture is known for lines in $\mathbb{C}^4$ or over $\mathbb{F}_p^4$. Moreover, nothing close to this conjecture is known in higher dimensions. I think that this is a natural question, and that if it is true, it would probably have significant applications. If it is false, that would also be interesting, and it would point to new subtleties in incidence geometry in higher dimensions.

In the next section, we discuss the proofs of the known results. Afterwards, we come back and discuss how much these proofs can tell us about higher dimensions, and what new issues arise.

3 How Does Ruled Surface Theory Help in the Proof

In this section, we discuss some of the ideas in the proofs of the results from the last section. The ideas we want to discuss are easiest to explain over $\mathbb{C}$, so we first state a version of Theorem 2.1 over $\mathbb{C}$.

Theorem 3.1 *There is a large constant K so that the following holds. Suppose that $\mathfrak{L}$ is a set of L lines in $\mathbb{C}^3$. Then either*

- $|P_2(\mathfrak{L})| \le KL^{3/2}$ *or*
- *there is a plane or degree 2 algebraic surface that contains at least $L^{1/2}$ lines of $\mathfrak{L}$.*

To get started, we think a little about the role of planes and degree 2 algebraic surfaces. What is special about planes and degree 2 algebraic surfaces that makes them appear here? Planes and degree 2 surfaces are doubly ruled. A ruled surface is an algebraic surface that contains a line through every point. A doubly ruled surface is a surface that contains two distinct lines through every point.

At this point, we can say a little about the connection between ruled surface theory and incidence geometry. A doubly ruled surface can be roughly thought of as an algebraic family of lines with many 2-rich points. In incidence geometry, one tries to classify finite sets of lines with many 2-rich points. In ruled surface theory, one tries to classify doubly ruled surfaces – that is, algebraic families of lines with many 2-rich points. To prove Theorem 3.1, we begin with a finite set of lines with many 2-rich points, and we build around it a whole doubly ruled surface. Tools from ruled surface theory help to build this surface and they help to analyze the surface once it is built, ultimately leading to information about the original finite set of lines.

Doubly ruled algebraic surfaces in $\mathbb{C}^3$ were classified in the nineteenth century. It turns out that planes and degree 2 surfaces are the only irreducible doubly ruled surfaces. These surfaces appear in the statement of Theorem 3.1 because they are the only irreducible doubly ruled surfaces. Roughly speaking, a doubly ruled surface is an algebraic family of lines with many 2-rich points. Theorem 3.1 is telling us that a finite configuration of lines with many 2-rich points must be related to an algebraic family of lines with many 2-rich points.

At this point, let us pause to review some vocabulary from algebraic geometry that we will use in the rest of the essay. After we set up this vocabulary, we can state things precisely, starting with the classification of doubly ruled surfaces in $\mathbb{C}^3$.

An algebraic set in $\mathbb{C}^n$ is the set of common zeroes of a finite list of polynomials in $\mathbb{C}[z_1, \ldots, z_n]$. An algebraic set is called reducible if it is the union of two proper algebraic subsets. Otherwise it is called irreducible. An irreducible algebraic set in $\mathbb{C}^n$ is also called an affine variety.

Any affine variety V in $\mathbb{C}^n$ has a dimension. The dimension of V is the largest number r so that there is a sequence of proper inclusions of non-empty varieties $V_0 \subset \ldots \subset V_r = V$. The dimension of an algebraic set in $\mathbb{C}^n$ is the maximum dimension of any irreducible subset. An algebraic curve is a variety of dimension 1.

Using the dimension, we can define a useful notion of the generic behavior of points in a variety. We say that a generic point of an algebraic variety V obeys condition (X) if the set of points $p \in V$ where (X) does not hold is contained in an algebraic subset $E \subset V$ with $\dim E < \dim V$. For instance, we say that a 2-dimensional algebraic variety $\Sigma \subset \mathbb{C}^3$ is generically doubly ruled if there is a 1-dimensional algebraic set $\gamma \subset \Sigma$, and every point of $\Sigma \setminus \gamma$ is contained in two lines in Σ.

An affine variety also has a degree. There is a non-trivial theorem which says that for any affine variety V in $\mathbb{C}^n$ there is unique choice of r and d so that a generic $(n-r)$-plane in $\mathbb{C}^n$ intersects V in exactly d points. The value of r is the dimension of V, as defined above. The value of d is called the degree of V.

There is a nice short summary of facts about dimension and degree in Section 4 of [24], which contains everything we have mentioned here. A fuller treatment appears in Harris's book on algebraic geometry [8].

This is all the terminology that we will need, and we now return to discussing doubly ruled surfaces. We can now state a classification theorem for double ruled surfaces in $\mathbb{C}^3$.

Theorem 3.2 (Classification of doubly ruled surfaces, cf. Proposition 13.30 in [4]) *Suppose that $P \in \mathbb{C}[z_1, z_2, z_3]$ is an irreducible polynomial and that $Z(P)$ is generically doubly ruled. Then P has degree 1 or 2, and so $Z(P)$ is a plane or a degree 2 algebraic surface.*

There are three somewhat different proofs of Theorem 3.1 in the literature – in [5], in [11], and in [6]. All three proofs use ruled surface theory in a crucial way, and this is the aspect that we will focus on. Other parts of the argument are somewhat different in the three proofs. The proof I want to outline here is the one from [6]. Another reference is my book on polynomial methods in combinatorics, [4], which will be published in the near future by the AMS. In the chapter on ruled surfaces in [4], I give a detailed proof of Theorem 3.1 using this method.

For this sketch, let us suppose that each line of $\mathfrak{L}$ contains about the same number of 2-rich points. This is the most interesting case of Theorem 3.1. So each line contains about $KL^{1/2}$ points of $P_2(\mathfrak{L})$. I want to highlight three stages in the proof, which I discuss in three subsections.

3.1 Degree Reduction

The first step of the argument is to find a (non-zero) polynomial P that vanishes on the lines of $\mathfrak{L}$ with a good bound on the degree of P. For reference, given any set of N points in $\mathbb{C}^3$, there is a non-zero polynomial that vanishes on all these points with degree at most about $N^{1/3}$. For a generic set of points, this bound is sharp. By a similar argument, for any set of L lines in $\mathbb{C}^3$, there is a non-zero polynomial that vanishes on the lines with degree at most about $L^{1/2}$. For a generic set of lines, this bound is also sharp.

Given that each line of $\mathfrak{L}$ contains about $KL^{1/2}$ lines of $P_2(\mathfrak{L})$, we show that there is a non-zero polynomial P vanishing on all the lines of $\mathfrak{L}$ with degree $O(K^{-1}L^{1/2})$. As long as K is large enough, this degree is well below the degree required for a generic set of lines. This shows that, compared to a generic set of lines, the set $\mathfrak{L}$ has a little algebraic structure.

Even though the degree of P is only a little smaller than the trivial bound $L^{1/2}$, this small improvement turns out to be a crucial clue to the structure of $\mathfrak{L}$, and it eventually leads to a much more precise description of P: P is a product of irreducible polynomials of degrees 1 and 2. Once we know this structure for the polynomial P, the conclusion of the theorem is easy: the lines of $\mathfrak{L}$ are contained in $O(K^{-1}L^{1/2})$ planes and degree 2 algebraic surfaces. By pigeonholing, one of these surfaces must contain at least $L^{1/2}$ lines of $\mathfrak{L}$.

Here is the idea of the degree bound for P. We randomly pick a subset $\mathfrak{L}' \subset \mathfrak{L}$ with $L' \leq L$ lines, where L' is a parameter that we can tune later. Then we find a non-zero polynomial P that vanishes on the lines of $\mathfrak{L}'$ with degree at most $C(L')^{1/2}$. (We will eventually choose L' so that this bound is $CK^{-1}L^{1/2}$.) If L' is big enough, then with high probability the polynomial P actually vanishes on all the lines of $\mathfrak{L}$. Here is the mechanism that makes this vanishing happen, which I call contagious vanishing. By hypothesis, each line $l \in \mathfrak{L}$ contains at least $KL^{1/2}$ 2-rich points of $\mathfrak{L}$. With high probability many of these points will lie in lines of $\mathfrak{L}'$. The polynomial P vanishes at every point where l intersects a line of $\mathfrak{L}'$. If the number of these points is more than the degree of P, then P must vanish on the line l also. If we choose L' carefully, then this mechanism will force P to vanish on all the lines of $\mathfrak{L}$. Carrying out the details of this argument, the numbers work out so that the degree of P is at most $CK^{-1}L^{1/2}$ – cf. Proposition 11.5 in [4].

At this point, we factor P into irreducible factors $P = \prod_j P_j$. Each line of $\mathfrak{L}$ must lie in $Z(P_j)$ for at least one j. We let $\mathfrak{L}_j \subset \mathfrak{L}$ be the set of lines of $\mathfrak{L}$ that lie in $Z(P_j)$. We subdivide the 2-rich points as

$$P_2(\mathfrak{L}) = \cup_j P_2(\mathfrak{L}_j) \bigcup \text{“mixed 2-rich points”},$$

where a mixed 2-rich point is the intersection point of some line $l \in \mathfrak{L}_j$ with some line $l' \notin \mathfrak{L}_j$. A line not in $\mathfrak{L}_j$ can intersect $Z(P_j)$ at most $\operatorname{Deg} P_j$ times. Therefore, the total number of mixed 2-rich points is at most $L(\sum_j \operatorname{Deg} P_j) = L \operatorname{Deg} P = O(K^{-1}L^{3/2})$, only a small fraction of the total number of 2-rich points. By factoring the polynomial P we have broken the original problem of understanding $\mathfrak{L}$ into essentially separate subproblems of understanding each set $\mathfrak{L}_j$.

The most difficult case is when P is irreducible. The general case can be reduced to this case by studying the set of lines $\mathfrak{L}_j$ and the polynomial P_j. From now on we assume that P is irreducible. It remains to show that P has degree 1 or 2.

3.2 Ruled Surface Theory

In this subsection, we discuss some tools from ruled surface theory and how they help up to understand the polynomial P in our proof sketch.

At this point, we know that there is a polynomial P that vanishes on the lines of $\mathfrak{L}$ with degree significantly smaller than $L^{1/2}$, and we are focusing on the case where P is irreducible. Using this little bit of structure, we are going to find out a lot more about the polynomial P and its zero set $Z(P)$. Ultimately, we will see that P has degree 1 or 2. In this subsection, we sketch how to prove that $Z(P)$ is generically doubly ruled.

For each 2-rich point $x \in P_2(\mathfrak{L})$, the point x lies in two lines in $Z(P)$. Since $\mathfrak{L}$ has many 2-rich points, we know that there are many points in $Z(P)$ that lie in two lines in $Z(P)$ – there are many points where $Z(P)$ “looks doubly-ruled”. Based on this we will show that almost every point of $Z(P)$ lies in two lines in $Z(P)$. Loosely speaking, the property of “looking doubly-ruled” is contagious – it spreads from the 2-rich points of $\mathfrak{L}$ and fills almost every point of $Z(P)$. The tools to understand why this property is contagious come from ruled surface theory.

The first topic from ruled surface theory that we introduce is flecnodal points. A point $z \in Z(P)$ is flecnodal if there is a line l through z which is tangent to $Z(P)$

to third order. Here is a more formal definition, which also makes sense if z is a singular point of $Z(P)$, where it's not immediately obvious what tangent to $Z(P)$ means. A point $z \in Z(P)$ is flecnodal if there is a line l with tangent vector v so that

$$0 = P(z) = \partial_v P(z) = \partial_v^2 P(z) = \partial_v^3 P(z).$$

Here we write ∂_v for the directional derivative in direction v:

$$\partial_v := \sum_{i=1}^{3} v_i \frac{\partial}{\partial z_i},$$

and we write ∂_v^k to denote repeatedly applying this differentiation – for instance,

$$\partial_v^2 P := \partial_v \left(\partial_v P\right).$$

If a point z lies in a line in $Z(P)$, then it follows immediately that z is flecnodal. Flecnodal points are useful because they also have a more algebraic description. A basic theme of algebraic geometry is to take any geometric property of a surface, and describe it in an algebraic way, in terms of the vanishing of some polynomials.

Theorem 3.3 (Salmon [20] Art. 588 pages 277–78) *For any polynomial* $P \in \mathbb{C}[z_1, z_2, z_3]$*, there is a polynomial* $\operatorname{Flec} P \in \mathbb{C}[z_1, z_2, z_3]$ *so that*

- *A point* $z \in Z(P)$ *is flecnodal if and only if* $\operatorname{Flec} P(z) = 0$.
- $\operatorname{Deg} \operatorname{Flec} P \leq 11 \operatorname{Deg} P$.

(For some discussion of the history of this result, see the paragraph after Remark 12 in [11].)

Our goal is to connect 2-rich points and doubly-ruled surfaces, so we introduce a doubly-ruled analogue of being flecnodal. We say that a point $z \in Z(P)$ is doubly flecnodal if there are two (distinct) lines l_1, l_2 through z, with tangent vectors v_1, v_2, so that for each $i = 1, 2$,

$$0 = P(z) = \partial_{v_i} P(z) = \partial_{v_i}^2 P(z) = \partial_{v_i}^3 P(z).$$

Doubly flecnodal points were first introduced in [6] and [4]. There is an analogue of Salmon's theorem for doubly flecnodal polynomials – cf. Proposition 13.3 in [4]. It is a little more complicated to state. Instead of one flecnodal polynomial, there is a finite list of them.

Theorem 3.4 *There are universal constants* J *and* C *so that the following holds. For any polynomial* $P \in \mathbb{C}[z_1, z_2, z_3]$*, there is a finite list of polynomials* $\operatorname{Flec}_{2,j} P$*, with* $1 \leq j \leq J$*, and a Boolean function* $\Phi : \{0, 1\}^J \to \{0, 1\}$ *so that the following holds.*

- *For each* j, $\operatorname{Deg} \operatorname{Flec}_{2,j} P \leq C \operatorname{Deg} P$.

- *Let* $V_{2,j}P(z)$ *be equal to zero if* $\mathrm{Flec}_{2,j}P(z) = 0$ *and equal to 1 otherwise. Then* z *is a doubly flecnodal point of* $Z(P)$ *if and only if*

$$\Phi\left(V_{2,1}P(z), \dots, V_{2,J}P(z)\right) = 0.$$

This theorem sounds more complicated than Salmon's theorem, but in the applications we're about to describe, it is essentially equally useful.

Because flecnodal and doubly flecnodal points have this algebraic description, they behave contagiously. We start with the flecnodal points and then discuss the doubly flecnodal points. We know that each line contains $KL^{1/2}$ 2-rich points of $\mathfrak{L}$. At each of these points, $\mathrm{Flec}\, P$ vanishes. The degree of P is at most $CK^{-1}L^{1/2}$, and the degree of $\mathrm{Flec}\, P$ is at most $11\, \mathrm{Deg}\, P \leq C'K^{-1}L^{1/2}$. As long as we choose K large enough, the number of points is larger than $\mathrm{Deg}\, \mathrm{Flec}\, P$ and it follows that $\mathrm{Flec}\, P$ vanishes along each line of $\mathfrak{L}$. Actually, since the lines of $\mathfrak{L}$ are contained in $Z(P)$, we already know that every point of each line is flecnodal, but we included the last discussion as a warmup for doubly flecnodal points. Now we know that $\mathrm{Flec}\, P$ vanishes on all L lines of $\mathfrak{L}$. By a version of the Bezout theorem (cf. Theorem 6.7 in [4]), $Z(P) \cap Z(\mathrm{Flec}\, P)$ can contain at most $\mathrm{Deg}\, P \cdot \mathrm{Deg}\, \mathrm{Flec}\, P$ lines, unless P and $\mathrm{Flec}\, P$ have a common factor. Because $\mathrm{Deg}\, P$ and $\mathrm{Deg}\, \mathrm{Flec}\, P$ are much less than $L^{1/2}$, we see that P and $\mathrm{Flec}\, P$ must indeed have a common factor. Since P is irreducible, P must divide $\mathrm{Flec}\, P$. Therefore $\mathrm{Flec}\, P$ vanishes on $Z(P)$, and every point of $Z(P)$ is flecnodal!

Doubly flecnodal points are contagious for a similar reason. We just do the first step of the argument. There are J polynomials $\mathrm{Flec}_{2,j}\, P$. For each point z, there are 2^J possible values for the vector $(V_{2,1}P(z), \dots, V_{2,J}P(z))$. Fix a line $l \in \mathfrak{L}$. By hypothesis, l contains at least $KL^{1/2}$ points of $P_2(\mathfrak{L})$. Now, by the pigeonhole principle, we can find a vector $\sigma \in \{0,1\}^J$ and a subset $X_\sigma \subset P_2(\mathfrak{L}) \cap l$ so that

- for each point $z \in X_\sigma$, $V_{2,j}P(z) = \sigma_j$.
- $|X_\sigma| \geq 2^{-J}KL^{1/2}$.

Because every point of X_σ is doubly flecnodal, we see that $\Phi(\sigma) = 0$. We choose the constant K significantly larger than 2^{-J}, and so $|X_\sigma| > \mathrm{Deg}\, \mathrm{Flec}_{2,j}\, P$ for each j. Therefore, if $\sigma_j = 0$, then $\mathrm{Flec}_{2,j}\, P$ vanishes on the whole line l. If $\sigma_j = 1$, then $\mathrm{Flec}_{2,j}\, P$ does not vanish on the whole line l, and so it vanishes at only finitely many points of l. Therefore, for almost every $z \in l$, $\mathrm{Flec}_{2,j}\, P(z)$ vanishes if and only if $\sigma_j = 0$. In other words, at a generic point of the line l, $V_{2,j}P(z) = \sigma_j$. Therefore, at a generic point of l, $\Phi(V_{2,1}P(z), \dots, V_{2,J}P(z)) = \Phi(\sigma) = 0$, and so a generic point of l is doubly flecnodal. Next, by making a similar argument to the flecnodal case above, one can show that a generic point of $Z(P)$ is doubly flecnodal.

We have now sketched the proof that $Z(P)$ is generically doubly flecnodal. We are starting to see how the combinatorial information that $\mathfrak{L}$ has many 2-rich points implies that $Z(P)$ must have a special structure.

Just because a point $z \in Z(P)$ is flecnodal, it doesn't mean that z lies in a line in $Z(P)$. For instance, let P be the polynomial $P(z) = z_1^{10} + z_2^{10} + z_3^{12} - 1$ and let z be the point $(1,0,0) \in Z(P)$. If l is a line through z parallel to the (z_2, z_3)-plane, then l is tangent to $Z(P)$ to ninth order. So there are infinitely many different lines through z that are tangent to $Z(P)$ to ninth order, but none of them lies in $Z(P)$. This kind of behavior can indeed occur at some special points of $Z(P)$, but it turns out that it cannot happen at a generic point of $Z(P)$.

Theorem 3.5 (Cayley–Salmon–Monge) *If $P \in \mathbb{C}[z_1, z_2, z_3]$, and if every point of $Z(P)$ is flecnodal, then $Z(P)$ is a ruled surface – every point of $Z(P)$ lies in a line in $Z(P)$.*

(For the history of this theorem and a sketch of the proof, see the discussion around Theorem 13 in [11].)

There is also a version of this result for doubly flecnodal points (and in fact it is a little easier):

Theorem 3.6 (cf. Proposition 13.30 in [4]) *If $P \in \mathbb{C}[z_1, z_2, z_3]$, and if $Z(P)$ is generically doubly flecnodal, then $Z(P)$ is generically doubly ruled.*

This theorem implies that our surface $Z(P)$ is generically doubly ruled.

There are several sources to read more about ruled surface theory and about the details of the arguments we have sketched here. I tried to write readable self-contained proofs in the chapter on ruled surface theory in [4]. In Kollar's paper [11], there is a discussion of the proof of Theorem 3.5 and also some history. In Katz's ICM talk [10], there is another discussion of the proof of Theorem 3.5. Finally, [6] gives a quite different proof of Theorem 3.5 which generalizes to algebraic curves in place of straight lines. For ruled surfaces in general, the referee suggested the classical work of Plucker [15] and the modern book [16].

This may be a good moment to say a bit more about the theorem in [6]. Suppose that Γ is a set of L circles in $\mathbb{R}^3$. For the case of circles, what kind of surfaces should play the role of planes and degree 2 surfaces? We say that a surface $Z(P)$ is generically doubly ruled by circles if a generic point of $Z(P)$ lies in two distinct circles in $Z(P)$. In [6], it is proven that either $|P_2(\Gamma)| \le KL^{3/2}$ or Γ contains at least $L^{1/2}$ circles in an algebraic surface $Z(P)$ which is generically doubly ruled by circles. The same holds if circles are replaced by other classes of curves, such as parabolas, degree 3 curves, etc. The proof follows the same outline that we have given here, and the main difficulty in the paper is to generalize the tools of ruled surface to other classes of curves.

The definition of flecnodal and doubly flecnodal involve three derivatives. The reader may wonder why we use three derivatives. In fact, using more than three derivatives would also work. Using r derivatives instead of three derivatives, we can define r-flecnodal points and doubly r-flecnodal points. Theorem 3.4 holds for any choice of r – only the constants C and J depend on r – cf. Proposition 13.3 in [4]. Three derivatives is the minimum number of derivatives necessary to prove Theorems 3.5 and 3.6. These theorems would be false if we assumed that only two derivatives vanish. Here is a dimensional heuristic why three derivatives are important (suggested by the referee). Fix a point z in $Z(P)$. In three dimensions, the space of lines through x is a 2-dimensional space. If we insist that r derivatives of P vanish in the tangent direction of a line, this gives us r equations on the space of lines. For $r = 2$, dimensional heuristics suggest that there will typically be such a line. But for $r = 3$, dimensional heuristics suggest that there will not be typically be such a line. Indeed the theory of ruled surfaces shows that these heuristics are correct – for a generic polynomial $P \in \mathbb{C}[z_1, z_2, z_3]$, every point of $Z(P)$ is 2-flecnodal, but the subset of 3-flecnodal points is a lower-dimensional subvariety.

3.3 Classification of Doubly Ruled Surfaces

At this point in our sketch, we have shown that $Z(P)$ is generically double ruled, and we know that P is irreducible. To finish the proof of Theorem 3.1, we have to prove that P has degree 1 or 2. This follows from the classification of (generically) doubly ruled surfaces in Theorem 3.2.

To end our sketch, we briefly discuss the proof of the classification Theorem 3.2. In fact, there is a more general classification theorem for degree d algebraic curves, which we discuss at the same time.

Theorem 3.7 ([6]) *Suppose that $P \in \mathbb{C}[z_1, z_2, z_3]$ is an irreducible polynomial, and that $Z(P)$ is generically doubly ruled by algebraic curves of degree at most d. Then* $\mathrm{Deg}\, P \leq 100d^2$.

Because a generic point of $Z(P)$ lies in two algebraic curves in $Z(P)$, it is not hard to find many algebraic curves in $Z(P)$ that intersect each other in many places. More precisely, we can find two arbitrarily large families of curves $\gamma_{1,i}$ and $\gamma_{2,j}$ in $Z(P)$, so that for each pair i, j, $\gamma_{1,i}$ intersects $\gamma_{2,j}$, and all the intersection points are distinct – cf. Lemma 11.8 in [6]. The proof strongly uses the fact that $Z(P)$ is 2-dimensional. The idea of the argument is to study the curves passing through a small ball in $Z(P)$. For the sake of this sketch, let us suppose that each point $z \in Z(P)$ lies in exactly two algebraic curves of degree d, $\gamma_1(z)$ and $\gamma_2(z)$. Let us suppose that these curves vary smoothly with z, and let us suppose that $\gamma_1(z)$ and $\gamma_2(z)$ intersect transversely at z. (This is the moment where we use that the dimension of $Z(P)$ is 2 – if the dimension of $Z(P)$ is greater than 2, then two curves can never intersect transversely.) We fix a smooth point $z_0 \in Z(P)$, and then we let z_i and w_j be a generic sequence of points of $Z(P)$ very close to z_0. The curves $\gamma_{1,i}$ and $\gamma_{2,j}$ are just $\gamma_1(z_i)$ and $\gamma_2(w_j)$. Since z_i and w_j are very close to z_0, then $\gamma_{1,i}$ and $\gamma_{2,j}$ are small perturbations of $\gamma_1(z_0)$ and $\gamma_2(z_0)$. Since $\gamma_1(z_0)$ and $\gamma_2(z_0)$ intersect transversely at z_0, then $\gamma_{1,i}$ and $\gamma_{2,j}$ must intersect at a point close to z_0.

Once we have the curves $\gamma_{1,i}$ and $\gamma_{2,j}$ we can bound the degree of P by using a contagious vanishing argument. For any degree D, we can choose a polynomial Q of degree at most D that vanishes on roughly $D^2 d^{-1}$ of the curves $\gamma_{1,i}$. On the other hand, if $\gamma_{2,j}$ does not lie in $Z(Q)$, then Q can vanish on at most dD points of $\gamma_{2,j}$. We choose D so that $D^2 d^{-1} \gg dD$. Choosing $D = 100d^2$ is big enough. Since Q vanishes on $D^2 d^{-1}$ curves $\gamma_{1,i}$, it vanishes at $D^2 d^{-1}$ points of each curve $\gamma_{2,j}$, and so it vanishes on each curve $\gamma_{2,j}$. Now we see that $Z(Q) \cap Z(P)$ contains infinitely many algebraic curves $\gamma_{2,j}$. By the Bezout theorem, P and Q must have a common factor. Since P is irreducible, P must divide Q. But then $\mathrm{Deg}\, P \leq \mathrm{Deg}\, Q \leq 100d^2$.

This degree reduction argument is essentially the same as the one in Sect. 3.1, but we get a better bound for the degree because the curves $\gamma_{1,i}$ and $\gamma_{2,j}$ have so many 2-rich points. Here is a big picture summary of the proof of Theorem 3.1. First we used the combinatorial information to prove that the set of lines $\mathfrak{L}$ has a little algebraic structure – the lines lie in $Z(P)$ where the degree of P is a bit smaller than for generic lines. If P is reducible, we divide the problem into essentially disjoint subproblems, and we assume from now on that P is irreducible. Second, we use the degree bound on P and the combinatorial information about the lines to prove that $Z(P)$ is generically doubly ruled. So our finite set of lines $\mathfrak{L}$ fits into an algebraic family of lines with many 2-rich points. Third, we extend $\mathfrak{L}$ by adding a lot of other lines from the surface $Z(P)$. By doing this, we can amplify the number of 2-rich points. We get a new set of $N \gg L$ lines in $Z(P)$ with around N^2 2-rich points. Finally, we apply degree reduction to this bigger set of lines, and we get a much stronger estimate for the degree of P.

4 Thoughts About Higher Dimensions

In this last section, we reflect on how much ruled surface theory can tell us about incidence geometry in higher dimensions, and we point out some open problems. What happens if we try to adapt the proof of Theorem 3.1 that we just sketched to higher dimensions? We broke the proof of Theorem 3.1 into three stages. We discuss each stage, but especially focusing on the last stage – the classification of doubly ruled surfaces.

We suppose that $\mathfrak{L}$ is a set of L lines in $\mathbb{C}^n$. We suppose that $|P_2(\mathfrak{L})| \geq KL^{\frac{n}{n-1}}$. We also make the minor assumption that each line contains about the same number of 2-rich points: so each line contains at least $KL^{\frac{1}{n-1}}$ points of $P_2(\mathfrak{L})$.

4.1 *Degree Reduction*

The degree reduction stage works in any dimension. In n dimensions, for any set of N points, there is a polynomial of degree at most $C_n N^{1/n}$ vanishing on the set, and this bound is sharp for generic sets. For any set of L lines, there is a polynomial of degree at most $C_n L^{\frac{1}{n-1}}$ vanishing on each line, and this bound is sharp for generic sets of lines. But if each line of $\mathfrak{L}$ contains at least $KL^{\frac{1}{n-1}}$ 2-rich points of $\mathfrak{L}$, then there is a polynomial P vanishing on the lines of $\mathfrak{L}$ with degree at most $C_n K^{\frac{-1}{n-2}} L^{\frac{1}{n-1}}$. So we see that in any number of dimensions, if K is large enough then $\mathfrak{L}$ has some algebraic structure. I think this suggests that it is a promising avenue to try to study $\mathfrak{L}$ using algebraic geometry.

4.2 Ruled Surface Theory

Some of the tools we used in the second stage have generalizations to higher dimensions. Landsberg [12] has proven a version of Theorem 3.5 in any number of dimensions. Sharir and Solomon [22] generalized the flecnode polynomial to four dimensions and proved the four-dimensional analogue of Theorem 3.3. Double-flecnode polynomials have so far only been defined in three dimensions. In higher dimensions, there is one technical point which will be more difficult. In $\mathbb{C}^n$, there are doubly ruled varieties of every dimension between 2 and $n-1$. Therefore, it is not enough to consider algebraic hypersurfaces, which can be written in the form $Z(P)$ for a single polynomial P – we have to consider algebraic varieties of all dimensions. If one could generalize the methods in this second stage to higher dimensions, it might be possible to prove the following conjecture.

Conjecture 4.1 *Suppose that $\mathfrak{L}$ is a set of L lines in $\mathbb{C}^n$. Then either*

- $|P_2(\mathfrak{L})| \leq C(n)L^{\frac{n}{n-1}}$ *or*
- *There is a dimension $2 \leq m \leq n-1$, and a generically double-ruled affine variety Z of dimension m so that Z contains at least $L^{\frac{m-1}{n-1}}$ lines of $\mathfrak{L}$. (Recall that an affine variety is irreducible by definition.)*

We can generalize this conjecture to algebraic curves as follows.

Conjecture 4.2 *Suppose that Γ is a set of L irreducible algebraic curves in $\mathbb{C}^n$ of degree at most d. Then either*

- $|P_2(\Gamma)| \leq C(d,n)L^{\frac{n}{n-1}}$ *or*
- *There is a dimension $2 \leq m \leq n-1$, and an (irreducible) affine variety Z of dimension m which is generically doubly ruled by algebraic curves of degree at most d and contains at least $L^{\frac{m-1}{n-1}}$ curves of Γ.*

If Conjectures 4.1 and/or 4.2 is true, it would point to a strong connection between incidence geometry and ruled surface theory. On the other hand, it would probably not be useful in applications unless we could also prove a classification of doubly ruled varieties – at least a very rough classification. So let us turn now to the problem of the classification of doubly ruled varieties.

4.3 Classification of Doubly Ruled Varieties

The classification of doubly ruled surfaces in $\mathbb{C}^3$ was fairly simple, but in higher dimensions, this part of the problem may become a lot more complex. I would like to propose a question about doubly ruled varieties that could be useful to understand for applications to incidence geometry.

To get started, we might ask, if $Y^m \subset \mathbb{C}^n$ is a generically doubly ruled (irreducible) variety, does it follow that $\operatorname{Deg} Y \leq C(n)$? The answer to this question is no. It may happen that every point of Y lies in a 2-plane in Y. Such a variety is clearly doubly ruled, and it may have an arbitrarily high degree. For a high degree example, suppose that Y is a graph of the form

$$z_4 = P_1(z_3)z_1 + P_2(z_3)z_2 + Q(z_3),$$

where P_1, P_2, and Q are polynomials of high degree. If $w = (w_1, w_2, w_3, w_4) \in Y$, then w lies in the following 2-plane in Y:

$$z_3 = w_3; z_4 = P_1(w_3)z_1 + P_2(w_3)z_2 + Q(w_3).$$

If P_1, P_2, or Q have high degree, then Y will have high degree also. (Also the algebraic set Y is in fact irreducible for any chocie of P_1, P_2, Q.)

Suppose for a moment that the variety Y that we find in the second stage is a graph of this form, and suppose for simplicity that every line of $\mathfrak{L}$ lies in Y. For a typical P_1, P_2, Q, every line in Y is contained in one of the planes above. Suppose for a moment that our variety Y has this convenient property. Then we can separate the lines of $\mathfrak{L}$ into subsets corresponding to different 2-planes. Since each line of $\mathfrak{L}$ contains at least $KL^{\frac{1}{n-1}}$ 2-rich points, one of the 2-planes must contain at least $KL^{\frac{1}{n-1}}$ lines of $\mathfrak{L}$, and this satisfies the conclusion of Conjecture 2.5.

I don't know whether there are more exotic examples of doubly ruled varieties than this one. Let me introduce a little vocabulary so that I can make an exact question. We say that a variety Y is ruled by varieties with some property $(*)$ if each point $y \in Y$, lies in a variety $X \subset Y$ where X has property $(*)$. We say that a variety Y is doubly ruled by varieties with property $(*)$ if each point y lies in two distinct varieties $X_1, X_2 \subset Y$ with property $(*)$. We say that Y is generically ruled by varieties with property $(*)$ if a generic point $y \in Y$ lies in a variety $X \subset Y$ with property $(*)$, and so on.

Question 4.3 *Suppose that $Y \subset \mathbb{C}^n$ is a variety which is generically doubly ruled (by lines). Does it follow that Y is generically ruled by varieties with dimension at least 2 and degree at most $C(n)$?*

To the best of my knowledge this question is open. Noam Solomon pointed me to a relevant paper in the algebraic geometry literature by Mezzetti and Portelli [14]. Under a technical condition, this paper gives a classification of doubly ruled 3-dimensional varieties in $\mathbb{CP}^4$ – see Theorem 0.1. The technical condition is that the Fano scheme of lines of Y is generically reduced. If Y is generically doubly ruled and obeys this condition, then the classification from Theorem 0.1 of [14] implies that either Y has degree at most 16 or Y is generically ruled by 2-dimensional varieties of degree at most 2.

We can also pose more general questions in a similar spirit to Question 4.3.

Question 4.4 *Suppose that $Y \subset \mathbb{C}^n$ is generically doubly ruled by (irreducible) algebraic curves of degree at most d. Does it follow that Y is generically ruled by varieties with dimension at least 2 and degree at most $C(d, n)$?*

Question 4.5 *Suppose that $Y \subset \mathbb{C}^n$ is generically doubly ruled by varieties of dimension m and degree at most d. Does it follow that Y is generically ruled by varieties with dimension at least $m + 1$ and degree at most $C(d, m, n)$?*

If the answers to Questions 4.3 and 4.4 are affirmative, then I think it would be promising to try to prove Conjecture 2.5 using tools from ruled surface theory. If the answer to Question 4.3 is no, then it means that there are some exotic doubly ruled varieties $Y \subset \mathbb{C}^n$. These varieties would be a potential source of new examples in incidence geometry, and could possibly lead to counterexamples to Conjecture 2.5.

For a given variety Y containing many lines, it looks interesting to explore incidence geometry questions for sets of lines in Y. This circle of questions was raised by Sharir and Solomon in [22]. In particular, they raised the following question.

Question 4.6 *Suppose that Y is the degree 2 hypersurface in $\mathbb{R}^4$ defined by the equation*

$$x_1 = x_2^2 + x_3^2 - x_4^2.$$

For a given r, what is the maximum possible size of $|P_r(\mathfrak{L})|$?

This question was studied by Solomon and Zhang in [23], building on earlier work of Zhang [27]. They constructed an example with many r-rich points. Counting the number of r-rich points in the example is non-trivial and they used tools from analytic number theory to do so. Their construction gives $\sim L^{3/2}r^{-3}$ r-rich points. Since a generic point of Y lies in infinitely many lines in Y, it is easy to produce examples with $\sim Lr^{-1}$ r-rich points, so their example is interesting when r is smaller than $L^{1/4}$. The best known upper bound on $|P_r(\mathfrak{L})|$ is based on a random projection argument. Rudnev used a closely related random projection argument in [18] – cf. the bottom of page 6 of [18]. We note that Y does not contain any 2-plane. Since Y is the zero set of a degree 2 polynomial, the intersection of Y with a 2-plane may contain at most two lines. Therefore, we see that $\mathfrak{L}$ contains at most two lines in any 2-plane. Now we let $\mathfrak{L}'$ be the projection of $\mathfrak{L}$ to a generic 3-plane. For a generic choice of the projection we see that $|\mathfrak{L}'| = |\mathfrak{L}|$, $|P_r(\mathfrak{L}')| = |P_r(\mathfrak{L})|$, and $\mathfrak{L}'$ contains at most two lines in any 2-plane. We then bound $|P_r(\mathfrak{L}')|$ using Theorem 4.5 from [5], giving the bound $|P_r(\mathfrak{L})| = |P_r(\mathfrak{L}')| \lesssim L^{3/2}r^{-2} + Lr^{-1}$. There is a large gap between the upper and lower bounds. The random projection argument does not seem to use much of the structure of Y: as Rudnev points out in [18], the space of lines in $\mathbb{R}^3$ is 4-dimensional while the space of lines in Y is only 3-dimensional.

We can ask the same question over the complex numbers. The example of [23] is still the best lower bound. For upper bounds, the random projection argument still works, but Theorem 4.5 from [5] is not known over the complex numbers. In the complex case, the best upper bound comes from applying Theorem 2 of [11], giving the bound $|P_r(\mathfrak{L})| \lesssim L^{3/2}r^{-3/2} + Lr^{-1}$.

In Question 4.6, the interesting case is when $r > 2$. The variety Y contains the subvariety $x_1 = x_3^2 - x_4^2$, $x_2 = 0$. It is not difficult to construct a set of L lines in this subvariety with $L^2/4$ 2-rich points by modifying the example at the start of Sect. 2.

But for cubic hypersurfaces, it looks hard to estimate the number of 2-rich points. For example, we can ask the following question.

Question 4.7 *Suppose that Y is the degree 3 hypersurface in $\mathbb{C}^4$ defined by the equation*

$$z_1^3 + z_2^3 + z_3^3 + z_4^3 = 1.$$

Suppose that $\mathfrak{L}$ is a set of L lines in Y. What is the maximum possible size of $P_2(\mathfrak{L})$?

(I believe that a generic point of this cubic hypersurface Y lies in six lines in Y. Here is a heuristic argument for this guess. Fix a point $p \in Y$ and translate the coordinate system so that $p = 0$. In the new coordinate system, Y is given as the zero set of a polynomial P of the form $P = P_3(z) + P_2(z) + P_1(z)$, where $P_i(z)$ is homogeneous of degree i. (There is no zeroth order term because we have arranged that $0 \in Z(P)$ and so $P(0) = 0$.) For a non-zero z, the line from 0 through z lies in $Y = Z(P)$ if and only if $P_3(z) = P_2(z) = P_1(z) = 0$. So the set of lines in Y through p is given by intersecting a degree 3 hypersurface, a degree 2 hypersurface, and a degree 1 hypersurface in $\mathbb{CP}^3$. For a generic choice of these hypersurfaces, the intersection will consist of six elements of $\mathbb{CP}^3$, and I believe that this occurs at a generic point of Y.)

Note that it does matter which cubic hypersurface we consider. The cubic hypersurface $z_4 = z_1 z_2 z_3$ contains a 2-dimensional degree 2 surface defined by $z_3 = 1$, $z_4 = z_1 z_2$, and this surface contains L lines with $L^2/4$ 2-rich points, as in the example at the start of Sect. 2. I believe that the cubic hypersurface $z_1^3 + z_2^3 + z_3^3 + z_4^3 = 1$ does not contain any 2-dimensional variety of degree 2. If this is the case, then we can get a non-trivial upper bound by a random projection argument. By a version of the Bezout theorem, the intersection of Y with any degree 2 2-dimensional variety will contain at most 6 lines. Randomly projecting $\mathfrak{L}$ to $\mathbb{C}^3$, we get a set of lines $\mathfrak{L}'$ with at most 6 lines of $\mathfrak{L}'$ in any 2-plane or degree 2 surface. Then applying Theorem 3.1, we see that $|P_2(\mathfrak{L})| = |P_2(\mathfrak{L}')| \lesssim L^{3/2}$. But I suspect that the maximum size of $|P_2(\mathfrak{L})|$ is much smaller than $L^{3/2}$.

I think that these questions about lines in low degree varieties are a natural direction of research in incidence geometry. If there are more exotic doubly-ruled varieties Y, then it would also be natural to study analogous questions for them.

References

1. J. Bourgain, N. Katz, T. Tao, A sum-product estimate in finite fields, and applications. Geom. Funct. Anal. **14**(1), 27–57 (2004)
2. Gy. Elekes, M. Sharir, Incidences in three dimensions and distinct distances in the plane, in *Proceedings 26th ACM Symposium on Computational Geometry* (2010), pp. 413–422
3. J. Ellenberg, M. Hablicsek, An incidence conjecture of Bourgain over fields of positive characteristic. Forum Math. Sigma **4**, e23, 9pp (2016). arXiv:1311.1479

4. L. Guth, *Polynomial Methods in Combinatorics*. University Lecture Series, vol. 64 (AMS, 2016)
5. L. Guth, N. Katz, On the Erdős distinct distance problem in the plane. Ann. Math. **181**, 155–190 (2015)
6. L. Guth, J. Zahl, Algebraic curves, rich points, and doubly-ruled surfaces. arXiv:1503.02173
7. L. Guth, J. Zahl, Curves in $\mathbb{R}^4$ and 2-rich points. arXiv:1512.05648
8. J. Harris, *Algebraic Geometry, a First Course*. Corrected reprint of the 1992 original. Graduate Texts in Mathematics, vol. 133 (Springer, New York, 1995)
9. H. Kaplan, J. Matoušek, M. Sharir, Simple proofs of classical theorems in discrete geometry via the Guth-Katz polynomial partitioning technique. Discrete Comput. Geom. **48**(3), 499–517 (2012)
10. N. Katz, The flecnode polynomial: a central object in incidence geometry, in *Proceedings of the 2014 ICM*. arXiv:1404.3412
11. J. Kollár, Szemerédi-Trotter-type theorems in dimension 3. Adv. Math. **271**, 30–61 (2015)
12. J.M. Landsberg, Is a linear space contained in a submanifold – on the number of the derivatives needed to tell. Journal für die reine und angewandte Mathematik **508**, 53–60 (1999)
13. J. Matoušek, *Using the Borsuk-Ulam Theorem* (Springer). Universitext, 2nd printing 2008
14. E. Mezzetti, D. Portelli, On threefolds covered by lines. Abh. Math. Sem. Univ. Hamburg **70**, 211–238 (2000)
15. J. Plucker, Neue Geometrie des Raumes gegründet auf die Betrachtung der geraden Linie als Raumelemente (Leipzig 1869)
16. H. Pottmann, J. Wallner, *Computational Line Geometry* (Springer, Heidelberg/Berlin, 2001)
17. O. Roche-Netwon, M. Rudnev, I. Shkredov, New sum-product type estimates over finite fields. Adv. Math. **293**, 589–605 (2016). arXiv:1408.0542
18. M. Rudnev, On the number of incidences between points and planes in three dimensions. arXiv:1407.0426
19. M. Rudnev, J. Selig, On the use of Klein quadric for geometric incidence problems in two dimensions. SIAM J. Discrete Math. **30**(2), 934–954 (2016). arxiv:1412.2909
20. G. Salmon, *A Treatise on the Analytic Geometry of Three Dimensions*, vol. 2, 5th edn. (Hodges, Figgis And Co. Ltd., 1915)
21. M. Sharir, A. Sheffer, N. Solomon, Incidences with curves in $\mathbb{R}^d$, in *Algorithms–ESA 2015*. Lecture Notes in Computer Science, vol. 9294 (Springer, Heidelberg, 2015), pp. 977–988. arXiv:1512.08267
22. M. Sharir, N. Solomon, Incidences between points and lines in $\mathbb{R}^4$. arXiv:1411.0777
23. N. Solomon, R. Zhang, Highly incidental patterns on a quadratic hypersurface in $\mathbb{R}^4$. arXiv:1601.01817
24. J. Solymosi, T. Tao, An incidence theorem in higher dimensions. Discrete Comput. Geom. **48**(2), 255–280 (2012)
25. The Szemerédi-Trotter theorem in the complex plane. Combinatorica **35**(1), 95–126 (2015). aXiv:math/0305283, 2003
26. J. Zahl, A Szemerédi-Trotter type theorem in $\mathbb{R}^4$. Discrete Comput. Geom. **54**(3), 513–572 (2015)
27. R. Zhang, Polynomials with dense zero sets and discrete models of the Kakeya conjecture and the Furstenberg set problem. arXiv:1403.1352

Approximating the k-Level in Three-Dimensional Plane Arrangements

Sariel Har-Peled, Haim Kaplan, and Micha Sharir

Abstract Let H be a set of n non-vertical planes in three dimensions, and let $r < n$ be a parameter. We give a construction that approximates the (n/r)-level of the arrangement $\mathcal{A}(H)$ of H by a terrain consisting of $O(r/\varepsilon^3)$ triangular faces, which lies entirely between the levels n/r and $(1 + \varepsilon)n/r$. The proof does not use sampling, and exploits techniques based on planar separators and various structural properties of levels in three-dimensional arrangements and of planar maps. This leads to conceptually cleaner constructions of shallow cuttings in three dimensions.

On the way, we get two other results that are of independent interest: (a) We revisit an old result of Bambah and Rogers (J Lond Math Soc 1(3):304–314, 1952) about triangulating a union of convex pseudo-disks, and provide an alternative proof that yields an efficient algorithmic implementation. (b) We provide a new construction of cuttings in two dimensions.

A preliminary version of this paper appeared in *Proceedings of the 27th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2016, 1193–1212 [31]. Work by Sariel Har-Peled was partially supported by NSF AF awards CCF-1421231 and CCF-1217462. Work by Haim Kaplan was partially supported by grant 1161/2011 from the German-Israeli Science Foundation, by grant 1841/14 from the Israel Science Foundation, and by the Israeli Centers for Research Excellence (I-CORE) program (center no. 4/11). Work by Micha Sharir has been supported by Grant 2012/229 from the U.S.-Israel Binational Science Foundation, by Grant 892/13 from the Israel Science Foundation, by the Israeli Centers for Research Excellence (I-CORE) program (center no. 4/11), by the Blavatnik Research Fund at Tel Aviv University, and by the Hermann Minkowski–MINERVA Center for Geometry at Tel Aviv University.

S. Har-Peled
Department of Computer Science, University of Illinois, 201 N. Goodwin Avenue,
61801 Urbana, IL, USA
e-mail: sariel@illinois.edu

H. Kaplan (✉) • M. Sharir
School of Computer Science, Tel Aviv University, 69978 Tel Aviv, Israel
e-mail: haimk@post.tau.ac.il; michas@post.tau.ac.il

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_19

1 Introduction

A tribute to Jirka Matoušek We were very fortunate to have Jirka as a friend and colleague. He has entered our community in the late 1980s, and has been a giant lighthouse ever since, showing us the way into new discoveries, solving mysteries for us, and providing us with new tools, ideas, and techniques, that have made our work much more interesting and productive. He has been everywhere, making seminal contributions to so many topics in computational and discrete geometry (and to other fields too). We have been avid readers of his many books, most notably *Lectures on Discrete Geometry*, and have been admiring his clear yet precise style of exposition and presentation. We have also learned to appreciate his personality, his dry but touching sense of humor, his love for nature, his infinite devotion to science on one hand, and to his family and friends on the other hand. His departure has been painful to us, and we will miss him badly. We thank you, Jirka, for all the gifts you gave us, and may your soul be blessed.

This paper is about a topic that Jirka has worked on, rather extensively, during the early 1990s, concerning *cuttings* and related techniques for decompositions of arrangements or of point sets, and their applications to range searching and other algorithmic and combinatorial problems in geometry. In particular, in 1992 he has written a seminal paper on "Reporting points in halfspaces" [41], where he introduced and analyzed *shallow cuttings*, a technique that had many applications during the following decades.

In a later paper, following his earlier work [38] (probably his first entry into computational geometry), Jirka [42] presented a construction of $(1/r)$-cuttings, for a set of lines in the plane, with $\le 8r^2 + 6r + 4$ cells. This construction uses, as a basic building block, a strikingly simple procedure for approximating a level in a line arrangement: Since a specific level is an x-monotone polygonal chain, one can pick every qth vertex, for $q \approx n/r$, and connect these vertices consecutively to form an approximate level, which is at line-crossing distance at most $q/2$ from the original level. As is well known, this construction is asymptotically optimal for any arrangement of lines in general position. This elegant level approximation algorithm, in two dimensions, raises the natural question of whether one can approximate a level in three dimensions for a given set of planes, by an xy-monotone polyhedral terrain constructed directly, in an analogous manner, from the original level.

This paper provides an affirmative answer to this question, thereby pushing Jirka's work further, for the special case of three-dimensional arrangements of planes. Our new scheme for approximating a level by a terrain, while significantly more involved than Jirka's two-dimensional construction, still echoes and generalizes his basic idea of "shortcutting" the original level by a coarser triangular mesh (instead of a simplified polygonal chain in the plane) spanned by selected vertices of the level.

Cuttings Let H be a set of n (non-vertical) hyperplanes in $\mathbb{R}^d$, and let $r < n$ be a parameter. A $(1/r)$-*cutting* of the arrangement $\mathcal{A}(H)$ is a collection of pairwise openly disjoint simplices (or other regions of constant complexity) such that the closure of their union covers $\mathbb{R}^d$, and each region is crossed (intersected in its interior) by at most n/r hyperplanes of H.

Cuttings have proved to be a powerful tool for a variety of problems in discrete and computational geometry, because they provide an effective divide-and-conquer mechanism for tackling such problems; see Agarwal [7] for an early survey. Applications include a variety of range searching techniques [10], partition trees [39], incidence problems involving points and lines, curves, and surfaces [26], and many more.

The first (albeit suboptimal) construction of cuttings is due to Clarkson [24]. This concept was formalized later on by Chazelle and Friedman [22], who gave a sampling-based construction of optimal-size cuttings (see below). An optimal deterministic construction algorithm was provided by Chazelle [20]. Matoušek [42] studied the number of cells in a $(1/r)$-cutting in the plane (see also [29]). See Agarwal and Erickson [10] and Chazelle [21] for comprehensive reviews of this topic.

To be effective, it is imperative that the number of simplices in the cutting be asymptotically as small as possible. Chazelle and Friedman [22] were the first to show the existence of a $(1/r)$-cutting of the entire arrangement of n hyperplanes in $\mathbb{R}^d$, consisting of $O(r^d)$ simplices, which is asymptotically the best possible bound. (We note in passing that cuttings of optimal size are not known for arrangements of (say, constant-degree algebraic) surfaces in $\mathbb{R}^d$, except for $d = 2$, where the known bound, $O(r^2)$, is tight, and for $d = 3, 4$, where nearly tight bounds, i.e., nearly cubic and quartic in r, respectively, are known [23, 35, 36].)

For additional works related to cuttings and their applications, see [1–6, 8, 13, 19, 29, 39, 40, 48].

Shallow cuttings The *level* of a point p in the arrangement $\mathcal{A}(H)$ of H is the number of hyperplanes lying vertically below it (that is, in the $(-x_d)$-direction). For a given parameter $0 \le k \le n-1$, the *k-level*, denoted as L_k, is the closure of all the points that lie on some hyperplane of H and are at level exactly k, and the $(\le k)$-level, denoted as $L_{\le k}$, is the union of all the j-levels, for $j = 0, \dots, k$. A collection of pairwise openly disjoint simplices such that the closure of their union covers $L_{\le k}$, and such that each simplex is crossed by at most n/r hyperplanes of H, is a *k-shallow* $(1/r)$*-cutting*. Naturally, the parameters k and r can vary independently, but the interesting case, which is the one that often arises in many applications, is the case where $k = \Theta(n/r)$. Furthermore, shallow cuttings for any value of k can be reduced to this case—see Chan and Tsakalidis [19, Section 5].

In his paper on reporting points in halfspaces [41], Matoušek has proved the existence of small-size shallow cuttings in arrangements of hyperplanes in any dimension, showing that the bound on the size of the cutting can be significantly improved for shallow cuttings. Specifically, he has shown the existence of a k-shallow $(1/r)$-cutting, for n hyperplanes in $\mathbb{R}^d$, whose size is $O\left(q^{\lceil d/2\rceil} r^{\lfloor d/2\rfloor}\right)$, where $q = k(r/n) + 1$. For the interesting special case where $k = \Theta(n/r)$, we have $q = 1$ and the size of the cutting is $O\left(r^{\lfloor d/2\rfloor}\right)$, a significant improvement over the general bound $O(r^d)$. (For example, in three dimensions, we get $O(r)$ simplices, instead of $O(r^3)$ simplices for the whole arrangement.) This has lead to improved solutions of many range searching and related problems.

In his paper, Matoušek presented a deterministic algorithm that can construct such a shallow cutting in polynomial time; the running time improves to $O(n \log r)$ but only when r is small, i.e., $r < n^\delta$ for a sufficiently small constant δ (that depends on the dimension d). Later, Ramos [48] presented a (rather complicated) randomized algorithm for $d = 2, 3$, that constructs a hierarchy of shallow cuttings for a geometric sequence of $O(\log n)$ values of r, where for each r the corresponding cutting is a $(1/r)$-cutting of the first $\Theta(n/r)$ levels of $\mathcal{A}(H)$. Ramos's algorithm runs in $O(n \log n)$ total expected time. Recently, Chan and Tsakalidis [19] provided a deterministic $O(n \log r)$-time algorithm for computing an $O(n/r)$-shallow $(1/r)$-cutting. Their algorithm can also construct a hierarchy of shallow cuttings for a geometric sequence of $O(\log n)$ values of r, as above, in $O(n \log n)$ time (deterministically). Interestingly, they use Matoušek's theorem on the existence of an $O(n/r)$-shallow $(1/r)$-cutting of size $O(r)$ in the analysis of their algorithm.

Each simplex Δ in the cutting has a *conflict list* associated with it, which is the set of hyperplanes intersecting Δ. The algorithms mentioned above for computing cuttings also compute the conflict lists associated with the simplices of the cutting. Alternatively, given the cutting, one can produce the conflict lists in $O(n \log r)$ time using a result of Chan [16], as we outline in Sect. 4.3.

Matoušek's proof of the existence of small-size shallow cuttings, as well as subsequent studies of this technique, are technically involved. They rely on random sampling, combined with a clever variant of the so-called exponential decay lemma of [22], and with several additional (and rather intricate) techniques.

Approximating a level An early study of Matoušek [38] gives a construction of a $(1/r)$-cutting of small (optimal) size in arrangements of lines in the plane. The construction chooses a sequence of r levels, n/r apart from one another, and approximates each of them by a coarser polygonal line, by choosing every $n/(2r)$-th vertex of the level, and by connecting them by an x-monotone polygonal path. Each approximate level does not deviate much from its original level, so they remain disjoint from one another. Then, partitioning the region between every pair of consecutive approximate levels into vertical trapezoids produces a total of $O(r^2)$ such trapezoids, each crossed by at most $O(n/r)$ lines.

It is thus natural to ask whether one can approximate, in a similar fashion, a k-level of an arrangement of a set H of n planes in 3-space. This is significantly more challenging, as the k-level is now a polyhedral terrain, and while it is reasonably easy to find a good (suitably small) set of vertices that "represent" this level (in an appropriate sense, detailed below), it is less clear how to triangulate them effectively to form an xy-monotone terrain, such that (i) none of its triangles is crossed by too many planes of H, and (ii) it remains close to the original level. To be more precise, given k and $\varepsilon > 0$, we want to find a polyhedral terrain with a small number of faces, which lies entirely between the levels k and $(1 + \varepsilon)k$ of $\mathcal{A}(H)$. A simple tweaking of Matoušek's technique produces such an approximation in the planar case, but it is considerably more involved to do it in 3-space.

Algorithms for terrain approximation, such as in [9, 12], do not apply immediately in this case, as they produce a suboptimal output, of size larger than the optimal by a logarithmic factor. More importantly, they are not geared to handle our measure of approximation (in terms of lying close to a specified level, in the sense that no point on the approximation is separated by too many planes from the level). Nevertheless, we note that the algorithm in [12] can be modified to provide a logarithmic approximation in our sense, but the resulting running time (at least $\Omega(n^8)$, and probably much worse) is quite large. Perhaps more significantly, without the results in the present work, it is not even clear that such an optimal-size approximation exists at all.

Such an approximation to the k-level, whose size is almost optimal up to a polylogarithmic factor, can be obtained by using a *relative-approximation* sample of H, and by extracting the appropriate level in the sample [30]. A more natural approach, of using the triangular faces of an optimal-size shallow cutting to form an approximate k-level, seems to fail in this case, as the shallow cutting is in general just a collection of simplices, stacked on top of one another, with no clearly defined xy-monotonicity. Such a monotonicity is obtained in Chan [17], by replacing a standard shallow cutting by the upper convex hull of its simplices. However, the resulting cuttings do not lead to a sharp approximation of the level, of the sort we seek.

In short, an effective and optimal technique for approximating a level in three dimensions as a terrain (let alone in higher dimensions) does not follow easily from existing techniques.

An additional advantage of such an approximation is that it immediately yields a simply-shaped shallow cutting of the first k levels of $\mathcal{A}(H)$, by replacing each triangle Δ of the approximate level by the vertical semi-unbounded triangular prism Δ^* having Δ as its top face, and consisting of all points that lie on or vertically below Δ. Such a cutting (by prisms) has already been constructed by Chan [17], but it does not yield (that is, come from) a $(1 + \varepsilon)$-approximation to the level. Such a shallow cutting, by vertical semi-unbounded triangular prisms, was a central tool in Chan's algorithm for dynamic convex hulls in three dimensions [18].

This discussion suggests that resolving the question of approximating the k-level by an xy-monotone terrain of small, optimal size is not a mere technical issue, but rather a tool that will shed more light on the geometry of arrangements of planes in three dimensions.

1.1 Our Results

In this paper we give an alternative and constructive proof of the existence of optimal-size shallow cuttings in a three-dimensional plane arrangement, by vertical semi-unbounded triangular prisms. We obtain this cutting in a straightforward manner from an optimal-size approximate level, as discussed above. Specifically, we show that given r and ε, one can approximate the (n/r)-level in an arrangement of n non-vertical planes in $\mathbb{R}^3$, by a polyhedral terrain with $O(r/\varepsilon^3)$ triangular faces, that lies entirely between the levels n/r and $(1+\varepsilon)n/r$.

The construction does not use sampling, nor does it use the exponential decay lemma of [22, 41]. It is based on the planar separator theorem of Lipton and Tarjan [37], or, more precisely, on recent separator-based decomposition techniques of planar maps, as in Klein et al. [34] (see also Frederickson [28]), and on several insights into the structure and properties of levels in three dimensions and of planar maps, which we believe to be of independent interest.

Sketch of our technique The k-level in a plane arrangement in three dimensions is an xy-monotone polyhedral terrain. After triangulating each of its faces, its xy-projection forms a (straight-edge) triangulated biconnected planar map. Since the overall complexity of the first k levels is $O(nk^2)$ (see, e.g., [25]), we may assume, by moving from a specified level to a nearby one, that the complexity of our level is near the average value $O(nk)$. The decomposition techniques of planar graphs mentioned above (as in [28]) allow us to partition the level into $O(n/k)$ clusters, where each cluster has $O(k^2)$ vertices and $O(k)$ boundary vertices (vertices that also belong to other clusters). In the terminology of [28], this is a k^2*-division* of the graph. Each such cluster, projected to the xy-plane, is a polygon with $O(k)$ boundary edges (and with $O(k^2)$ interior projected edges of the original level). We show that replacing each such projected polygon by its convex hull results in a collection of $O(n/k)$ convex *pseudo-disks*, namely, each hull is (trivially) simply connected, and the boundaries of any pair of hulls cross at most twice. Moreover, the decomposition has the property that, for each triangle Δ that is fully contained in such a pseudo-disk, lifting its vertices back to the k-level yields a triple of points that span a triangle Δ' with a small number of planes crossing it, so it lies close to the k-level.

An old result of Bambah and Rogers [15], proving a statement due to L. Fejes-Tóth, and reviewed in [47, Lemma 3.9] (and also briefly below), shows that a union of m convex pseudo-disks that covers the plane induces a triangulation of the plane by $O(m)$ triangles, such that each triangle is fully contained inside one of the pseudo-disks. (As a matter of fact, it shows that each pseudo-disk can be shrunk into a convex polygon so that these polygons are pairwise openly disjoint, with the same union, and the total number of edges of the polygons is at most $6m$; the desired triangulation is obtained by simply triangulating, arbitrarily, each of these polygons.) Lifting (the vertices of) this triangulation to the k-level, with a corresponding lifting of its triangular faces, results in the desired terrain approximating the level. A shallow cutting of the first k levels is obtained by simply replacing each triangle Δ of the approximate level by the semi-unbounded vertical prism of points lying below Δ.

Planar cuttings Interestingly, a simplified version of the algorithm for approximating the k-level in 3-space can also be applied to arrangement of lines in the plane, yielding a new construction of cuttings in the plane, which is different from previous approaches. We present this warm-up exercise in Sect. 3, and believe it to be of independent interest.

Confined triangulations One of the main contributions of this work is providing an alternative proof of the aforementioned result of Bambah and Rogers [15]. The original proof in [15], and its simplified presentation in [47], do not seem to lead to a sufficiently efficient construction. In contrast, the new proof leads to an algorithm with near linear running time that constructs a triangulation with the desired properties; see Sect. 2.

The idea of decomposing the union of objects (pseudo-disks here) into pairwise openly disjoint simply-shaped fragments, each fully contained in some original object, is implicit in algorithms for efficiently computing the union of objects; see the work of Ezra et al. [27], which was in turn inspired by Mulmuley's work on hidden surface removal [45]. Mustafa et al. [46] use a more elaborate version of such a decomposition, for situations where the objects are weighted. While these decompositions are useful for a variety of applications, they still suffer from the problem that the complexity of a single region in the decomposition might be arbitrarily large. In contrast, the triangulation scheme that we use (following [15]) is simpler, optimal, and independent of the complexity of the relevant pseudo-disks. We are pleased that this nice property of convex pseudo-disks is (effectively) applicable to the problems studied here, and expect it to have many additional potential applications.

Our analysis extends to a collection of convex pseudo-disks whose union does not cover the plane. In this setting our triangulation consists of triangles and caps (where a cap is the intersection of an input pseudo-disk with a halfplane). This provides a representation of "most" of the union by triangles, where the more complicated caps are only used to fill in the "fringe" of the union (and are absent when the union covers the entire plane, as in [15]). We believe that this triangulation could be useful in practice, in situations where, given a query point q, one wants to decide whether q is inside the union, and if so, provide a witness shape that contains q. For this, we simply locate the triangle or cap that contains q in our triangulation, from which the desired witness shape is immediately available.

We also extend our analysis, and show that such a decomposition exists for arbitrary convex shapes, with the number of pieces being proportional to the union complexity.

Additional applications Two additional applications of our construction, that are described in the arXiv version of this paper [32], are the following:

(a) We extend Matoušek's construction [38] of cuttings in planar arrangements to three dimensions. That is, we construct a "layered" $(1/r)$-cutting of the entire arrangement $\mathcal{A}(H)$ of a set H of n non-vertical planes in $\mathbb{R}^3$, of optimal size $O(r^3)$, by approximating each level in a suitable sequence of levels, and then by triangulating each layer between consecutive levels in the sequence.

(b) We present yet another construction of cuttings in two-dimensional line arrangements that is based on a packing argument combined with the new techniques of this paper.

Paper organization We start by presenting the construction of the confined triangulation in Sect. 2. As a warm-up exercise, we use this result in Sect. 3 to present a new algorithm for constructing cuttings of arrangements of lines in the plane. We then describe the construction of approximate levels, and the construction of shallow cuttings that it leads to, in Sect. 4.

2 Triangulating the Union of Convex Pseudo-disks and Other Shapes

In this section we show that, given a finite collection of m convex pseudo-disks covering the plane, one can construct a triangulation of the plane, consisting of $O(m)$ triangles, such that each triangle is contained in a single original pseudo-disk—see Theorem 2.4 below for details. Our result can be extended to situations where the union of the pseudo-disks is not the entire plane; see below. This claim is a key ingredient in our construction of approximate k-levels, detailed in Sect. 4, but it is not new, as it is an immediate consequence of an old result of Bambah and Rogers [15] (proving a statement by L. Fejes-Tóth), whose proof is sketched below. Our analysis provides an alternative constructive proof.

We use this result (i.e., Theorem 2.4) as a black box later on in the paper, and the impatient reader might want to skip this (somewhat tedious) section for later reading, and go directly to Sect. 3.

Bambah and Rogers' proof For the sake of completeness, we briefly sketch the proof of Bambah and Rogers (as presented in Pach and Agarwal [47, Lemma 3.9]). Let $\mathcal{K}$ be a collection of m convex pseudo-disks in the plane, and assume, for simplicity, that their union is a triangle T (extending this simpler scenario to the more general case is straightforward). We may also assume that no pseudo-disk of $\mathcal{K}$ is contained in the union of the other regions of $\mathcal{K}$, as one can simply throw away any such redundant pseudo-disk. Finally, since the construction will create regions with overlapping boundaries, we use the more general definition of pseudo-disks, requiring, for each pair $C, D \in \mathcal{K}$, that $C \setminus D$ and $D \setminus C$ be both connected. See Fig. 1A.

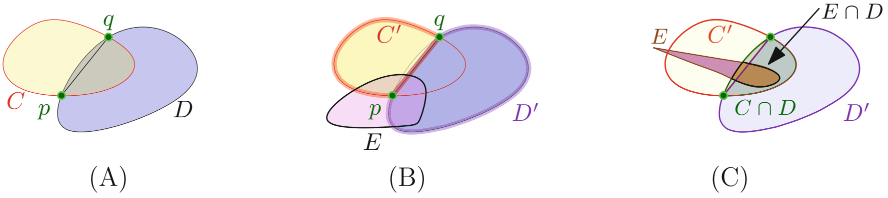


Fig. 1 The proof of Bambah and Rogers

Let C and D be two pseudo-disks of $\mathcal{K}$, such that the intersection $\mathrm{int}(C) \cap \mathrm{int}(D)$ of their interiors is nonempty and minimal in terms of containment (that is, it does not contain any other such intersection). Let p and q be the two intersection points of ∂C and ∂D (since $C \cap D$ has a nonempty interior, ∂C and ∂D cannot overlap, so p and q are well defined). Cut C and D along the segment pq, and let $C' \subseteq C$ and $D' \subseteq D$ be the two resulting pieces whose union is $C \cup D$, see Fig. 1B. Let $\mathcal{K}' = (\mathcal{K} \setminus \{C, D\}) \cup \{C', D'\}$. We claim that $\mathcal{K}'$ is a collection of m pseudo-disks covering T.

Indeed, consider a pseudo-disk $E \in \mathcal{K}'$ other than C', D'. We need to show that $E \setminus C'$ and $C' \setminus E$ are both connected, and similarly for E and D'. The pseudo-disk property is immediate if $E \cap pq$ is empty. If E contains exactly one endpoint, say p, of pq, then it must intersect $\partial C \cap D$ at exactly one point, which is eliminated in C' and is replaced by the single intersection point of ∂E with pq. Finally, assume that E does not contain p or q, but still intersects the segment pq at two points. If $E \subset C$ then the pseudo-disk property for E and C' is obvious, so assume that ∂E intersects ∂C at two points. These points must either both lie in $\partial C \cap D$ or both lie in $\partial C \setminus D$. In the former case, ∂E crosses $\partial C'$ only at the two points on pq. In the latter case, $E \cap C \subset C \cap D$, contradicting the minimality of $C \cap D$; see Fig. 1C.

We thus replace $\mathcal{K}$ by $\mathcal{K}'$, and repeat this process till all the pseudo-disks in the resulting collection are pairwise interior disjoint. At this point, $\mathcal{K}$ is a pairwise openly disjoint cover of the triangle T, by m convex polygons (each contained inside its original pseudo-disk). By Euler's formula, these polygons have a total of $O(m)$ edges, and can thus be triangulated into $O(m)$ triangles with the desired property.

This elegant proof is significantly simpler than what follows, but it does *not* seem to lead to an efficient algorithm for constructing the desired triangulation in near-linear running time. We present here a different alternative (efficiently) constructive proof, which leads to an $O(m \log m)$-time algorithm for constructing the triangulation for a set of m pseudo-disks, in a suitable model of computation. (As an aside, we also think that such a nice property deserves more than one proof.) We also establish extensions of this result to the case where the union of the pseudo-disks does not cover the plane, and for more general convex shapes, not necessarily pseudo-disks.

2.1 Preliminaries

The notion of a triangulation that we use here is slightly non-standard, as it might be a triangulation of the entire plane, and not just of the convex hull of some input set of points. As such, it contains unbounded triangles, where the boundary of each such triangle consists of one bounded segment and two unbounded rays (where the segment might degenerate into a single point, in which case the triangle becomes a wedge).

Given a convex shape D, a ***cap*** of D is the region formed by the intersection of D with a halfplane. A ***crescent*** is a portion of a cap obtained by removing from

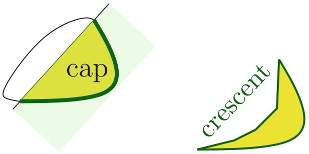


Fig. 2 A cap and a crescent

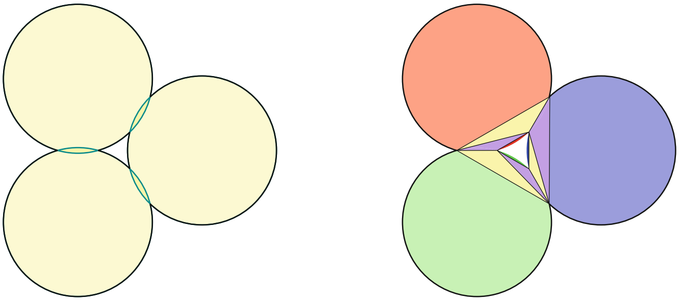


Fig. 3 A union of three disks, and its decomposition into triangles and caps. Note that the decomposition computed by our algorithm is somewhat different for this case

it a convex polygon that has the base chord of the cap as an edge, but is otherwise contained in the interior of the cap. See Fig. 2.

Definition 2.1 Given a collection $\mathcal{D}$ of convex shapes in the plane, a decomposition $\mathcal{T}$ of their union into pairwise openly disjoint regions is a ***confined triangulation*** if (i) every region in $\mathcal{T}$ is either a triangle or a cap, and (ii) every such region is fully contained in one of the original input shapes. See Fig. 3.

2.2 Construction

We are given a collection $\mathcal{D}$ of m convex pseudo-disks, and our goal is to construct a confined triangulation for $\mathcal{D}$, as described above, with $O(m)$ pieces. In what follows we consider both the case where the union of $\mathcal{D}$ covers the plane, and the case where it does not.

2.2.1 Painting the Union from Front to Back

A basic property of a collection $\mathcal{D}$ of m pseudo-disks is that the combinatorial complexity of the boundary of the union $\mathcal{U} := \mathcal{U}(\mathcal{D}) = \bigcup_{C \in \mathcal{D}} C$ of $\mathcal{D}$ is at most $6m-12$, where we ignore the complexity of individual members of $\mathcal{D}$, and just count the number of intersection points of pairs of boundaries of members of $\mathcal{D}$ that lie on

$\partial\mathcal{U}$; see [33]. For convenience, we also (i) include the leftmost and rightmost points of each $D \in \mathcal{D}$ in the set of intersection points (if they lie on the union boundary), thus increasing the complexity of the union by at most $2m$, and (ii) assume general position of the pseudo-disks.

An intersection point v of a pair of boundaries is at *depth k* (of the arrangement $\mathcal{A}(\mathcal{D})$ of $\mathcal{D}$) if it is contained in the interiors of exactly k members of $\mathcal{D}$. The boundary intersection points are thus at depth 0, and a simple application of the Clarkson–Shor technique [25] implies that the number of boundary intersection points that lie at depth 1 is also $O(m)$. Hence there exists at least one pseudo-disk $D \in \mathcal{D}$ that contains at most c intersection points at depths 0 or 1 (including leftmost and rightmost points of disks), for some suitable absolute constant c. Clearly, these considerations also apply to any subset of $\mathcal{D}$.

This allows us to order the members of $\mathcal{D}$ as $D_1, \ldots, D_m$, so that the following property holds. Set $\mathcal{D}_i := \{D_1, \ldots, D_i\}$, for $i = 1, \ldots, m$. Then D_i contains at most c intersection points at depths 0 and 1 of $\mathcal{A}(\mathcal{D}_i)$. Equivalently, for each i, the boundary of $D_i^0 := D_i \setminus \mathcal{U}(\mathcal{D}_{i-1})$ contains at most c intersection points.

To prepare for the algorithmic implementation of the construction in this proof, which will be presented later, we note that this ordering is not easy to obtain efficiently in a deterministic manner. Nevertheless, a random insertion order (almost) satisfies the above property: As we will show, the expected sum of the complexities of the regions D_i^0, for a random insertion order, is $O(m)$, which is the property that our analysis really needs. See later for more details.

We thus have $\mathcal{U}(\mathcal{D}_j) = \bigcup_{i \le j} D_i^0$ (as a pairwise openly disjoint union), for each j; for the convenience of presentation (and for the algorithm to follow), we interpret this ordering as an incremental process, where the pseudo-disks of $\mathcal{D}$ are inserted, one after the other, in the order $D_1, \ldots, D_m$, and we maintain the partial unions $\mathcal{U}(\mathcal{D}_j)$, after each insertion, by the formula $\mathcal{U}(\mathcal{D}_j) = \mathcal{U}(\mathcal{D}_{j-1}) \cup D_j^0$.

2.2.2 Decomposing the Union into Vertical Trapezoids

Since the boundary of $D_i^0 = D_i \setminus \mathcal{U}(\mathcal{D}_{i-1})$ contains at most c intersection points, we can decompose D_i^0 into $O(1)$ *vertical pseudo-trapezoids*, using the standard vertical decomposition technique; see, e.g., [50]. Let $\mathcal{T}_j$ be the collection of pseudo-trapezoids in the decomposition of $\mathcal{U}(\mathcal{D}_j)$, collected from the decompositions of the regions D_i^0, for $i = 1, \ldots, j$, and let V_j be the set of vertices of these pseudo-trapezoids, each of which is either an intersection point (more precisely, a boundary intersection or an x-extreme point) of $\mathcal{A}(\mathcal{D}_j)$, or an intersection between some $\partial\mathcal{D}_i$ and a vertical segment erected from an intersection point of $\mathcal{A}(\mathcal{D}_j)$.

Each of the pseudo-trapezoids in $\mathcal{T}_j$ is bounded by (at most) two vertical segments, a portion of the boundary of a single pseudo-disk as its top edge, and a portion of the boundary of (another) single pseudo-disk as its bottom edge; see the top parts of the subfigures in Fig. 4. We have $D_1^0 = D_1$, which we regard as a single pseudo-trapezoid, in which the vertical sides degenerate to the leftmost and rightmost points of ∂D_1; see Fig. 4(1). Note that in the vertical decomposition of D_i^0

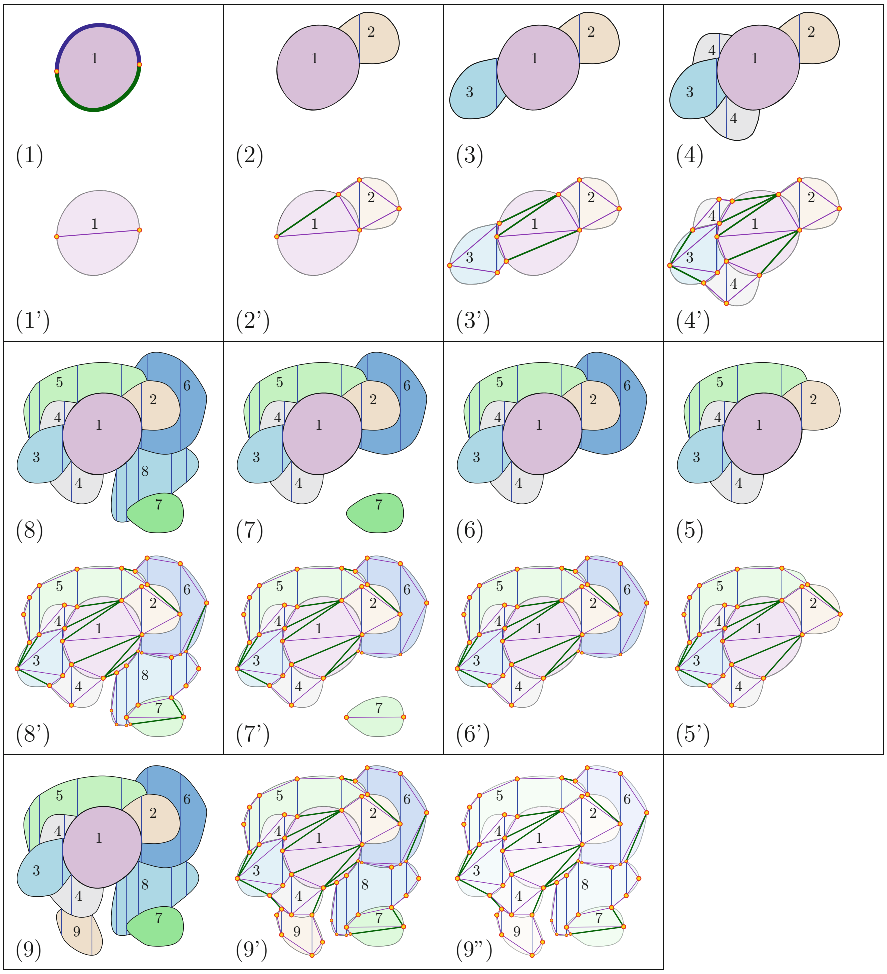


Fig. 4 A step-by-step illustration of the decomposition $\mathcal{T}$ into pseudo-trapezoids and of the polygonalization of the union. See Sect. 2.2.4. An animation of the steps depicted in this figure is available online at http://sarielhp.org/blog/?p=8920, see also the animated figure in the arXiv version of the paper [32]

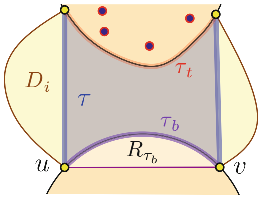
we split it by vertical segments through the intersection points on its boundary, but not through vertices of V_{i-1} on $\mathcal{U}(\mathcal{D}_{i-1})$ that are not intersection points of $\mathcal{A}(\mathcal{D})$. (Informally, these vertices are "internal" to $\mathcal{U}(\mathcal{D}_{i-1})$, and are not "visible" from the outside.) See, e.g., Fig. 4(4). The set V_i is obtained by adding to V_{i-1} the vertices of the pseudo-trapezoids in the decomposition of D_i^0.

If D_i^0 is bounded then each pseudo-trapezoid τ in its decomposition has a top boundary and a bottom boundary, but one or both of the vertical sides may be missing (see, e.g., Fig. 4(1) for the single pseudo-trapezoid $D_1^0 = D_1$ and Fig. 4(3) for the left pseudo-trapezoid of 3). From the point of view of τ, each of the top and bottom boundaries of τ may be either convex (if it is a subarc of ∂D_i on ∂D_i^0), or concave (if it is part of the boundary of some previously inserted pseudo-disk); If D_i^0 is not bounded then some of the vertical pseudo-trapezoids covering D_i^0 will also be unbounded and missing some of their boundaries. Note that D_i^0 is not necessarily connected; in case it is not connected we separately decompose each of its connected components into vertical pseudo-trapezoids in the above manner; see Fig. 4(4).

At the end of the incremental process, after inserting all the m pseudo-disks in $\mathcal{D}$, the pseudo-trapezoids in $\mathcal{T} := \mathcal{T}_m$ cover $\mathcal{U}(\mathcal{D})$, which may or may not be the entire plane, and they are pairwise openly disjoint. By construction, each pseudo-trapezoid in $\mathcal{T}$ is contained in a single pseudo-disk of $\mathcal{D}$. Moreover, since the complexity of each D_i^0 is $O(1)$, the total number of pseudo-trapezoids in $\mathcal{T}$ is $O(m)$. So $\mathcal{T}$ possesses some of the properties that we want, but it is not a triangulation.

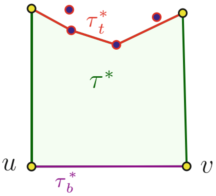
2.2.3 Polygonalizing the Pseudo-trapezoids

To get a triangulation, we associate a polygonal vertical pseudo-trapezoid τ^* with each pseudo-trapezoid $\tau \in \mathcal{T}$. We obtain τ^* from τ by replacing the bottom boundary τ_b and the top boundary τ_t of τ by respective polygonal chains τ_b^* and τ_t^*, that are defined as follows.[1] Let D_i be the pseudo-disk during whose insertion τ was created; in particular, $\tau \subseteq D_i^0$. Let u and v denote the endpoints of τ_b. Consider the region R_{τ_b} between τ_b and the straight segment uv; clearly, by the convexity of D_i, R_{τ_b} is fully contained in D_i. See figure on the below.

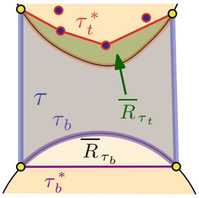


If R_{τ_b} contains no vertices of V_i, other than u and v (this will always be the case when $R_{\tau_b} \subseteq \tau$), we replace τ_b by $\tau_b^* = uv$. Otherwise, we replace τ_b by the chain τ_b^* of edges of the convex hull of $V_i \cap R_{\tau_b}$, other than the edge uv. We define τ_t^* analogously, and take τ^* to be the polygonal vertical pseudo-trapezoid that has the same vertical edges as τ, and its top (resp., bottom) part is τ_t^* (resp., τ_b^*). See figure on the below.

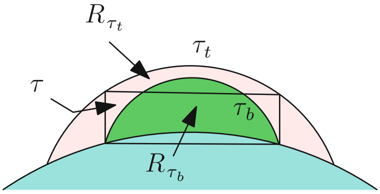
[1] The term "polygonal" is somewhat misleading, as some of the boundaries of the pseudo-disks of $\mathcal{D}$ may also be polygonal. To avoid confusion, think of the boundaries of the pseudo-disks of $\mathcal{D}$ as smooth convex arcs (as drawn in the figures) even though they might be polygonal.



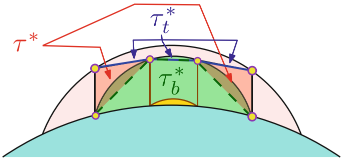
Note that, by construction, τ_b^* is a convex polygonal chain. From the point of view of τ, it is convex (resp., concave) if and only if τ_b is convex (resp., concave). (These statements become somewhat redundant when τ_b^* is the straight segnment uv.) An analogous property holds for τ_t^* and τ_t. We denote the crescent-like region bounded by τ_b and τ_b^* by $\overline{R}_{\tau_b}$; $\overline{R}_{\tau_t}$ is defined analogously. (Formally, $\overline{R}_{\tau_b} = R_{\tau_b} \setminus CH(V_i \cap R_{\tau_b})$ and $\overline{R}_{\tau_t} = R_{\tau_t} \setminus CH(V_i \cap R_{\tau_t})$.) Let $\mathcal{T}_i^*$ be the set of polygonal vertical pseudo-trapezoids associated in this manner with the pseudo-trapezoids in $\mathcal{T}_i$. See the figure below.



Note that R_{τ_b} and R_{τ_t} need not be disjoint, as illustrated in the figure on the below. Nevertheless, τ_b^* and τ_t^* cannot cross one another, as follows from Invariant (I2) that we establish below (in Lemma 2.2). This implies that τ^* is well defined. If τ_b^* and τ_t^* are not disjoint then they may only be pinched together at common vertices, or overlap in a single common connected portion (in the extreme case they may be identical).



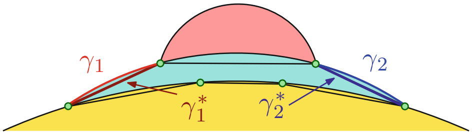
This pinching or overlap, if it occurs, causes the interior of τ^* to be disconnected (into at most two pieces, as depicted in the figure below; it may also be empty, as is the case for D_1^0, illustrated in Fig. 4(1)).



2.2.4 Filling the Cavities

The insertion of D_i may in general split some arcs of $\partial\mathcal{U}(\mathcal{D}_{i-1})$ into subarcs, whose new endpoints are either points of contact between ∂D_i and $\partial\mathcal{U}(\mathcal{D}_{i-1})$, or endpoints of vertical segments erected from other vertices of D_i^0. This can be seen all over Fig. 4. For example, see the subdivision of the top arc of D_7 caused by the insertion of D_8 in Fig. 4(8′). Some of these subarcs are boundaries of the new pseudo-trapezoids of D_i^0 and thus do not belong to $\partial\mathcal{U}(\mathcal{D}_i)$, and some remain subarcs of $\partial\mathcal{U}(\mathcal{D}_i)$. We refer to subarcs of the former kind as *hidden*, and to those of the latter kind as *exposed*. Note that, among the subarcs into which an arc of $\partial\mathcal{U}(\mathcal{D}_{i-1})$ is split, only the leftmost and rightmost extreme subarcs can be exposed (this follows from the pseudo-disk property of the objects of $\mathcal{D}$).

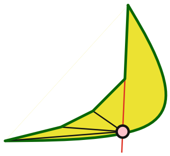
We take each new exposed arc γ, with endpoints u, v, and apply to it the same polygonalization that we applied above to τ_b and τ_t. That is, we take the region R_γ enclosed between γ and the segment uv, and define γ^* to be either uv, if R_γ does not contain any vertex of V_i, or else the boundary of $\mathrm{CH}\left(R_\gamma \cap V_i\right)$, except for uv. We note that γ^* is a convex polygonal chain that shares its endpoints with γ, and denote the region enclosed between γ and γ^* as $\overline{R}_\gamma$.



Let E_i denote the collection of all straight edges in the polygonal boundaries of the pseudo-trapezoids in $\mathcal{T}_i^*$ and in the polygonal chains γ^* corresponding to new exposed subarcs γ of $\partial\mathcal{U}(\mathcal{D}_{j-1})$, $1 \le j \le i$, which were created and polygonalized when adding the corresponding pseudo-disk D_j. See figure above.

2.2.5 Putting It All Together

When the pseudo-disks cover the plane When the polygonalization process terminates, there are no more regions $\overline{R}_\gamma$, for boundary arcs γ of the union (because there is no boundary), so we are left with a straight-edge planar map M with E_m as its set of edges. (Invariant (I1) in Lemma 2.2 below asserts that the edges in E_m do not cross each other.) By Euler's formula, the complexity of M is $O(m)$. We then triangulate each face of M, and, as the analysis in the next subsection will show, obtain the desired triangulation.



The general case In general, the construction decomposes the union into (pairwise openly disjoint) triangles and crescent regions. To complete the construction, we decompose each crescent region into triangles and caps. A crescent region with $t \geq 2$ vertices on its concave boundary can be decomposed into $t-2$ triangles and at most $t-1$ caps. The case $t = 2$ is vacuous, as the crescent is then a cap, so assume that $t \geq 3$. To get such a decomposition, take an extreme edge of the concave polygonal chain, and extend it till it intersects the convex boundary of the crescent, at some point w, thereby chopping off a cap from the crescent. We then create the triangles that w spans with all the concave edges that it sees, and then recurse on the remaining crescent; see the figure above. It is easily seen that this results in $t-2$ triangles and at most $t-1$ caps, as claimed. After this fix-up, we get a decomposition of the union into triangles and caps. Here too, by Euler's formula, the complexity of M is $O(m)$.

2.3 Analysis

The correctness of the construction is established in the following lemma.

Lemma 2.2 *The pseudo-trapezoids in $\mathcal{T}_i^*$ and the edges of E_i satisfy the following invariants:*

(I1) *The segments in E_i do not cross one another.*

(I2) *Each subarc γ of $\partial\mathcal{U}(\mathcal{D}_i)$ with endpoints u and v has an associated convex polygonal arc $\gamma^* \subseteq E_i$ between u and v. The chains γ^* are pairwise openly disjoint, and their union forms the boundary of a polygonal region $\mathcal{U}_i^* \subseteq \mathcal{U}(\mathcal{D}_i)$.*

(I3) *The pseudo-trapezoids in $\mathcal{T}_i^*$ are pairwise openly disjoint, and each of them is fully contained in some pseudo-disk of $\mathcal{D}_i$.*

(I4) $\mathcal{U}(\mathcal{D}_i) \setminus \bigcup_{\tau^* \in \mathcal{T}_i^*} \tau^*$ *consists of a collection of pairwise openly disjoint* holes. *Each hole is a region between two x-monotone convex chains or between two x-monotone concave chains, with common endpoints, where either both chains are polygonal, or one is polygonal and the other is a portion of the boundary of a single pseudo-disk that lies on $\partial\mathcal{U}(\mathcal{D}_i)$. (Each of the latter holes is a crescent-like region of the form $\overline{R}_{\tau_b}$, $\overline{R}_{\tau_t}$, for some trapezoid τ, or $\overline{R}_\gamma$, for some exposed* *arc γ, as defined above.) The union of the holes of the latter kind (crescents) is $\mathcal{U}(\mathcal{D}_i) \setminus \mathcal{U}_i^*$. Each hole, of either kind, is fully contained in some pseudo-disk D_j, $j \leq i$.*

We refer to holes of the former (resp., latter) kind in (I4) of the lemma as *internal polygonal* holes (resp., *external half-polygonal* holes).

Proof We prove that these invariants hold by induction on i. The invariants clearly hold for $\mathcal{T}_1^*$ and E_1 after starting the process with $D_1^0 = D_1$. Concretely, $\mathcal{T}_1^*$ consists of the single degenerate pseudo-trapezoid uv, where u and v are the leftmost and rightmost points of $\mathcal{D}_1$, respectively, and $E_1 = \{uv\}$. The (external half-polygonal) holes are the portions of D_1 lying above and below uv. It is obvious that (I1)–(I4) hold in this case.

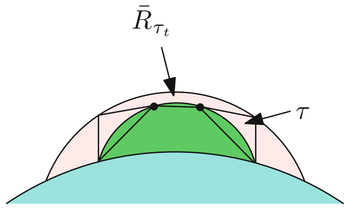
Suppose the invariants hold for $\mathcal{T}_{i-1}^*$ and E_{i-1}. We first prove (I1) for E_i. By construction, the new edges in $E_i \setminus E_{i-1}$ form a collection of convex or concave polygonal chains, where each chain γ^* starts and ends at vertices u, v of either ∂D_i^0 or $\partial \mathcal{U}(\mathcal{D}_{i-1})$. Moreover, by construction, u and v are connected to one another by a single arc γ of the respective boundary ∂D_i^0 or $\partial \mathcal{U}(\mathcal{D}_{i-1})$ (γ is either an exposed or a hidden subarc of $\partial \mathcal{U}(\mathcal{D}_{i-1})$, or a subarc of ∂D_i along ∂D_i^0), and the region $\overline{R}_\gamma$ between γ and γ^* does not contain any vertex of V_i in its interior.

Clearly, the edges in a single chain γ^* do not cross one another. Suppose to the contrary that an edge e of some (new) chain γ^* is crossed by an edge e' of some other (new or old) chain. Then either e' has an endpoint inside $\overline{R}_\gamma$, contradicting the construction, or e' crosses γ too, to exit from $\overline{R}_\gamma$, which again is impossible by construction, since no edge crosses ∂D_i^0 or $\partial \mathcal{U}(\mathcal{D}_{i-1})$. This establishes (I1).

(I2) follows easily from the construction and from the preceding discussion. Note that, for each polygonal chain γ^*, each of its endpoints is also an endpoint of exactly one neighboring arc $\hat{\gamma}^*$, so the union of these arcs consists of closed polygonal cycles, which bound some polygonal region, which we call $\mathcal{U}_i^*$, as claimed.

By construction, the vertical boundaries of the new polygonal pseudo-trapezoids of D_i^0 are contained in D_i^0 and do not cross any boundaries of other polygonal pseudo-trapezoids. This, together with (I1), imply that the new pseudo-trapezoids are pairwise openly disjoint, and are also openly disjoint from the polygonal pseudo-trapezoids in $\mathcal{T}_{i-1}^*$. It is also clear from the construction that each new pseudo-trapezoid $\sigma^* \in \mathcal{T}_i^* \setminus \mathcal{T}_{i-1}^*$ is contained in D_i. So (I3) follows.

Finally consider (I4). Each new hole that is created when adding D_i^0 is of one of the following kinds:



(a) The hole is a region of the form $\overline{R}_{\tau_b}$ or $\overline{R}_{\tau_t}$, for some $\tau \in \mathcal{T}_i \setminus \mathcal{T}_{i-1}$, such that $\overline{R}_{\tau_b}$ or $\overline{R}_{\tau_t}$ is contained in τ (if it lies outside τ, it becomes part of τ^*). Such a hole is contained in D_i, and is bounded by two concave or two convex chains, one of which, call it ζ^*, is polygonal, and the other, ζ, is part of ∂D_i^0. Moreover, ζ^*, if different from the chord e connecting the endpoints of ζ, passes through inner vertices of $\partial\mathcal{U}(\mathcal{D}_{i-1})$ that "stick into" the corresponding portion R_{τ_b} or R_{τ_t} of τ; see figure above.
(b) The hole is a region of the form $\overline{R}_\gamma$, for an exposed subarc γ of an arc of $\partial\mathcal{U}(\mathcal{D}_{i-1})$, that got delimited by a new vertex (an endpoint of some arc of ∂D_i). These holes are similar to those of type (a).
(c) The hole was part of a hole of type (a) or (b) in $\mathcal{U}(\mathcal{D}_{i-1})$, bounded by an arc γ of $\partial\mathcal{U}(\mathcal{D}_{i-1})$ and its associated polygonal chain γ^*, so that γ has been split into several subarcs (some hidden and some exposed) when adding D_i. For each of these subarcs ζ, we construct an associated polygonal chain ζ^*, either as a top or bottom side of some polygonal pseudo-trapezoid τ^* (constructed from a pseudo-trapezoid τ that has ζ as its top or bottom side), or as the polygonalization of an exposed subarc. The concatenation of the chains ζ^* results in a convex polygonal chain that is contained in $\overline{R}_\gamma$ and connects the endpoints of γ. The region enclosed between γ^* and ζ^* is an internal polygonal hole. Again, holes of type (c) can be seen all over Fig. 4; for example, see the top part of D_1 in Fig. 4(2′).

Holes of type (a) and (b) are *boundary half-polygonal holes*, whereas holes of type (c) are *internal polygonal holes*. Using the induction hypothesis that (I4) holds for $\mathcal{U}(\mathcal{D}_{i-1})$, we get that the union of the new holes of type (a) and (b), together with the old holes of type (a) and (b) corresponding to subarcs of $\partial\mathcal{U}(\mathcal{D}_i) \cap \partial\mathcal{U}(\mathcal{D}_{i-1})$, is $\mathcal{U}(\mathcal{D}_i) \setminus \mathcal{U}_i^*$. This completes the proofs of (I1)–(I4). □

Theorem 2.3

(a) Let $\mathcal{D}$ be a collection of $m \geq 3$ planar convex pseudo-disks, whose union covers the plane. Then there exists a set V of $O(m)$ points and a triangulation T of V that covers the plane, such that each triangle $\Delta \in T$ is fully contained in some member of $\mathcal{D}$.

(b) If $\mathcal{U}(\mathcal{D})$ is not the entire plane, it can be partitioned into $O(m)$ pairwise openly disjoint triangles and caps, such that each triangle and cap is fully contained in some member of $\mathcal{D}$.

Proof Since the number of vertices of M is $O(m)$, Euler's formula implies that $|E_m| = O(m)$ too. It is easily seen from the construction and from the invariants of Lemma 2.2, that each face of M is fully contained in some original pseudo-disk, so the same holds for each triangle. This establishes (a). Part (b) follows in a similar manner from the construction. □

2.4 *Efficient Construction of the Triangulation*

With some care, the proof of Theorem 2.3 can be turned into an efficient algorithm for constructing the required triangulation. This is a major advantage of the new proof over the older one. The algorithm is composed of building blocks that are variants of well-known tools, so we only give a somewhat sketchy description thereof.

2.4.1 Construction of the Original Pseudo-trapezoids

(A similar approach is mentioned in Matoušek et al. [43].) The construction proceeds by inserting the pseudo-disks of $\mathcal{D}$ in a *random* order, which, for simplicity, we denote as $D_1, \dots, D_m$. (Unlike the deterministic construction given above, here we do not guarantee that each D_i^0 has constant complexity. Nevertheless, as argued below, the random nature of the insertion order guarantees that this property holds on average.) As before, we put $\mathcal{D}_i = \{D_1, \dots, D_i\}$ for each i, and we maintain $\mathcal{U}(\mathcal{D}_i)$ after each insertion of a pseudo-disk. To do so efficiently, we maintain a vertical decomposition K_i of the *complement* $\mathcal{U}_i^c$ of the union $\mathcal{U}(\mathcal{D}_i)$ into vertical pseudo-trapezoids (as depicted in Fig. 5), and maintain, for each $\tau \in K_i$, a *conflict list*, consisting of all the pseudo-disks D_j that have not yet been inserted (i.e., with $j > i$), and that intersect τ.

Since the number of pseudo-trapezoids in the decomposition of the complement of the union of any k pseudo-disks (as depicted in Fig. 5) is $O(k)$ (an easy consequence of the linear bound on the union complexity [33]), a simple application of the Clarkson-Shor technique (similar to those used to analyze many other randomized incremental algorithms) shows that the expected overall number of these "complementary" pseudo-trapezoids that arise during the construction is $O(m)$, and that the expected overall size of their conflict lists is $O(m \log m)$.

When we insert a pseudo-disk D_i, we retrieve all the pseudo-trapezoids of K_{i-1} that intersect D_i. The union $\bigcup_{\tau \in K_{i-1}} (D_i \cap \tau)$ is precisely D_i^0. For each $\tau \in K_{i-1}$, the intersection $D_i \cap \tau$ (if nonempty) decomposes τ into $O(1)$ sub-trapezoids (this follows from the property that each of the four sides of τ crosses ∂D_i at most twice),

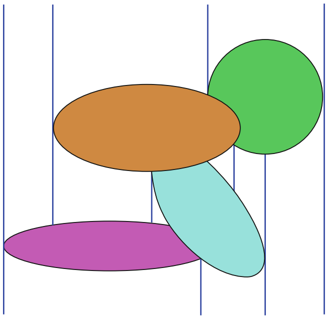


Fig. 5 The vertical decomposition of the complement of the union $\mathcal{U}(\mathcal{D}_i)$

some of which lie inside D_i (and, as just noted, form D_i^0), and some lie outside D_i, and form part of the new complement of the union $\mathcal{U}_i^c$.

Typically, the new pseudo-trapezoidal pieces are not necessarily real pseudo-trapezoids, as they may contain one or two "fake" vertical sides, because the feature that created such a side got "chopped off" by the insertion of D_i, and is no longer on the pseudo-trapezoid boundary. In this case, we "glue" these pieces together, across common fake vertical sides, to form the new real pseudo-trapezoids. We do it both for pseudo-trapezoids that are interior to D_i, and for those that are exterior. (This gluing step is a standard theme in randomized incremental constructions; see, e.g., [49].) This will produce (a) the desired vertical decomposition of D_i^0, and (b) the vertical decomposition K_i of the new union complement $\mathcal{U}_i^c$. The conflict lists of the new exterior pseudo-trapezoids (interior ones do not require conflict lists) are assembled from the conflict lists of the pseudo-trapezoids that have been destroyed during the insertion of D_i, again, in a fully standard manner.

To recap, this procedure constructs the vertical decompositions of all the regions D_i^0, so that the overall expected number of these pseudo-trapezoids is $O(m)$, and the total expected cost of the construction (dominated by the cost of handling the conflict lists) is $O(m \log m)$.

2.4.2 Construction of the Polygonal Chains and the Triangulation

By (I2) of Lemma 2.2, before D_i was inserted, each arc γ of $\partial\mathcal{U}(\mathcal{D}_{i-1})$ has an associated convex polygonal arc γ^* with the same endpoints. The union of the arcs γ^* forms a (possibly disconnected) polygonal curve within $\mathcal{U}(\mathcal{D}_{i-1})$, which partitions it into two subsets, the (polygonal) *interior*, $\mathcal{U}_{i-1}^*$, which is disjoint from $\partial\mathcal{U}(\mathcal{D}_{i-1})$ (except at the endpoints of the arcs γ^*), and the (half-polygonal) *exterior*, which is simply the (pairwise openly disjoint) union of the corresponding regions $\overline{R}_\gamma$.

To construct the triangulation, we maintain, for each polygonal chain γ^* of the boundary between the interior and the exterior, a list of its segments, sorted in left-to-right order of their x-projections, in a separate binary search tree (since the leftmost and rightmost points of each pseudo-disk are vertices in the construction, each chain γ^* is indeed x-monotone). We also maintain a triangulation of the interior. When we add D_i we update the lists representing the arcs γ and extend the triangulation of the interior to cover the "newly annexed" interior, as follows.

When D_i is inserted, some of the arcs γ of $\partial\mathcal{U}(\mathcal{D}_{i-1})$ are split into several subarcs. At most two of these arcs still appear on $\partial\mathcal{U}(\mathcal{D}_i)$, and each of them is an extreme subarc of γ (we call them, as above, *exposed* arcs). All the others are now contained in D_i (we call them *hidden*). Each endpoint of any new subarc is either an intersection point of ∂D_i with $\partial\mathcal{U}(\mathcal{D}_{i-1})$, or an endpoint of a vertical segment erected from some other vertex of D_i^0. (This also includes the case where an arc of $\partial\mathcal{U}(\mathcal{D}_{i-1})$ is fully "swallowed" by D_i and becomes hidden in its entirety.) In addition, $\partial\mathcal{U}(\mathcal{D}_i)$ contains ***fresh*** arcs, which are subarcs of ∂D_i along ∂D_i^0. The fresh subarcs and the hidden subarcs form the top and bottom sides of the new pseudo-trapezoids in the decomposition of D_i^0 (where each top or bottom side may be either fresh or

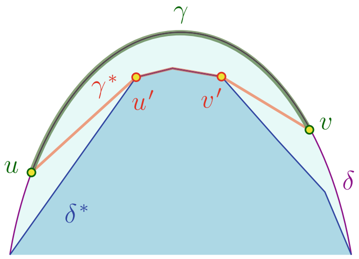


Fig. 6 Constructing the polygonal curve γ^* from γ

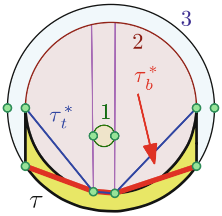
hidden). To obtain the top or bottom sides of some new pseudo-trapezoids we may have to concatenate several previously exposed subarcs of $\partial\mathcal{U}(\mathcal{D}_{i-1})$. These subarcs are connected at "inner" vertices of $\partial\mathcal{U}(\mathcal{D}_{i-1})$ which are not intersection points of the arrangement but intersections of vertical sides of pseudo-trapezoids which we already generated within $\mathcal{U}(\mathcal{D}_{i-1})$.)

The algorithm needs to construct, for each new exposed, hidden, and fresh arc γ, its associated polygonal curve γ^*. It does so in two stages, first handling exposed and hidden arcs, and then the fresh ones. Let γ be an exposed or hidden subarc, let δ denote the arc of $\partial\mathcal{U}(\mathcal{D}_{i-1})$, or the concatenation of several such arcs, containing γ, and let δ^* be its associated polygonal chain, or, in case of concatenation, the concatenation of the corresponding polygonal chains. As already noted, since the x-extreme points of each pseudo-disk boundary are vertices in the construction, δ and δ^* are both x-monotone.

If $\gamma = \delta$, we do nothing, as $\gamma^* = \delta^*$. Otherwise, let u and v be the respective left and right endpoints of γ. If uv does not intersect δ^* then γ^* is just the segment uv. Otherwise, γ^* is obtained from a portion of δ^*, delimited on the left by the point u' of contact of the right tangent from u to δ^*, and on the right by the point v' of contact of the left tangent from v to δ^*, to which we append the segments uu' on the left and $v'v$ on the right. See Fig. 6 for an illustration.

Note that the old arc δ may contain several new exposed or hidden arcs γ, so we apply the above procedure to each such arc γ. After doing this, the endpoints of δ (and of δ^*) are now connected by a new convex polygonal chain $\hat{\delta}^*$, which visits each of the new vertices along δ (the endpoints of the new arcs γ) and lies in between δ and δ^*. The region between $\hat{\delta}^*$ and δ^* is a new interior polygonal hole, and we partition it into simple cells, e.g., into vertical trapezoids, by a straightforward left-to-right scan.

Recall that some arcs τ_b and τ_t of new trapezoids τ may be concatenations of several hidden subarcs γ_i (connected at inner vertices which are not vertices of new trapezoids, as explained above). For each such arc, say τ_b, we obtain τ_b^* by concatenating the polygonal chains γ_i^* in x-monotone order.



We next handle the fresh arcs. Each such arc is the top or bottom side of some new pseudo-trapezoid τ, say it is the bottom side τ_b. If τ_t is also fresh, then τ is a convex pseudo-trapezoid, and we replace each of τ_b, τ_t by the straight segment connecting its endpoints. If τ_t is hidden, we take its associated chain τ_t^*, which we have constructed in the preceding stage, and form τ_b^* from it using the same procedure as above: Letting u and v denote the endpoints of τ_b, we check whether uv intersects τ_t^*. If not, τ_b^* is the segment uv. Otherwise, we compute the tangents from u and v to τ_t^*, and form τ_b^* from the tangent segments and the portion of τ_t^* between their contact points. See the figure above. We triangulate each polygonal pseudo-trapezoid τ once we have computed τ_b^* and τ_t^*.

2.4.3 Further Implementation Details

The actual implementation of the construction of the polygonal chains γ^* proceeds as follows. Given a new arc γ, which is a subarc of an old arc δ, we construct γ^* from δ^* as follows. Let u and v be the endpoints of γ. We (binary) search the list of edges of δ^* for the edge e_u whose x-projection contains the x-projection of u and for the edge e_v whose x-projection contains the x-projection of v. We then walk along the list representing δ^* from e_u towards e_v until we find the point u' of contact of the right tangent from u to δ^*. We perform a similar search from e_v towards e_u to find v'. (If we have traversed the entire portion of δ^* between e_u and e_v without encountering a tangent, we conclude that uv does not intersect δ^*, and set $\gamma^* := uv$.) We extract the sublist between u' and v' from δ^* by splitting δ^* at u' and v' and we insert the segments uu' and vv' at the endpoints of this sublist to obtain γ^*. We create the polygonalization of fresh arcs from their hidden counterparts in an analogous manner. Note that we destroy the representation of δ^* to produce the representation of γ^*. So in case the arc δ is split into several new subarcs, γ_i, some care has to be taken to maintain a representation of the remaining part of δ^* after producing each γ_j^*, from which we can produce the representation of the remaining subarcs γ_i.

For the analysis, we note that to produce γ^* we perform two binary searches to find e_u and e_v, each of which takes $O(\log m)$ time, and then perform linear scans to locate u' and v'. Each edge e traversed by these linear scans (except for $O(1)$ edges) drops off the boundary of the interior so we can charge this step to e and the total number of such charges is linear in the size of the triangulation.

2.5 The Result

The computation model In the preceding description, we implicitly assume a convenient model of computation, in which each primitive geometric operation that is needed by the algorithm, and that involves only a constant number of pseudo-disks (e.g., deciding whether two pseudo-disks or certain subarcs thereof intersect, computing these intersection points, and sorting them along a pseudo-disk boundary) takes constant time. In our application, described in the next section, the pseudo-disks are convex polygons, each having $O(k)$ edges. In this case, each primitive operation can be implemented in $O(\log k)$ time in the standard (say, real RAM) model, so the running time should be multiplied by this factor.

The preceding analysis implies the following theorem.

Theorem 2.4 *A confined triangulation of the union of m convex pseudo-disks, with $O(m)$ triangles and caps, can be computed in $O(m \log m)$ expected time, in a suitable model of computation where every primitive operation on a constant number of pseudo-disks takes $O(1)$ time. If the pseudo-disks are convex polygons each with at most k edges, then such a confined triangulation can be computed $O(m \log m \log k)$ expected time. If the pseudo-disks cover the plane then the triangulation consists only of triangles (each contained in a single pseudo-disk as required by a confined triangulation).*

2.6 Extension to General Convex Shapes

Theorem 2.4 uses only peripherally the property that the input shapes are pseudo-disks, and a simple modification (of the analysis, not of the construction itself) allows us to extend it to general convex shapes. Specifically, let $\mathcal{D}$ be a collection of m simply-shaped convex regions in the plane, such that the union complexity of any i of them is at most $u(i)$, where the complexity is measured, as before, by the number of boundary intersection points on the union boundary, and where $u(\cdot)$ is a monotone increasing function satisfying $u(i) = \Omega(i)$. We assume that the regions in $\mathcal{D}$ are simple enough so that the boundaries of any pair of them intersect only a constant number of times, and so that each primitive operation on them can be performed in reasonable time (which we take to be $O(1)$ in the statement below). The interesting cases are those in which $u(i)$ is small (that is, near-linear). They include, e.g., the case of fat triangles, or a low-density collection of convex regions; see [14] and references therein.

Deploying the algorithm of Theorem 2.4 results in the desired confined triangulation of $\mathcal{U}(\mathcal{D})$. Extending the analysis to this general setup (and omitting the straightforward technical details), we obtain the following theorem.

Theorem 2.5 *Let $\mathcal{D}$ be a collection of n convex shapes in the plane, such that the union complexity of any i of them is at most $u(i)$, where $u(i)$ is a monotone* *increasing function with $u(i) = \Omega(i)$. Then, a confined triangulation of $\mathcal{U}(\mathcal{D})$ with $O(u(m))$ triangles and caps (or just triangles if the union covers the entire plane), can be computed, in $O(u(m) \log m)$ expected time, under the assumption that every primitive geometric operation takes $O(1)$ time.*

3 Warm-Up Exercise: Constructing Cuttings in the Plane

In this section we apply the machinery developed in the previous section to obtain a new construction of $(1/r)$-cuttings in arrangements of lines in the planes.

Let L be a set of n lines in the plane in general position, and let $0 < r \leq n$ be a parameter. In the planar setup, a $(1/r)$-cutting for L is a partition of the plane into $O(r^2)$ pairwise openly disjoint triangles, such that (the interior of) each triangle is crossed by at most n/r lines of L.

The construction of cuttings in the plane that we present here is similar in spirit to the more involved scheme for approximating the level in arrangements of planes in three dimensions, as presented in the following Sect. 4.

3.1 Tools

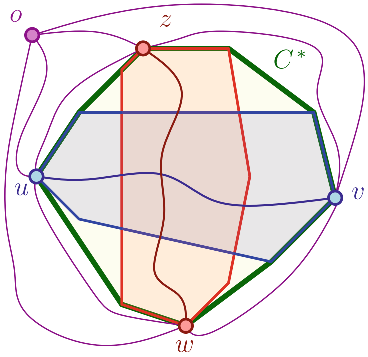
3.1.1 Divisions

We begin by reviewing the construct of a *κ-division* of a planar graph, which is a decomposition of such a graph into subgraphs, and a refined and stronger variant of the planar separator theorem of Lipton and Tarjan [37] and Miller's cycle separator theorem [44]. It goes back to Frederickson's 30-years-old work [28], and has eventually culminated in the fast κ-division algorithm of Klein et al. [34]. We remind the reader that a graph is *biconnected* if any pair of vertices are connected by at least two vertex-disjoint paths.

Definition 3.1 (Frederickson [28]) Given a non-crossing plane drawing of a planar triangulated and biconnected graph G with N vertices, and a parameter $\kappa < N$, a *κ-division* of G is a decomposition of G into m connected subgraphs $G_1, \ldots, G_m$, such that

(i) $m = O(N/\kappa)$,
(ii) each G_i has at most κ vertices,
(iii) each G_i has at most $\beta\sqrt{\kappa}$ *boundary vertices*, for some absolute constant β, namely, vertices that belong to at least one additional subgraph; and
(iv) each G_i has at most $O(1)$ *holes*, namely, faces of the induced drawing of G_i that are not faces of G (as they contain additional edges and vertices of G).

Fig. 7 Illustrating the proof of Lemma 3.2



As shown in Klein et al. [34], a κ-division of a planar triangulated and biconnected graph with N vertices can be computed in $O(N)$ time.[2]

3.1.2 The Convex Hulls of Pairwise Openly Disjoint Polygons Are Pseudo-disks

Another tool that we need is the following folklore result, whose proof is included for the sake of completeness.

Lemma 3.2 *Let P and P' be two connected polygons in the plane with disjoint interiors, and let C and C' denote their respective convex hulls. Then ∂C and $\partial C'$ intersect each other at most twice.*

Proof For simplicity of exposition, we assume that P and P' are in general position, in a sense that will become more concrete from the proof. The analysis easily extends to the more general case too.

Assume, for the sake of contradiction, that ∂C and $\partial C'$ cross more than twice (in general position, the boundaries do not overlap). This implies that each of $\partial C \setminus C'$, $\partial C' \setminus C$ is disconnected, and thus there exist four vertices u, w, v, and z of the boundary of the convex hull $C^* = \mathrm{CH}(C \cup C')$, that appear along ∂C^* in this circular order, so that $u, v \in \partial C \setminus C'$ and $w, z \in \partial C' \setminus C$, see Fig. 7. Clearly, u and v are also vertices of P, and w and z are vertices of P'.

We show that this scenario leads to an impossible planar drawing of K_5. For this, let o be an arbitrary point outside C^*. Connect o to each of u, v, w, z by noncrossing arcs that lie outside C^*, and connect u, w, v, and z by the four respective portions of ∂C^* between them. Finally, connect u to v by a path contained in P, and connect w to z by a path contained in P'. The resulting ten edges are pairwise noncrossing, where, for the last pair of edges, the property follows from the disjointness of (the interiors

[2] The algorithm of [34] constructs κ-divisions for a geometrically increasing sequence of values of the parameter κ, in overall $O(N)$ time.

of) P and P'. The contradiction resulting from this impossible planar drawing of K_5 establishes the claim. □

Note that the above proof does not require the polygons to be simply connected.

Corollary 3.3 *Let $\mathcal{P} = \{P_1, \ldots, P_m\}$ be a set of m pairwise openly disjoint connected polygons in the plane, and let C_i denote the convex hull of P_i, for $i = 1, \ldots, m$. Then $\mathcal{C} := \{C_1, \ldots, C_m\}$ is a collection of m convex (polygonal) pseudo-disks.*

3.2 Construction of Cuttings in Two Dimensions

Combining the tools from the previous subsection, we obtain the following new construction of cuttings in the plane.

Theorem 3.4 *Given a set L of n lines in the plane, in general position, and a parameter $0 < r \le n$, one can decompose the plane into $O(r^2)$ pairwise openly disjoint triangles, such that (the interior of) each triangle is crossed by at most n/r lines of L.*

Proof Consider the arrangement $\mathcal{A}(L)$, add a "fake" vertex at infinity, which serves as a common endpoint of all the unbounded rays in $\mathcal{A}(L)$, and triangulate every (bounded or unbounded) face of $\mathcal{A}(L)$ with more than three boundary edges, by adding diagonals. Let G be the resulting planar graph, whose vertices are the vertices of $\mathcal{A}(L)$, and each of whose edges is either an original (bounded or unbounded) edge of $\mathcal{A}(L)$, or one of the added diagonals. Clearly, G is planar, triangulated and, as is easily checked, also biconnected. It has $N := 1 + \binom{n}{2}$ vertices.

Construct a κ-division of G, for $\kappa = \left(\frac{n}{\beta r}\right)^2$, where β is the constant from the construction of κ-divisions. We get a partition of the plane into $m = O(N/\kappa) = O(n^2/\kappa) = O(r^2)$ subgraphs $G_1, \ldots, G_m$, and we turn each subgraph G_i into a (not necessarily simple) polygon P_i by forming the union of all the faces of G_i that are also faces of G. By the properties of κ-divisions, each P_i has at most $\beta\sqrt{\kappa} \le n/r$ vertices (and edges) of $\mathcal{A}(L)$ on its boundary.

We clean up the construction, as follows. If one of the polygons B in this collection has a hole, we remove the hole from B (i.e., add its area to B), and remove all the polygons contained inside the hole from the collection. We repeat this process until all the polygons are simple, and obtain a partition of the plane into $m = O(r^2)$ simple pairwise openly disjoint polygons $B_1, \ldots, B_m$, such that each polygon has at most $t := n/r$ vertices of $\mathcal{A}(L)$ on its boundary.

Since every line of L intersects ∂B at a vertex, and every line that intersects the interior of B must cross its boundary at least twice, it follows that the interior of B is crossed by at most t lines of L.

Now form the convex hulls $C_i = \mathrm{CH}(B_i)$, for $i = 1, \ldots m$. By Corollary 3.3, the set $\{C_1, \ldots, C_m\}$ is a set of m convex pseudo-disks. Hence, by Theorem 2.5, one can compute a triangulation of the plane into $O(m)$ triangles, such that each triangle is fully contained in one of these hulls. Since a line intersects the interior of C_i if and only if it intersects the interior of B_i, it follows that at most t lines of L can intersect the interior of C_i, for $i = 1, \ldots, m$, and therefore every triangle in the triangulation is crossed by at most $t = n/r$ lines. This shows that $\mathcal{T}$ is a $(1/r)$-cutting of $\mathcal{A}(L)$ of size $O(r^2)$, as desired. □

We remark that a disadvantage of this construction is that it takes $O(N) = O(n^2)$ time to perform. This also holds, by the way, for a naive implementation of Matoušek's deterministic construction of planar cuttings [38].

4 Construction of Shallow Cuttings and Approximate Levels in Three Dimensions

We begin by presenting a high-level description of the technique, filling in the technical details in subsequent subsections. The high-level part does not pay too much attention to the efficiency of the construction; this is taken care of later in this section.

4.1 Sketch of the Construction

Let H be a set of n planes in three dimensions in general position. Assume that, for a given parameter $0 < r \leq n$, we want to approximate level $k = n/r$ of $\mathcal{A}(H)$. Note that when r is too close to n, that is, when k is a constant, we can simply compute the k-level explicitly and use it as its own approximation. The complexity of such a level is $O(n)$, and it can be computed in $O(n \log n)$ time [11, 18] (better than what is stated in Theorem 4.6 below for such a large value of r). We therefore assume in the remainder of this section that $r \ll n$.

Put $k_1 := (1 + c)n/r$ and $k_2 := (1 + 2c)n/r$, for a suitable sufficiently small (but otherwise arbitrary) constant fraction c. The analysis of Clarkson and Shor [25] implies that the overall complexity of $L_{\leq k_2}$ (the first k_2 levels of $\mathcal{A}(H)$) is $O(nk^2)$. This in turn implies that there exists an index $k_1 \leq \xi \leq k_2$ for which the complexity $|L_\xi|$ of L_ξ is $O(nk^2/(cn/r)) = O(nk/c) = O(n^2/(cr))$. We fix such a level ξ, and continue the construction with respect to L_ξ (slightly deviating from the originally prescribed value of k). However, to simplify the notation for the current part of the analysis, we use k to denote the nearby level ξ, and will only later return to the original value of k.

The next step is to decompose the xy-projection of the k-level L_k, using the κ-division technique reviewed in Sect. 3.1.1. Specifically, we set

$$\kappa := \left(\frac{cn - 43.5r}{9\beta r}\right)^2,$$

where β is the constant from property (iii) of κ-divisions (see Sect. 3.1.1). Notice that since $r \ll n$ we have $\kappa > 1$. Let L'_k denote the projection of L_k onto the xy-plane. We turn L'_k into a triangulated and biconnected planar graph G'_k, similarly to the way in which we handled planar arrangements of lines in Sect. 3.2. That is, we add a new vertex v_∞ at infinity, replace each ray $[p, \infty)$ of L'_k by the edge (p, v_∞), and triangulate each bounded or unbounded face, if needed, by adding diagonals. The resulting graph is planar and triangulated, and, as is easily checked, is also biconnected. We can therefore apply to G'_k the planar κ-division algorithm of Klein et al. [34], as reviewed in Sect. 3.1.1, with the value of κ given above. The resulting κ-division of G'_k consists of

$$m := O(|L_k|/\kappa) = O\left(\frac{n^2/(cr)}{c^2n^2/r^2}\right) = O(r/c^3)$$

connected, possibly unbounded, polygons, $P_1, \ldots, P_m$, with pairwise disjoint interiors. The union of $P_1, \ldots, P_m$ covers the entire xy-plane, and the edges of these polygon are projections of (some) edges of L_k (including the diagonals drawn to triangulate the original faces of L_k).

By construction, each P_i is connected and has at most $\beta\sqrt{\kappa} = (cn - 43.5r)/(9r)$ edges (and also contains $O(\kappa)$ edges and vertices of the projected k-level in its interior). Let C_i denote the convex hull of P_i, for $i = 1, \ldots, m$. As shows in Corollary 3.3, $\mathcal{C} := \{C_1, \ldots, C_m\}$ is a collection of m (possibly unbounded) convex pseudo-disks whose union is the entire plane.

We then apply Theorem 2.3 to $\mathcal{C}$ and obtain a set S of $O(m)$ points in the xy-plane, and a triangulation T of S that covers the plane, such that each triangle $\Delta \in T$ is fully contained in some hull C_i in $\mathcal{C}$.

For a point p in the xy-plane, we denote by $\uparrow_k(p)$ the *lifting* of p to the k-level, i.e., the unique point on the level that is co-vertical with p. For a bounded triangle Δ of T, $\uparrow_k(\Delta)$ is defined as the triangle spanned by the lifted images of the three vertices of Δ. We lift an unbounded triangle Δ with vertices p, q, and v_∞ by lifting pq to $\uparrow_k(p)\uparrow_k(q)$, as before, and lifting each of its rays, say $[p, \infty)$, as follows. If $[p, \infty)$ is the projection of an original ray of L_k, we simply lift it to that ray. Otherwise, we lift $[p, \infty)$ to a ray $\uparrow([p, \infty))$ that emanates from $\uparrow_k(p)$ in a direction parallel to the plane which lies vertically above $[p, \infty)$ at infinity. If the liftings $\uparrow([p, \infty))$, and $\uparrow([q, \infty))$, and the edge $\uparrow_k(p)\uparrow_k(q)$ are not coplanar, we add another ray r emanating from $\uparrow_k(p)$ parallel to $\uparrow([q, \infty))$. We add to T' the unbounded triangle spanned by $\uparrow([q, \infty))$, $\uparrow_k(p)\uparrow_k(q)$, and r, and the unbounded wedge spanned by r and $\uparrow([p, \infty))$. Let T' denote the corresponding collection of lifted (bounded and unbounded) triangles and wedges in $\mathbb{R}^3$, given by $T' = \{\uparrow_k(\Delta) \mid \Delta \in T\}$.

Note that the triangles of T' are in general not contained in L_k. However, for each triangle $\Delta' \in T'$, its (finite) vertices lie on L_k, and, as we show in Lemma 4.5 below, at most $9\beta\sqrt{\kappa} + 43.5 = cn/r$ planes of H can cross Δ'. This implies, returning now to the original value of k, that Δ' fully lies between the levels $\xi \pm cn/r$ of $\mathcal{A}(H)$. In particular, Δ' lies fully above the level

$$\xi - cn/r \geq k_1 - cn/r = n/r = k,$$

and fully below the level

$$\xi + cn/r \leq k_2 + cn/r = (1 + 3c)n/r = (1 + 3c)k.$$

The lifted triangulation T' forms a polyhedral terrain that consists of $O(r/c^3)$ triangles and is contained between the levels $k = n/r$ and $(1 + 3c)k$. That is, for a given $\varepsilon > 0$, choosing $c = \varepsilon/3$ makes T' an ε-approximation of L_k, and we obtain the following result.

Theorem 4.1 *Let H be a set of n non-vertical planes in $\mathbb{R}^3$ in general position, and let $0 < r \leq n$, $\varepsilon > 0$ be given parameters. Then there exists a polyhedral terrain consisting of $O(r/\varepsilon^3)$ triangles, that is fully contained between the levels n/r and $(1 + \varepsilon)n/r$ of $\mathcal{A}(H)$, which can be computed in polynomial time.*

(The last assertion in the theorem is a consequence of the constructive nature of our analysis. Efficient implementation of this construction is described later in this section.)

To turn this approximate level into a shallow cutting, replace each $\Delta' \in T'$ (including each of the unbounded triangles and wedges, as constructed above) by the semi-unbounded vertical prism Δ^* consisting of all the points that lie vertically below Δ'. This yields a collection Ξ of prisms, with pairwise disjoint interiors, whose union covers $L_{\leq n/r}$, so that, for each prism τ of Ξ, we have (a) each vertex of τ lies at level (at least k and) at most $(1 + \frac{2}{3}\varepsilon)k$, and, as will be established in Lemma 4.5 below, (b) the top triangle of τ is crossed by at most $\frac{1}{3}\varepsilon k$ planes of H (in the preceding analysis, we wrote this bound as $\frac{cn}{r}$; this is the same value, recalling that $\varepsilon = 3c$ and $k = n/r$). Hence, as is easily seen, each prism of Ξ is crossed by at most $(1 + \varepsilon)n/r$ planes, so Ξ is the desired shallow cutting. That is, we have the following result.

Theorem 4.2 *Let H be a set of n non-vertical planes in $\mathbb{R}^3$ in general position, let $k < n$ and $\varepsilon > 0$ be given parameters, and put $r = n/k$. Then there exists a k-shallow $((1 + \varepsilon)/r)$-cutting of $\mathcal{A}(H)$, consisting of $O(r/\varepsilon^3)$ vertical prisms (unbounded from below). The top of each prism is a triangle or a wedge that is fully contained between the levels k and $(1 + \varepsilon)k$ of $\mathcal{A}(H)$, and these triangles form a polyhedral terrain (we say that such a terrain approximates the k-level L_k up to a relative error of ε).*

4.2 Crossing Properties of the Planar Subdivision

Recall that our construction computes a κ-division of the xy-projection L'_k of L_k where $\kappa := ((cn - 43.5r)/9\beta r)^2$ (and recall that $k = n/r$, $r \ll n$, and β is a constant). Our goal in the rest of this section is to show that the lifting $\uparrow_k(\Delta)$ of any triangle Δ contained in the convex hull C of a polygon P of this decomposition intersects at most ck planes of H. We prove this explicitly for bounded triangles, and the proof for unbounded triangles (or wedges) is similar.

Recall that, for a point p in the xy-plane, we denote by $\uparrow_k(p)$ the (unique) point that lies on L_k and is co-vertical with p. The *crossing distance* $\mathbf{cr}(p, q)$ between any pair of points $p, q \in \mathbb{R}^3$, with respect to H, is the number of planes of H that intersect the closed segment pq. The crossing distance is a quasi-metric, in that it is symmetric and satisfies the triangle inequality. For a connected set $X \subseteq \mathbb{R}^3$, the *crossing number* $\mathbf{cr}(X)$ of X is the number of planes of H intersecting X (thus $\mathbf{cr}(p, q)$ is the crossing number of the closed segment pq).

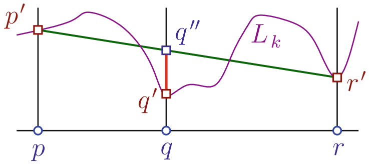
Lemma 4.3 *Let p, q, r be three collinear points in the xy-plane, such that $q \in pr$, and let $p' = \uparrow_k(p)$, $q' = \uparrow_k(q)$, and $r' = \uparrow_k(r)$; these points, which lie on the k-level, are in general not collinear. Let q'' be the intersection of the vertical line through q with the segment $p'r'$. Then we have $\mathbf{cr}(q'', q') \leq \frac{1}{2}\mathbf{cr}(p', r') + 8.5$. See figure below.*

Proof For a point u, we denote by level(u) the number of planes lying vertically strictly below u. Put $k'' =$ level(q''). The point p' lies at level k, which is the closure of all points of level k. Thus the number of planes lying vertically strictly below q is k if p' is in the relative interior of a face of level k, at least $k-1$ if p' is in the relative interior of an edge of level k, and at least $k - 2$ if p' is a vertex of level k. In either case, we have level$(p') \geq k - 2$, and similarly for r', and thus

$$\mathbf{cr}(p', q'') \geq |\text{level}(p') - \text{level}(q'')| \geq |k - k''| - 2 ,$$

and

$$\mathbf{cr}(q'', r') \geq |\text{level}(p') - \text{level}(q'')| \geq |k - k''| - 2 .$$



On the other hand we have

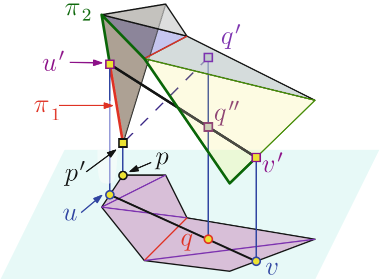
$$\mathbf{cr}(q', q'') \leq |k'' - \text{level}(q')| + 3 \leq |k - k''| + 5 .$$

(Indeed, if q'' lies above q' then $|k'' - \text{level}(q')| \leq |k'' - (k - 2)| \leq |k'' - k| + 2$, and if q' lies above q'' then $|k'' - \text{level}(q')| \leq |k'' - k|$. In addition, the difference in the levels of q' and q'' does not count the at most three planes that intersect $q'q''$ at q'', if q'' is above q', or at q', otherwise; this accounts for the terms 3 and 5 in the preceding inequality.) Hence,

$$\begin{aligned}\mathbf{cr}(q',q'') &\le \frac{1}{2}\Big(\mathbf{cr}(p',q'')+\mathbf{cr}(q'',r')+4\Big)+5 \le \frac{1}{2}\big(\mathbf{cr}(p',r')+3\big)+7\\ &= \frac{1}{2}\mathbf{cr}(p',r')+8.5,\end{aligned}$$

where the term 3 in the next-to-last expression is due to the potential double counting of the (up to) three planes passing through q'', in both terms $\mathbf{cr}(p',q'')$ and $\mathbf{cr}(q'',r')$. □

In what follows, we consider polygonal regions contained in L_k, where each such region R is a connected union of some of the faces of L_k. The xy-projection of R is a connected polygon in the xy-plane, and, for simplicity, we refer to R itself also as a polygon.



Lemma 4.4 *Let H be a set of n non-vertical planes in $\mathbb{R}^3$ in general position. Let P' be a bounded connected polygon with t edges that lies on the k-level L_k of $\mathcal{A}(H)$, such that all the boundary edges of P' are edges of L_k. Let p' be a vertex of the external boundary of P', and let q be any point in the convex hull C of the xy-projection P of P'. Then the crossing distance between p' and $q' = \uparrow_k(q)$ is at most* $3t + 14.5$.

Proof Since q lies in C, we can find two points u, v on the external boundary of P such that $q \in uv$. Put $q' = \uparrow_k(q)$, $u' = \uparrow_k(u)$, and $v' = \uparrow_k(v)$, and denote by q'' the point that lies on the segment $u'v'$ and is co-vertical with q. See figure above. We have

$$\mathbf{cr}(p',q') \le \mathbf{cr}(p',u')+\mathbf{cr}(u',q'')+\mathbf{cr}(q'',q') \le \mathbf{cr}(p',u')+\mathbf{cr}(u',v')+\mathbf{cr}(q'',q').$$

Let π_1 and π_2 be the two portions of the external boundary that connect p' and u', and u' and v', respectively, and that do not overlap. Now, by Lemma 4.3, we have

$\mathbf{cr}(q'', q') \leq \frac{1}{2}\mathbf{cr}(u', v') + 8.5$, so we get

$$\begin{aligned}\mathbf{cr}(p', q') &\leq \mathbf{cr}(p', u') + \frac{3}{2}\mathbf{cr}(u', v') + 8.5 \leq \mathbf{cr}(\pi_1) + \frac{3}{2}\mathbf{cr}(\pi_2) + 8.5 \\ &\leq \frac{3}{2}\mathbf{cr}(\partial P') + 13,\end{aligned}$$

where $\partial P'$ denotes the external boundary of P', and where the last inequality follows because $\frac{3}{2}\mathbf{cr}(\pi_1) + \frac{3}{2}\mathbf{cr}(\pi_2)$ double counts the planes that pass through u', adding at most $\frac{3}{2} \cdot 3 = 4.5$ to the bound.

To bound the number of planes of H that intersect $\partial P'$, consider its vertices $p_1, p_2, \ldots, p_t$ (the actual number of vertices might be smaller since P' may not be simply connected). Observe that p_1 is contained in three planes. For each i, p_i lies on at most two planes that do not contain p_{i-1} (there are two such planes when $p_{i-1}p_i$ is a diagonal of an original face of the untriangulated level L_k). Furthermore, the open segment $p_{i-1}p_i$ does not cross any plane, and each plane that contains it contains both its endpoints. Therefore, the number $\mathbf{cr}(\partial P')$ of planes of H that intersect $\partial P'$ satisfies $\mathbf{cr}(\partial P') \leq 3 + 2(t-1) = 2t + 1$, from which the lemma follows. (Note that this analysis is somewhat conservative—for example, if the polygon P' uses only original edges of the k-level, the bound drops to $t + 2$.) □

Lemma 4.5 *Let H be a set of n non-vertical planes in $\mathbb{R}^3$ in general position, and let P' be a connected polygon with t edges, such that P' lies on the k-level L_k of $\mathcal{A}(H)$, and such that all the boundary edges of P' are edges of L_k. Then, for any triangle $\Delta = \Delta pqr$ that is fully contained in the convex hull of the xy-projection of P', the number $\mathbf{cr}(\Delta')$ of planes of H that cross the triangle $\Delta' = \Delta p'q'r'$, where $p' = \uparrow_k(p)$, $q' = \uparrow_k(q)$, $r' = \uparrow_k(r)$, is at most $9t + 43.5$.*

Proof Let w be any vertex of the external boundary of P'. Any plane that crosses Δ' must also cross two of its sides. Moreover, by Lemma 4.4 and the triangle inequality,

$$\mathbf{cr}(p', q') \leq \mathbf{cr}(w, p') + \mathbf{cr}(w, q') \leq 2(3t + 14.5),$$

and similarly for $\mathbf{cr}(p', r')$ and $\mathbf{cr}(q', r')$. Adding up these bounds and dividing by 2, implies the claim. □

By Property (iii) of Definition 3.1 our polygons have at most $t = \beta\sqrt{\kappa} = (cn - 43.5r)/9r$ edges. Therefore by Lemma 4.5 any triangle of the polyhedral terrain of Theorem 4.1 is crossed by at most cn/r planes.

4.3 Efficient Implementation

We next turn our constructive proof into an efficient algorithm, and show:

Theorem 4.6 *Let H be a set of n non-vertical planes in $\mathbb{R}^3$ in general position, let $k < n$ and $\varepsilon > 0$ be given parameters, and put $r = n/k$. Then we have:*

(a) *One can construct the k-shallow $((1 + \varepsilon)/r)$-cutting of $\mathcal{A}(H)$ given in Theorem 4.2, or, equivalently, the ε-approximating terrain of the k-level in Theorem 4.1, in $O(n + r\varepsilon^{-6} \log^3 r)$ expected time. This algorithm computes a correct ε-approximating terrain with probability at least $1 - 1/r^{O(1)}$.*

(b) *Computing the conflict lists of the vertical prisms takes an additional $O(n(\varepsilon^{-3} + \log \frac{r}{\varepsilon}))$ expected time.*

(c) *If we also compute the conflict lists then we can verify, in $O(n/\varepsilon^3)$ time, that the cutting is indeed correct and thereby make the algorithm always succeed, at the cost of increasing its expected running time by a constant factor.*

Proof

(a) We first describe a straightforward implementation of the algorithm described in Sect. 4.1. We then apply this implementation to a random sample to get the desired time bound.

The first step of the algorithm is to construct level L_ξ of smallest complexity between levels $k_1 = (1 + \varepsilon)k$ and $k_2 = (1 + 2\varepsilon)k$ in $\mathcal{A}(H)$. To get this level we compute all the first k_2 levels in $\mathcal{A}(H)$, using a randomized algorithm of Chan [16],[3] which takes $O(n \log n + nk^2)$ expected time and then extract the desired level L_ξ. Recall that the complexity of L_ξ is $|L_\xi| := O(nk/\varepsilon)$.

We project L_ξ onto the xy-plane, and construct a κ-division of the projection for $\kappa = \Theta(k^2)$, in $O(|L_\xi|)$ time. This κ-division consists of $m = O(r/\varepsilon^3)$ pieces and the boundary of each piece consists of $O(\sqrt{\kappa})$ edges. We compute the convex hull of each piece of the κ-division, in $O(m\sqrt{\kappa})$ overall time and construct the confined triangulation of these convex hulls in $O(m \log m \log \sqrt{\kappa})$ time.

Finally, we lift the vertices of the resulting triangles to L_ξ. This can be done, using a point location data structure over the xy-projection of L_ξ, in $O(|L_\xi| \log |L_\xi| + m \log |L_\xi|)$ time. This completes the construction (excluding the construction of the conflict lists). Summing up over all stages, we obtain that the overall expected construction time is

$$O\left(n \log n + nk^2 + m\sqrt{\kappa} + m \log m \log \sqrt{\kappa} + |L_\xi| \log |L_\xi| + m \log |L_\xi|\right).$$

[3] The paper of Chan [16] does not use shallow cuttings.

Substituting $\sqrt{\kappa} = O(k)$, $k = n/r$, $|L_\xi| := O(nk/\varepsilon)$, and $m = O(r/\varepsilon^3)$, we obtain that the running time is

$$O\left(n\log n + \frac{n^3}{r^2} + \frac{n}{\varepsilon^3} + \frac{r}{\varepsilon^3}\log\frac{r}{\varepsilon}\log\frac{n}{r} + \frac{n^2}{r\varepsilon}\log\frac{n}{r\varepsilon} + \frac{r}{\varepsilon^3}\log\frac{n}{r\varepsilon}\right). \qquad (1)$$

The idea is to apply this construction to an approximate level $k' = \frac{n'}{r} = \frac{b}{\varepsilon^2}\log r$ of a random sample S of $n' = \frac{br}{\varepsilon^2}\log r$ planes of H, where b is a suitable constant. The dominant term in the running time bound is the second term in Eq. (1), which is $O(r\varepsilon^{-6}\log^3 r)$.

We prove the correctness of this procedure as follows.

Let $(H, \mathcal{R})$ denote the range space in which each range in $\mathcal{R}$ corresponds to some vertical segment or ray e, and is equal to the subset of the planes of H that cross e. Clearly, $(H, \mathcal{R})$ has finite VC-dimension (see, e.g., [50]). A random sample S of $n' = \frac{br}{\varepsilon^2}\log r$ planes from H, for a sufficiently large constant b, is a *relative* $\left(\frac{1}{r}, \varepsilon\right)$*-approximation* for $(H, \mathcal{R})$, with probability $\geq 1 - 1/r^{O(1)}$; see [30] for full details concerning the definition and properties of relative approximations. In our context, this means (assuming that the sample is indeed a relative approximation) that each vertical segment or ray that intersects $x \geq n/r$ planes of H intersects between $(1+\varepsilon)\frac{n'}{n}x$ and $(1-\varepsilon)\frac{n'}{n}x$ planes of S, and each vertical segment or ray that intersects $x < n/r$ planes of H intersects at most $\frac{n'}{n}x + \varepsilon\frac{n'}{r}$ planes of S. (This holds, with probability $\geq 1 - 1/r^{O(1)}$, for all vertical segments and rays.)

It follows from Theorem 4.1 that our approximation of level k' of $\mathcal{A}(S)$ is a terrain T of size $O(r/\varepsilon^3)$ that lies between level k' and level $(1+\varepsilon)k'$ of $\mathcal{A}(S)$. We claim that T also lies between levels k and $(1+4\varepsilon)k$ of $\mathcal{A}(H)$ and therefore (up to a scaling of ε) gives the desired level approximation.

To justify this claim, consider a point p on level k of $\mathcal{A}(H)$. By the properties specified above, of a relative $\left(\frac{1}{r}, \varepsilon\right)$-approximation, it follows that the level of p in $\mathcal{A}(S)$ is at most $(1+\varepsilon)(n'/n)(n/r) = k'$. Similarly, let p be a point at level larger than, say, $(1+4\varepsilon)k$ of $\mathcal{A}(H)$. Then the level of p in $\mathcal{A}(S)$ is at least $(1-\varepsilon)(n'/n)(1+4\varepsilon)(n/r) \geq (1+\varepsilon)k' = k' + t'$, for $\varepsilon \leq 1/2$. Since this holds with probability $\geq 1 - 1/r^{O(1)}$, for every point p, we conclude that T lies between levels k and $(1+4\varepsilon)k$ of $\mathcal{A}(H)$, with probability $\geq 1 - 1/r^{O(1)}$.

Our algorithm can fail only if S fails to be a relative approximation. As mentioned, this happens with probability at most $1/r^{O(1)}$.

(b) We now describe an algorithm that computes for every semi-unbounded vertical prism Δ^* stretching below a triangle Δ of our approximating terrain T, the set of planes of H that intersect it (i.e., the conflict list of the prism). To this end, we put the vertices of T into the range reporting data structure of Chan [16]. In this structure, after preprocessing, in $O(\frac{r}{\varepsilon^3}\log\frac{r}{\varepsilon})$ expected time, one can report, for any given query half-space h^+, the points in $h^+ \cap T$, in $O(\log\frac{r}{\varepsilon} + |h^+ \cap T|)$ expected time (we recall again that this data range reporting structure of Chan is simple and does not use shallow cuttings). We query this data structure with the set of halfspaces h^+, bounded from below by the respective planes $h \in H$, and, for each vertex x of T that we report, we add h to the conflict lists of the prisms incident to x. This takes $O(n \log \frac{r}{\varepsilon} + \frac{n}{\varepsilon^3})$ expected time, since the total size of the conflict lists is $O(\frac{r}{\varepsilon^3} \cdot \frac{n}{r}) = O(\frac{n}{\varepsilon^3})$ (in expectation and with probability $\geq 1 - 1/r^{O(1)}$).

(c) The probability that the sample S fails to be a relative $\left(\frac{1}{r}, \varepsilon\right)$-approximation for $(H, \mathcal{R})$ is at most $1/r^{O(1)}$. When the sample does indeed fail, T may fail to be the desired k-shallow $((1 + \varepsilon)/r)$-cutting. Such a failure happens if and only if there exists a vertex of T whose conflict list is of size smaller than k or larger than $(1 + \varepsilon)k$. When we detect such a conflict list, we repeat the entire computation. Since the failure probability is small the expected number of times we will repeat the computation is (a small) constant. □

Acknowledgements We thank János Pach for pointing out that a variant of Theorem 2.3 is already known. We also thank the anonymous referees for their useful feedback.

References

1. P. Afshani, T.M. Chan, Optimal halfspace range reporting in three dimensions, in *Proceedings of the 20th ACM-SIAM Symposium on Discrete Algorithms (SODA)* (2009), pp. 180–186
2. P. Afshani, K. Tsakalidis, Optimal deterministic shallow cuttings for 3d dominance ranges, in *Proceedings of the 25th ACM-SIAM Symposium on Discrete Algorithms (SODA)* (2014), pp. 1389–1398
3. P. Afshani, C.H. Hamilton, N. Zeh, A general approach for cache-oblivious range reporting and approximate range counting. Comput. Geom. Theory Appl. **43**(8), 700–712 (2010)
4. P. Afshani, T.M. Chan, K. Tsakalidis, Deterministic rectangle enclosure and offline dominance reporting on the RAM, in *Proceedings of the 41st International Colloquium on Automata, Languages and Programming (ICALP).* Volume 8572 of Lecture Notes in Computer Science (Springer, 2014), pp. 77–88
5. P.K. Agarwal, Partitioning arrangements of lines I: an efficient deterministic algorithm. Discrete Comput. Geom. **5**, 449–483 (1990)
6. P.K. Agarwal, Partitioning arrangements of lines: II. Applications. Discrete Comput. Geom. **5**, 533–573 (1990)
7. P.K. Agarwal, Geometric partitioning and its applications, in *Computational Geometry: Papers from the DIMACS Special Year*, ed. by J.E. Goodman, R. Pollack, W. Steiger (American Mathematical Society, Providence, 1991), pp. 1–37
8. P.K. Agarwal, *Intersection and Decomposition Algorithms for Planar Arrangements* (Cambridge University Press, New York, 1991)
9. P.K. Agarwal, P.K. Desikan, An efficient algorithm for terrain simplification, in *Proceedings of the 8th ACM-SIAM Symposium on Discrete Algorithms (SODA)* (1997), pp. 139–147
10. P.K. Agarwal, J. Erickson, Geometric range searching and its relatives, in *Advances in Discrete and Computational Geometry*, ed. by B. Chazelle, J.E. Goodman, R. Pollack (American Mathematical Society, Providence, 1999), pp. 1–56
11. P.K. Agarwal, J. Matoušek, Dynamic half-space range reporting and its applications. Algorithmica **13**, 325–345 (1995)
12. P.K. Agarwal, S. Suri, Surface approximation and geometric partitions. SIAM J. Comput. **27**(4), 1016–1035 (1998)

13. P.K. Agarwal, B. Aronov, T.M. Chan, M. Sharir, On levels in arrangements of lines, segments, planes, and triangles. Discrete Comput. Geom. **19**, 315–331 (1998)
14. B. Aronov, M. de Berg, E. Ezra, M. Sharir, Improved bounds for the union of locally fat objects in the plane. SIAM J. Comput. **43**(2), 543–572 (2014)
15. R.P. Bambah, C.A. Rogers, Covering the plane with convex sets. J. Lond. Math. Soc. **1**(3), 304–314 (1952)
16. T.M. Chan, Random sampling, halfspace range reporting, and construction of ($\leq k$)-levels in three dimensions. SIAM J. Comput. **30**(2), 561–575 (2000)
17. T.M. Chan, Low-dimensional linear programming with violations. SIAM J. Comput. **34**(4), 879–893 (2005)
18. T.M. Chan, A dynamic data structure for 3-d convex hulls and 2-d nearest neighbor queries. J. Assoc. Comput. Mach. **57**(3), 1–15 (2010). Art. 16
19. T.M. Chan, K. Tsakalidis, Optimal deterministic algorithms for 2-d and 3-d shallow cuttings, in *Proceedings of the 31st International Annual Symposium on Computational Geometry (SoCG)* (2015), pp. 719–732
20. B. Chazelle, Cutting hyperplanes for divide-and-conquer. Discrete Comput. Geom. **9**(2), 145–158 (1993)
21. B. Chazelle, Cuttings (chapter 25), in *Handbook of Data Structures and Applications*, ed. by D.P. Mehta, S. Sahni (Chapman and Hall/CRC, Boca Raton, 2004)
22. B. Chazelle, J. Friedman, A deterministic view of random sampling and its use in geometry. Combinatorica **10**(3), 229–249 (1990)
23. B. Chazelle, H. Edelsbrunner, L.J. Guibas, M. Sharir, A singly-exponential stratification scheme for real semi-algebraic varieties and its applications. Theoret. Comput. Sci. **84**, 77–105 (1991). Also in *Proceedings of the 16th International Colloquium on Automata, Languages and Programming*, pp. 179–193
24. K.L. Clarkson, New applications of random sampling in computational geometry. Discrete Comput. Geom. **2**, 195–222 (1987)
25. K.L. Clarkson, P.W. Shor, Applications of random sampling in computational geometry, II. Discrete Comput. Geom. **4**, 387–421 (1989)
26. K.L. Clarkson, H. Edelsbrunner, L.J. Guibas, M. Sharir, E. Welzl, Combinatorial complexity bounds for arrangements of curves and spheres. Discrete Comput. Geom. **5**, 99–160 (1990)
27. E. Ezra, D. Halperin, M. Sharir, Speeding up the incremental construction of the union of geometric objects in practice. Comput. Geom. Theory Appl. **27**(1), 63–85 (2004)
28. G.N. Frederickson, Fast algorithms for shortest paths in planar graphs, with applications. SIAM J. Comput. **16**(6), 1004–1022 (1987)
29. S. Har-Peled, Constructing planar cuttings in theory and practice. SIAM J. Comput. **29**(6), 2016–2039 (2000)
30. S. Har-Peled, M. Sharir, Relative (p, ε)-approximations in geometry. Discrete Comput. Geom. **45**(3), 462–496 (2011)
31. S. Har-Peled, H. Kaplan, M. Sharir, Approximating the k-level in three-dimensional plane arrangements, in *Proceedings of the 27th ACM-SIAM Symposium on Discrete Algorithms (SODA)* (2016), pp. 1193–1212
32. S. Har-Peled, H. Kaplan, M. Sharir, Approximating the k-level in three-dimensional plane arrangements. CoRR (2016). abs/1601.04755
33. K. Kedem, R. Livne, J. Pach, M. Sharir, On the union of Jordan regions and collision-free translational motion amidst polygonal obstacles. Discrete Comput. Geom. **1**, 59–71 (1986)
34. P.N. Klein, S. Mozes, C. Sommer, Structured recursive separator decompositions for planar graphs in linear time, in *Proceedings of the 45th Annual ACM Symposium on Theory Computing (STOC)* (2013), pp. 505–514
35. V. Koltun, Almost tight upper bounds for vertical decompositions in four dimensions. J. Assoc. Comput. Mach. **51**(5), 699–730 (2004)
36. V. Koltun, M. Sharir, Curve-sensitive cuttings. SIAM J. Comput. **34**(4), 863–878 (2005)
37. R.J. Lipton, R.E. Tarjan, A separator theorem for planar graphs. SIAM J. Appl. Math. **36**, 177–189 (1979)

38. J. Matoušek, Construction of ε-nets. Discrete Comput. Geom. **5**, 427–448 (1990)
39. J. Matoušek, Efficient partition trees. Discrete Comput. Geom. **8**, 315–334 (1992)
40. J. Matoušek, Range searching with efficient hierarchical cutting. Discrete Comput. Geom. **10**, 157–182 (1992)
41. J. Matoušek, Reporting points in halfspaces. Comput. Geom. Theory Appl. **2**(3), 169–186 (1992)
42. J. Matoušek, On constants for cuttings in the plane. Discrete Comput. Geom. **20**, 427–448 (1998)
43. J. Matoušek, N. Miller, J. Pach, M. Sharir, S. Sifrony, E. Welzl, Fat triangles determine linearly many holes, in *Proceedings of the 32nd Annual IEEE Symposium on Foundations of Computer Science (FOCS)* (1991), pp. 49–58
44. G.L. Miller, Finding small simple cycle separators for 2-connected planar graphs. J. Comput. Syst. Sci. **32**(3), 265–279 (1986)
45. K. Mulmuley, An efficient algorithm for hidden surface removal, II. J. Comput. Syst. Sci. **49**, 427–453 (1994)
46. N.H. Mustafa, R. Raman, S. Ray, Settling the APX-hardness status for geometric set cover, in *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science (FOCS)* (2014), pp. 541–550
47. J. Pach, P.K. Agarwal, *Combinatorial Geometry* (Wiley, New York, 1995)
48. E.A. Ramos, On range reporting, ray shooting and k-level construction, in *Proceedings of the 15th Annual Symposium on Computational Geometry (SoCG)* (ACM, 1999), pp. 390–399
49. R. Seidel, A simple and fast incremental randomized algorithm for computing trapezoidal decompositions and for triangulating polygons. Comput. Geom. Theory Appl. **1**, 51–64 (1991)
50. M. Sharir, P.K. Agarwal, *Davenport-Schinzel Sequences and Their Geometric Applications* (Cambridge University Press, New York, 1995)

Schrijver Graphs and Projective Quadrangulations

Tomáš Kaiser and Matěj Stehlík

Abstract In a recent paper, Kaiser and Stehlík (J Combin Theory Ser B 113:1–17, 2015) have extended the concept of quadrangulation of a surface to higher dimension, and showed that every quadrangulation of the n-dimensional projective space $\mathbb{P}^n$ is at least $(n+2)$-chromatic, unless it is bipartite. They conjectured that for any integers $k \geq 1$ and $n \geq 2k+1$, the Schrijver graph $SG(n,k)$ contains a spanning subgraph which is a non-bipartite quadrangulation of $\mathbb{P}^{n-2k}$. The purpose of this paper is to prove the conjecture.

1 Introduction

Given any integers $k \geq 1$ and $n \geq 2k$, the *Kneser graph* $KG(n,k)$ is the graph whose vertex set consists of all k-subsets of $[n] = \{1,\dots,n\}$, and with edges joining pairs of disjoint subsets. It was conjectured by Kneser [5], and proved by Lovász [6] in 1978, that the chromatic number of $KG(n,k)$ is $n-2k+2$.

Schrijver [10] found a vertex-critical subgraph $SG(n,k)$ of $KG(n,k)$ whose chromatic number is also $n-2k+2$. (Recall that a graph is *vertex-critical* if the deletion of any vertex decreases the chromatic number.) Let C_n be the cycle with vertices $1,\dots,n$ (in this order). The vertices of the *Schrijver graph* $SG(n,k)$ are all

Supported by project GA14-19503S of the Czech Science Foundation.

Partially supported by ANR project Stint(ANR-13-BS02-0007) and by LabEx PERSYVAL-Lab (ANR-11-LABX-0025).

T. Kaiser (✉)
Department of Mathematics, Institute for Theoretical Computer Science (CE-ITI), and European Centre of Excellence NTIS (New Technologies for the Information Society), University of West Bohemia, Univerzitní 8, 306 14 Pilsen, Czech Republic
e-mail: kaisert@kma.zcu.cz

M. Stehlík
Laboratoire G-SCOP, Université Grenoble Alpes, Grenoble, France
e-mail: matej.stehlik@grenoble-inp.fr

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_20

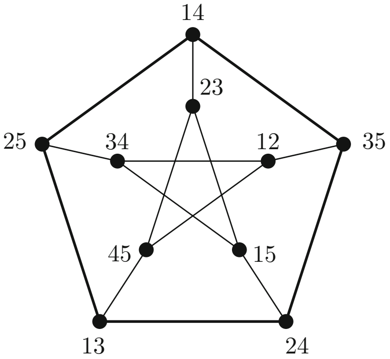


Fig. 1 The Kneser graph $KG(5,2)$ with its induced subgraph $SG(5,2)$ drawn by *thick lines*. Vertex labels such as $\{1,3\}$ are abbreviated to 13

the independent sets of C_n of size k, and the edges of $SG(n,k)$ join disjoint subsets. Thus, $SG(n,k)$ is an induced subgraph of $KG(n,k)$. See Fig. 1 for an example.

In [4], a *quadrangulation* of a space triangulated by a (generalised) simplicial complex K is defined as a spanning subgraph G of the 1-skeleton $\mathsf{K}^{(1)}$ such that the induced subgraph of G on the vertex set of any maximal simplex of K is complete bipartite with at least one edge.

Particular attention was given in [4] to quadrangulations of projective spaces, and it was shown that if G is a non-bipartite quadrangulation of the (real) projective space $\mathbb{P}^n$, then the chromatic number of G is at least $n+2$. By constructing suitable projective quadrangulations of $\mathbb{P}^n$ homomorphic to Schrijver graphs, an alternative proof of Schrijver's result was obtained.

The purpose of this paper is to prove Conjecture 7.1 from [4] by establishing the following result:

Theorem 1 *For any $k \geq 1$ and $n > 2k$, the graph $SG(n,k)$ contains a spanning subgraph $QG(n,k)$ that embeds in $\mathbb{P}^{n-2k}$ as a non-bipartite quadrangulation. In particular, $\chi(QG(n,k)) = n-2k+2$.*

To prove Theorem 1, we need to construct a suitable triangulation of the sphere S^{n-2k}. We first review some topological preliminaries (Sect. 2) and explore combinatorial relations among the vertices of Schrijver graphs (Sect. 3).

In Sect. 4, the required properties of the sought triangulation of S^{n-2k} are formulated in Theorem 10, which is then proved by giving an explicit recursive construction. Theorem 1 is derived at the end of Sect. 4.

In Sect. 5, two open problems are given to conclude the paper. In particular, we conjecture that the graph $QG(n,k)$ of Theorem 1 is edge-critical.

2 Topological Preliminaries

In this section, we recall the necessary topological concepts. For a background on topological methods in combinatorics, we refer the reader to Matoušek [7]. For an introduction to algebraic topology, consult Hatcher [3] or Munkres [8].

A *simplicial complex* C with vertex set V is a hereditary set system on V; the elements of this set system are the *faces* of C. A *geometric simplicial complex* K in $\mathbb{R}^d$ is obtained if we associate each vertex in V with a point in $\mathbb{R}^d$ in such a way that

1. the set of points P_σ associated with each face σ is in convex position, and
2. for distinct faces σ and τ, the relative interiors of the convex hulls of P_σ and P_τ are disjoint.

The convex hulls of the sets P_σ, where σ is a face of the underlying simplicial complex C, will be referred to as the *faces* of K. Since we will be dealing exclusively with geometric simplicial complexes in this paper, we will often drop the adjectives 'geometric' and 'simplicial'. Throughout the paper, we use sans-serif symbols such as C or QK to denote complexes.

A face such as $\{a, b, c\}$ is also written as abc. Two vertices v, w of K are *adjacent* if vw is a face of K. The *dimension* of a face σ is $|\sigma| - 1$. Faces of dimension one are called *edges*. The vertex set of a geometric simplicial complex K will be referred to as $V(\mathsf{K})$.

The *space* $\|\mathsf{K}\|$ of a geometric simplicial complex K in $\mathbb{R}^d$ is the subspace of $\mathbb{R}^d$ obtained as the union of all faces of K. If a space $X \subseteq \mathbb{R}^d$ is homeomorphic to $\|\mathsf{K}\|$, we say that K *triangulates* X.

The *induced subcomplex* of K on a set $X \subseteq V(\mathsf{K})$, denoted by $\mathsf{K}[X]$, has vertex set X and its faces are all the faces of K contained in X.

The *closed star* of a set X of vertices in K is the subcomplex of K consisting of all the faces of K containing a vertex of X, together with their subfaces. If X is contained in a subcomplex L of K, then the closed star of X in L is the intersection of the closed star of X (in K) with L. The closed star of a vertex v of K is defined as the closed star of $\{v\}$.

The *link* of a face σ of K, denoted by $\mathrm{lk}(\sigma)$, is the subcomplex consisting of all faces τ such that $\sigma \cup \tau$ is a face of K and $\sigma \cap \tau = \emptyset$.

A 2-*coloured complex* K in $\mathbb{R}^d$ is a geometric simplicial complex in $\mathbb{R}^d$, with each vertex coloured black or white. For any point $p \in \mathbb{R}^d$, its *antipode* is the point $-p$. The complex K is *antisymmetric* if the following holds:

- for every vertex v of K, the antipode $-v$ of v is also a vertex of K, and the colours of v and $-v$ are different,
- for each face σ of K, the antipodes of the vertices of σ form a face of K.

Suppose that a 2-coloured complex K triangulates the ball B^d. The *boundary* of K is the subcomplex triangulating the boundary sphere $S^{d-1} = \partial B^d$. We will say that K is *boundary-antisymmetric* if its boundary is antisymmetric.

We recall several notions related to homotopy (cf. [3, Chapter 0]). Let X and Y be topological spaces. Continuous maps $f, g : X \to Y$ are *homotopic* if there exists a continuous map $H : X \times [0, 1] \to Y$ such that $H(x, 0) = f(x)$ and $H(x, 1) = g(x)$ for all $x \in X$. A *homotopy equivalence* between spaces X and Y is a continuous map $f : X \to Y$ such that there is a continuous map $g : Y \to X$ with the property that each of $f \circ g$ and $g \circ f$ is homotopic to an identity map. Homotopy equivalent spaces are also said to have the same *homotopy type*.

Closely related to homotopy equivalence is the notion of deformation retraction. Given a subspace A of a space X, a family of continuous maps $f_t : X \to X$ (where $t \in [0, 1]$) is a *deformation retraction* of X onto A if f_0 is the identity, so is the restriction of each f_t to A, the image of f_1 is A, and the family is continuous when viewed as a map from $X \times [0, 1] \to X$. If such a deformation retraction exists, A is said to be a *deformation retract of* X (and X is said to *deformation retract to* A). It is easy to see from the definitions that the space X is homotopy equivalent to any deformation retract of X.

A space X is *contractible* if the identity map on X is *nullhomotopic* (homotopic to a constant map), which is somewhat weaker than the property of having a deformation retraction to a single point.

Next, let K be a 2-coloured complex whose space is a deformation retract of the thickened sphere $S^d \times I$ in $\mathbb{R}^{d+1}$, where $d \geq 1$, S^d is the unit d-sphere and I is a short interval in $\mathbb{R}$. Thus, we can define the *interior* of K as the bounded component of $\mathbb{R}^{d+1} \setminus \|\mathsf{K}\|$, and similarly for the *exterior* of K. Note that the origin of $\mathbb{R}^{d+1}$ is contained in the interior. We define the *interior boundary* of K, $\mathsf{IB}(\mathsf{K})$, as the subcomplex of K consisting of all the faces of K contained in the closure of the interior of K. The *exterior boundary* $\mathsf{EB}(\mathsf{K})$ is defined analogously. Note that $\mathsf{IB}(\mathsf{K})$ and $\mathsf{EB}(\mathsf{K})$ need not be disjoint.

In the above setting, we will utilise the operation of *adding the cone* over a subcomplex S of $\mathsf{IB}(\mathsf{K})$. We add a vertex v_{S} and all faces $\sigma \cup \{v_{\mathsf{S}}\}$, where σ is a face of S. The vertex v_{S} is the *apex* of the cone. By placing v_{S} suitably in the interior of K and deforming K slightly if necessary, we obtain a realisation of the resulting complex K' as a geometric complex. While this operation may change the homotopy type of the complex, we will always use it in cases where K' is again a deformation retract of the thickened sphere. In addition, the colour of v_{S} will always be specified. The *cone* over S is the complex consisting of all the added faces (including $\{v_{\mathsf{S}}\}$) and all the faces of S. It is well known that the space of this complex is contractible.

A vertex z of $\mathsf{IB}(\mathsf{K})$ is an *inside* vertex of S if the closed star of z (within $\mathsf{IB}(\mathsf{K})$) is contained in S.

Observation 2 *With* K *as above, let* K' *be obtained by adding the cone* v_{S} *over a subcomplex* S *of* $\mathsf{IB}(\mathsf{K})$. *Then* $\mathsf{IB}(\mathsf{K}')$ *contains* v_{S} *and does not contain any inside vertex of* S. *In fact, a face* σ *of* $\mathsf{IB}(\mathsf{K})$ *is a face of* $\mathsf{IB}(\mathsf{K}')$ *if and only if* $\sigma \setminus \{v_{\mathsf{S}}\}$ *is a (possibly empty) face of* $\mathsf{IB}(\mathsf{K})$ *containing no inside vertex of* S.

Let K be a 2-coloured complex and u, v two adjacent vertices of K of the same colour. The *contraction* of the edge uv is the operation replacing each incident face σ of K with $\sigma \cup \{w\} \setminus \{u, v\}$, where w is a new vertex (assigned the colour of u and v).

Geometrically, it corresponds to shrinking the segment uv to a point. By definition, the operation does not introduce multiple copies of any face. Although contraction may in general change the topological properties of the complex, we will only apply it in situations where the contraction yields a complex of the same homotopy type. The key property is the *link condition* for the edge uv (cf. [1, 9]):

$$\mathrm{lk}(u) \cap \mathrm{lk}(v) = \mathrm{lk}(uv). \tag{1}$$

Let K and L be 2-coloured complexes. A mapping $f : V(\mathsf{K}) \to V(\mathsf{L})$ is a *homomorphism* (of 2-coloured complexes) from K to L if f preserves vertex colours and for any face σ of K, its image $f[\sigma]$ is a face of L. (We stress that $f[\sigma]$ is a set, without repeated elements.)

A homomorphism f from K to L is an *isomorphism* if f is a bijection and f^{-1} is a homomorphism.

For an antisymmetric 2-coloured complex K, we define its *associated graph* $G(\mathsf{K})$ as the graph with vertex set $V(\mathsf{K})$ and with the edge set consisting of all edges of K with one end black and the other white.

3 Combinatorial Preliminaries

Before we present the construction proving Theorem 1, we need to do some preparatory work. In this section, we introduce some terminology and notation that is useful for the classification of the vertices of the Schrijver graph $SG(n,k)$.

Let $k \geq 1$ and $n \geq 2k+1$. Recall from Sect. 1 that C_n denotes the n-cycle on the vertex set $[n] = \{1,\dots,n\}$. Let $V(n,k)$ be the set of all subsets of $[n]$ of size k that are independent sets in C_n. (Thus, $V(n,k)$ is the vertex set of $SG(n,k)$.) Note that $V(n-1,k)$ is a subset of $V(n,k)$.

Addition and subtraction on $[n]$ are defined 'with wrap-around': for instance, if $i,j \in [n]$ and $(i-1)+(j-1) \equiv \ell - 1 \pmod{n}$, where $\ell \in [n]$, then $i+j$ is defined as ℓ. The *core* of a set $A \in V(n,k)$ is the set

$$\mathrm{core}(A) = \begin{cases} A \setminus \{1\} & \text{if } 1 \in A, \\ A \setminus \{\max(A)\} & \text{otherwise.} \end{cases}$$

Thus, $\mathrm{core}(\{1,3,5\}) = \{3,5\}$, while $\mathrm{core}(\{2,4,6\}) = \{2,4\}$.

Observation 3 *For $A \in V(n,k)$,*

$$\mathrm{core}(A) \cap \{1,n\} = \emptyset.$$

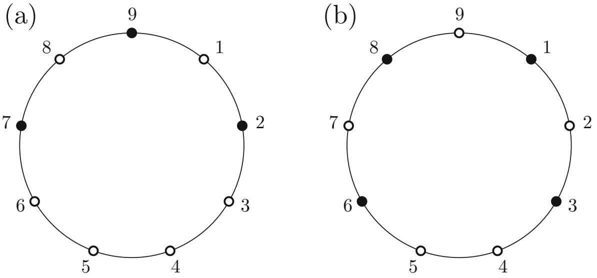


Fig. 2 Examples of the sets $\Lambda_{n,i}$ for $n = 9$, pictured as subsets of $V(C_9)$. *Black dots* represent elements included in the subset, *white dots* show the other vertices of C_9. (**a**) $\Lambda_{9,3}$. (**b**) $\Lambda_{9,4}$

Let $0 \le i \le n/2$. We define the set $\Lambda_{n,i} \subseteq [n]$ as follows:

$$\Lambda_{n,i} = \begin{cases} \{n-i+1, n-i+3, \dots, n, 2, 4, \dots, i-1\} & \text{if } i \text{ is odd,} \\ \{n-i+1, n-i+3, \dots, n-1, 1, 3, \dots, i-1\} & \text{if } i \text{ is even.} \end{cases}$$

See Fig. 2 for examples. Note that for each i, $\Lambda_{n,i} \in V(n,i)$. For small i, the sets $\Lambda_{n,i}$ are given in the following table:

$\Lambda_{n,0}$	$\Lambda_{n,1}$	$\Lambda_{n,2}$	$\Lambda_{n,3}$	$\Lambda_{n,4}$
$\emptyset$	$\{n\}$	$\{1, n-1\}$	$\{2, n-2, n\}$	$\{1, 3, n-3, n-1\}$

The *n-level* of a set $A \in V(n,k)$, $\ell^n(A)$, is the maximum i such that $\Lambda_{n,i} \subseteq A$. Note that $0 \le \ell^n(A) \le k$. For $0 \le i \le k$, we define

$$V_i(n,k) = \{A \in V(n,k) : \ell^n(A) = i\}.$$

Furthermore, we let $V_+(n,k)$ be the union of all $V_i(n,k)$ with $i \ge 1$.

Lemma 4 *We have*

$$V_0(n,k) = V(n-1,k).$$

Proof We need to show that for any set $A \in V(n,k)$, we have $\ell^n(A) = 0$ if and only if $A \in V(n-1,k)$. By definition, $\ell^n(A) = 0$ if and only if A contains neither $\{n\}$ nor $\{1, n-1\}$ as a subset. In turn, this holds if and only if $A \in V(n-1,k)$. □

Let $B \in V(n-1,k)$. We define the set $B\langle n\rangle \in V(n,k)$ by

$$B\langle n\rangle = \text{core}(B) \cup \{n\}.$$

By Observation 3, the operation is well defined. The following lemma will be useful:

Lemma 5 *For $2k+1 \leq i < m$ and $A \in V(i-1,k)$, $(A\langle i\rangle)\langle m\rangle = A\langle m\rangle$.*

Proof By definition, $A\langle i\rangle = \mathrm{core}(A) \cup \{i\}$. Since $1 \notin \mathrm{core}(A)$, $\mathrm{core}(A\langle i\rangle) = \mathrm{core}(A)$. Thus, $(A\langle i\rangle)\langle m\rangle = A\langle m\rangle$. □

For a set $A \in V(n,k)$ such that $1 \notin A$, we define $A - 1$ as the set obtained by subtracting 1 from each element of A (and similarly for $A+1$ when $n \notin A$, or when the result is interpreted in $V(n+1,k)$).

Let us define a mapping f from $V(n,k)$ to $V(n-2,k-1)$, and a mapping g_n in the inverse direction. Let $X \in V(n,k)$ and $Y \in V(n-2,k-1)$. The mappings are as follows:

$$f(X) = \mathrm{core}(X) - 1,$$

$$g_n(Y) = \begin{cases} (Y+1) \cup \{1\} & \text{if } n-2 \in Y, \\ (Y+1) \cup \{n\} & \text{otherwise.} \end{cases}$$

Lemma 6 *The restriction of f to $V_+(n,k)$ is a bijection*

$$f : V_+(n,k) \to V(n-2,k-1),$$

and g_n is its inverse. Furthermore, f maps disjoint pairs of sets to disjoint pairs.

Proof The first assertion follows from the fact that the image of g_n is contained in $V_+(n,k)$, and from the easily verified equalities

$$f(g_n(Y)) = Y \qquad \text{and} \qquad g_n(f(X)) = X$$

for $X \in V_+(n,k)$, $Y \in V(n-2,k-1)$.

The assertion that the images of disjoint sets under f are disjoint follows directly from the definition of f. □

Corollary 7 *All the sets in $V_+(n,k)$ have distinct cores.*

Define an equivalence $\approx$ on $V(n,k)$ by putting $A \approx B$ if $\mathrm{core}(A) = \mathrm{core}(B)$. Corollary 7 can be strengthened as follows:

Lemma 8 *If $A \in V(n,k)$ and $B \in V_+(n,k)$ are distinct sets such that $A \approx B$, then $A \in V_0(n,k)$ and $B \in V_1(n,k)$.*

Proof By Corollary 7, $A \in V_0(n,k)$. Suppose that $\ell^n(B) \geq 2$. We have either $\{2, n-2, n\} \subseteq B$ or $\{1, n-1\} \subseteq B$. In the former case, $\{2, n-2\} \subseteq \mathrm{core}(B) = \mathrm{core}(A)$, which is impossible as $A \in V_0(n,k)$ (so $n \notin A$). In the latter case, $n-1 \in \mathrm{core}(B) = \mathrm{core}(A)$ and hence necessarily $1 \in A$, which would imply that $\ell^n(A) \geq 2$, a contradiction. We have shown that $A \in V_0(n,k)$ and $B \in V_1(n,k)$. □

Observation 9 *For any set $B \in V(n-1,k)$, we have $f(B\langle n\rangle) = f(B)$. Thus, the appropriate restriction of f is a bijection*

$$\{B\langle n\rangle : B \in V_{+}(n-1,k)\} \to V(n-3,k-1).$$

Finally, we define a set $A \in V(n,k)$ to be *singular* if $A \in V(2k,k)$. Thus, the singular sets in $V(7,3)$ are $\{1,3,5\}$ and $\{2,4,6\}$.

4 Constructing the Embedding

In this section, we shall construct the antisymmetric 2-coloured complex $\mathsf{QK}(n,k)$ in $\mathbb{R}^{n-2k+1}$ triangulating the sphere S^{n-2k}. The vertices will be coloured black and white; both the black vertices and the white vertices will be labelled bijectively with elements of $V(n,k)$. We will identify each vertex with its label and speak, for instance, of the black copy of $\{1,3,5\}$ or the white copy of $\{2,6,8\}$. For a set $A \in V(n,k)$, its black copy will be denoted by $A^{\bullet}$ and its white copy by A°. (In all the complexes we construct, each label will appear at most once at a black vertex and at most once at a white vertex.)

We extend some of the combinatorial notions defined in Sect. 3 to vertices. Thus, we say that a vertex $A^{\bullet}$ is *singular* if A is singular (and similarly for A°). Likewise, the *core* of $A^{\bullet}$ is $\mathrm{core}(A)$.

Theorem 10 *For any $k \geq 1$ and $n \geq 2k+1$, there is a 2-coloured geometric complex $\mathsf{QK}(n,k)$ in $\mathbb{R}^{n-2k+1}$ with the following properties:*

(i) *$\mathsf{QK}(n,k)$ is an antisymmetric triangulation of the sphere S^{n-2k} such that no face contains a pair of antipodal vertices.*
(ii) *$\mathsf{QK}(n,k)$ contains no monochromatic maximal faces.*
(iii) *The graph obtained from the associated graph of $\mathsf{QK}(n,k)$ by identifying each pair of antipodal vertices is a spanning subgraph of $SG(n,k)$.*
(iv) *For $n > 2k+1$, $\mathsf{QK}(n,k)$ contains $\mathsf{QK}(n-1,k)$ as an antisymmetric subcomplex.*

Let us embark on the construction of $\mathsf{QK}(n,k)$ which eventually proves Theorem 10. We start by introducing some more notation. Let L be a 2-coloured complex whose vertices are labelled with elements of $V(n,k)$ (which will be the case most of the time). For a vertex $A^{\bullet}$ of L, we define $[A^{\bullet}]$ as the set of all black vertices $B^{\bullet}$ of L such that $A \approx B$. The *region* of A in L, denoted by $\mathsf{Reg}(A^{\bullet},\mathsf{L})$, is defined as the closed star of $[A^{\bullet}]$. The region of a white vertex is defined in an analogous way.

In the construction, we will ensure that the following (more technical) conditions hold in addition to those in Theorem 10:

(P1) If $A^{\bullet}B^{\circ}$ is a face of $\mathsf{QK}(n,k)$, then $|\ell^n(A)-\ell^n(B)| \leq 1$, and $\ell^n(A)=\ell^n(B)$ only if $\ell^n(A)=\ell^n(B)=0$.

(P2) For $k \geq 2$, and $A \in V(n,k)$, every nonsingular vertex of $\mathsf{QK}(n,k)$ belongs to a maximum face of $\mathsf{QK}(n,k)$ containing no black vertex $B^{\bullet}$ with $\mathrm{core}(B) = \mathrm{core}(C)$. An analogous condition holds with the colours inverted.

(P3) For any vertex $A^{\bullet}$ of $\mathsf{QK}(n,k)$, the induced subcomplex of $\mathsf{QK}(n,k)$ on $[A^{\bullet}]$ is either a face, or (in the case that $[A^{\bullet}]$ contains a singular vertex) the join of a face with a subcomplex consisting of two points.

The definition of $\mathsf{QK}(n,k)$ is straightforward in case that $n = 2k+1$. For $j \in [2k+1]$, let

$$I_j = \{j, j+2, \ldots, j+2k-2\} \in V(2k+1,k).$$

The complex $\mathsf{QK}(2k+1,k)$ is 1-dimensional, so we can describe it as a graph: it is the cycle of length $2(2k+1)$ with vertices

$$I_1^{\bullet}, I_2^{\circ}, I_3^{\bullet}, \ldots, I_{2k+1}^{\bullet}, I_1^{\circ}, I_2^{\bullet}, \ldots, I_{2k+1}^{\circ}, I_1^{\bullet}$$

in this order. See Fig. 3 for an illustration.

Suppose thus that $n > 2k+1$ and that $\mathsf{QK}(n-1,k)$ has already been constructed. Recall that $\mathsf{QK}(n-1,k)$ is an antisymmetric triangulation of S^{n-2k-1} in $\mathbb{R}^{n-2k}$. A quick summary of the construction of $\mathsf{QK}(n,k)$ (illustrated in Fig. 4) is as follows:

- we extend $\mathsf{QK}(n-1,k)$ to a 2-coloured complex $\mathsf{QR}^{\bullet}(n,k)$ triangulating a 'partially thickened' sphere S^{n-2k-1} if $k \geq 2$,

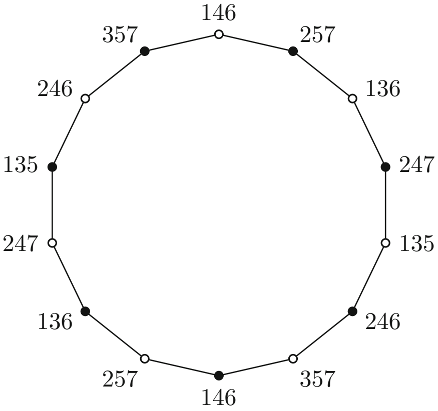


Fig. 3 The complex $\mathsf{QK}(7,3)$. Set brackets are omitted in vertex labels such as $\{1,3,5\}$

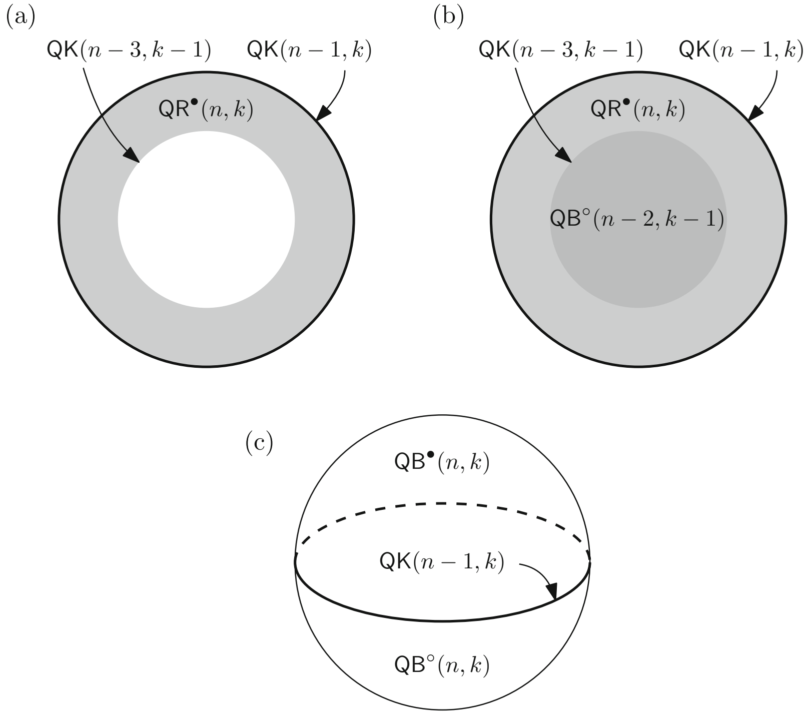


Fig. 4 A schematic illustration of the construction of $\mathsf{QK}(n,k)$. (**a**) $\mathsf{QR}^\bullet(n,k)$ (*light gray*) has exterior boundary $\mathsf{QK}(n-1,k)$ and interior boundary $\mathsf{QK}(n-3,k-1)$. (**b**) $\mathsf{QB}^\bullet(n,k)$ triangulates B^{n-2k}; it can be decomposed into $\mathsf{QR}^\bullet(n,k)$ (outer layer) and $\mathsf{QB}^\circ(n-2,k-1)$ (filling, *dark gray*). (**c**) $\mathsf{QK}(n,k)$ triangulates S^{n-2k}; it is obtained by pasting $\mathsf{QB}^\bullet(n,k)$ and $\mathsf{QB}^\circ(n,k)$ together along their common boundary $\mathsf{QK}(n-1,k)$

- we fill in the interior of $\mathsf{QR}^\bullet(n,k)$ using a complex $\mathsf{QB}^\circ(n-2,k-1)$ (constructed at an earlier stage of this recursive procedure) to obtain a 2-coloured boundary-antisymmetric triangulation $\mathsf{QB}^\bullet(n,k)$ of the $(n-2k)$-ball B^{n-2k},
- we obtain the complex $\mathsf{QB}^\circ(n,k)$ in an analogous way, inverting the colours,
- we form an antisymmetric 2-coloured triangulation $\mathsf{QK}(n,k)$ of S^{n-2k} by pasting $\mathsf{QB}^\bullet(n,k)$ and $\mathsf{QB}^\circ(n,k)$ together.

We remark that the letter B in $\mathsf{QB}^\bullet(n,k)$ stands for ‘ball’, while the R in $\mathsf{QR}^\bullet(n,k)$ was chosen to represent ‘rind’, the outer layer of $\mathsf{QB}^\bullet(n,k)$.

As the first step of the construction, we now extend $\mathsf{QK}(n-1,k)$ to the 2-coloured complex $\mathsf{QR}^\bullet(n,k)$ by adding cones over certain subcomplexes and contracting some of the edges. For $k \geq 2$, the exterior boundary $\mathsf{QK}(n-1,k)$ of $\mathsf{QR}^\bullet(n,k)$ as well as its interior boundary will be deformation retracts of $\mathsf{QR}^\bullet(n,k)$. The interior boundary of $\mathsf{QR}^\bullet(n,k)$ will be shown to be isomorphic (as a complex) to $\mathsf{QK}(n-3,k-1)$, enabling us to fill in the interior by recursion.

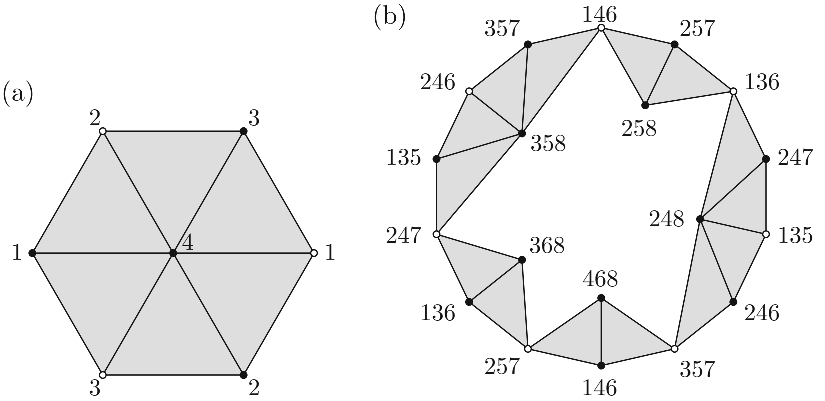


Fig. 5 Cases $k = 1$ and $n = 2k + 2$ of the construction of $\mathsf{QR}^\bullet(n, k)$. Set brackets in vertex labels are omitted. (**a**) $\mathsf{QR}^\bullet(4, 1)$. (**b**) $\mathsf{QR}^\bullet(8, 3)$

In the special case $k = 1$, the construction is particularly simple (see Fig. 5a): $\mathsf{QR}^\bullet(n, 1)$ is obtained just by adding the cone over $\mathsf{QK}(n - 1, 1)$, with the apex coloured black and labelled n.

To construct $\mathsf{QR}^\bullet(n, k)$ from $\mathsf{QK}(n - 1, k)$ for $k > 1$, we proceed as follows (we urge the reader to consult the illustration in Figs. 5b and 6):

(B1) For each class of the equivalence $\approx$ on $V(n-1, k)$, we choose a representative A, add the cone over $\mathsf{Reg}(A^\bullet, \mathsf{QK}(n - 1, k))$, colour the apex black and label it by $A\langle n\rangle$. Let K_1 denote the interior boundary of the resulting complex after all the classes of $\approx$ are processed.

(B2) For each class of $\approx$ on $V_0(n - 1, k)$, we choose a nonsingular representative B, we add the cone over $\mathsf{Reg}(B^\circ, \mathsf{K}_1)$, colour the apex white and label it by B_*. We contract the edge $B^\circ_* B\langle n - 1\rangle^\circ$. The resulting vertex retains the label $B\langle n - 1\rangle^\circ$. (This step is justified by Observation 11.)

An application of the above rules to a particular equivalence class is referred to as *processing* that class.

By switching colours in the above description (for example, adding the cone over $\mathsf{Reg}(A^\circ, \mathsf{QK}(n - 1, k))$ in step (B1)), we obtain the complex $\mathsf{QR}^\circ(n, k)$.

The following observation provides a justification for step (B2).

Observation 11 *Each class of $\approx$ on $V_0(n - 1, k)$ contains a nonsingular set. Moreover, the following hold for a nonsingular $B \in V_0(n - 1, k)$:*

(i) The vertex B° is contained in the complex K_1 defined in step (B1) above, so $\mathsf{Reg}(B^\circ, \mathsf{K}_1)$ is nonempty.

(ii) In step (B2), B°_* is adjacent to $B\langle n - 1\rangle^\circ$ and the edge joining them satisfies the link condition (1).

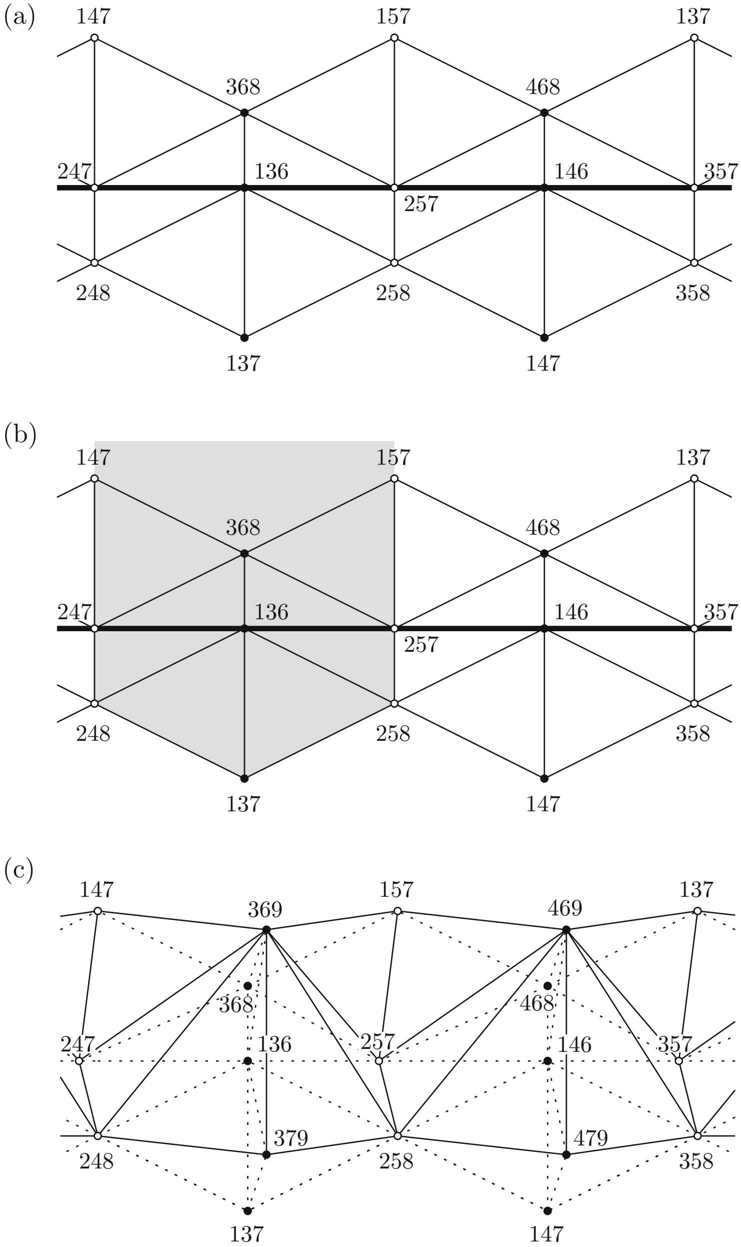


Fig. 6 The construction of $\mathsf{QR}^\bullet(9,3)$ from $\mathsf{QK}(8,3)$. (**a**) A portion of the complex $\mathsf{QK}(8,3)$ which triangulates B^2. The equator $\mathsf{QK}(7,3)$ is shown by a *thick line*. Note that the vertex set of $\mathsf{QK}(7,3)$ equals $V_0(8,3)$. (**b**) Part of the subcomplex $\mathsf{Reg}(368^\bullet, \mathsf{QK}(8,3))$ in $\mathsf{QK}(8,3)$ (*gray*). (**c**) The result of step (B1). The complex $\mathsf{QK}(8,3)$ should be pictured in a base plane, and the added apex vertices (such as $369^\bullet$) above it. *Dotted* and *solid lines* represent visibility

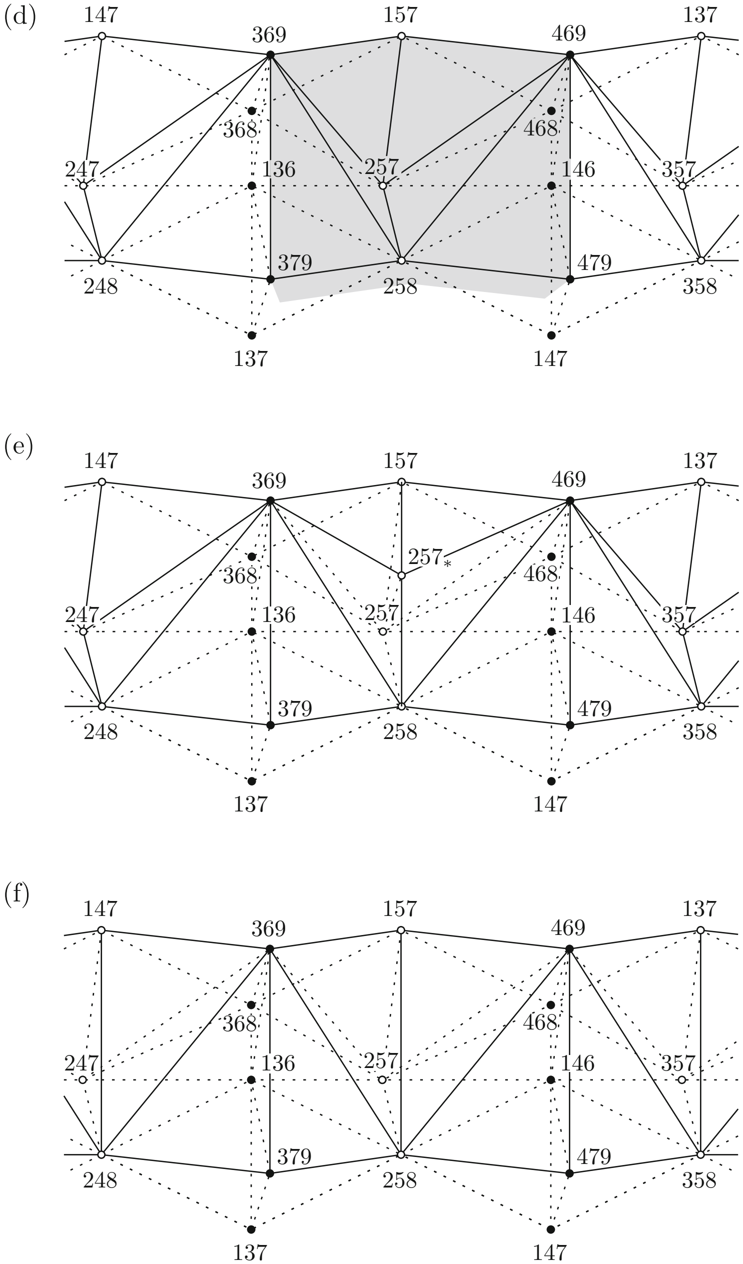


Fig. 6 The construction of $\mathsf{QR}^\bullet(9,3)$ from $\mathsf{QK}(8,3)$ (continued). (**d**) Part of the subcomplex $\mathsf{Reg}(257^\circ, \mathsf{K}_1)$ (*gray*) in the complex resulting from step (B1). (**e**) The complex obtained in step (B2) after adding the cone over $\mathsf{Reg}(257^\circ, \mathsf{K}_1)$ (assuming 257 is the first set whose equivalence class is processed in this step). (**f**) The result of step (B2)

Proof The first assertion follows from the fact that if A is a singular set in $V_0(n-1,k)$, then $A\langle 2k+1\rangle$ is nonsingular and $A \approx A\langle 2k+1\rangle$. Furthermore, $A\langle 2k+1\rangle \in V_0(n-1,k)$ since $n \geq 2k+2$.

(i) Since B is nonsingular, B° is not an inside vertex of $\mathsf{Reg}(A^\bullet, \mathsf{QK}(n-1,k))$ for any $A \in V(n-1,k)$ by property (P2). Observation 2 implies that B° is a vertex of K_1.

(ii) Similarly as in (i), $B\langle n-1\rangle$ is contained in K_1. Since $\mathrm{core}(B\langle n-1\rangle) = \mathrm{core}(B)$, $B\langle n-1\rangle$ is a vertex of $\mathsf{Reg}(B^\circ, \mathsf{K}_1)$. It follows that it is adjacent to B°_* after step (B2).

To verify the link condition, let $u = B\langle n-1\rangle$ and $v = B^\circ_*$. Since $\mathrm{lk}(uv) \subseteq \mathrm{lk}(u) \cap \mathrm{lk}(v)$, it suffices to establish the other inclusion. However, if σ is a face of $\mathrm{lk}(u)$, then the definition of the cone with apex v implies that σ is a face of $\mathrm{lk}(uv)$. □

Let us summarise the crucial topological properties of $\mathsf{QR}^\bullet(n,k)$:

Lemma 12 *For $k \geq 2$, the following hold:*

(i) The space of $\mathsf{QR}^\bullet(n,k)$ *is homotopy equivalent to* $\mathsf{QK}(n-1,k)$.
(ii) The space of the interior boundary of $\mathsf{QR}^\bullet(n,k)$ *is homeomorphic to the sphere* S^{n-2k-1}.

Proof (i) Property (P3) and the fact that $\mathsf{QK}(n-1,k)$ is a geometric complex implies that each complex $\mathsf{Reg}(A^\bullet, \mathsf{QK}(n-1,k))$ is contractible for each $A \in V(n-1,k)$. Furthermore, it is not hard to see that the latter property remains true if $\mathsf{QK}(n-1,k)$ is replaced by the interior boundaries of complexes obtained in subsequent applications of rule (B1). Thus, each of these applications of rule (B1) adds a cone over a contractible subcomplex of the interior boundary, which does not change the homotopy type.

The discussion of rule (B2) is similar; the fact that the edge contractions do not change the homotopy type follows, e.g., from [1, Theorem 2] together with the link condition verified in Observation 11(ii).

To prove (ii), we observe that as we apply rules (B1) and (B2), the effect on the interior boundary is equivalent to contracting the subcomplexes of condition (P3). The statement essentially follows by checking the link condition for these contractions and using [9, Theorem 4]; we omit the details. □

It is easy to describe the vertex set of $\mathsf{QR}^\bullet(n,k)$:

Observation 13 *The vertex set of* $\mathsf{QR}^\bullet(n,k)$ *is*

$$\bigcup_{A\in V(n-1,k)} \{A^\bullet, A^\circ\} \cup \{A\langle n\rangle^\bullet : A \in V_+(n-1,k)\}.$$

Proof The vertex set of $\mathsf{QK}(n-1,k)$ is

$$\bigcup_{A\in V(n-1,k)} \{A^\bullet, A^\circ\}.$$

Step (B1) adds one vertex per equivalence class of $\approx$, which accounts for the remaining vertices in the statement as each equivalence class intersects $V_+(n-1,k)$ in precisely one vertex (cf. Corollary 7). No new vertices are added in step (B2). □

We can now state the following lemma, which completely describes the interior boundary of $\mathsf{QR}^\bullet(n,k)$ and enables us to use a recursive construction.

Lemma 14 *Let* IB *be the interior boundary of* $\mathsf{QR}^\bullet(n,k)$*, where* $n \geq 2k+2$ *and* $k > 1$*. The following properties hold:*

(i) *The vertex set of* IB *is*

$$W = \bigcup_{A \in V_+(n-1,k)} \{A\langle n\rangle^\bullet, A^\circ\},$$

and IB *is the induced subcomplex of* $\mathsf{QR}^\bullet(n,k)$ *on* W*.*

(ii) IB *is isomorphic to* $\mathsf{QK}(n-3,k-1)$*, with the isomorphism determined by the mapping* $f : A \mapsto \mathrm{core}(A) - 1$ *and preserving the colours (where* A *is a vertex of* IB*).*

Proof (i) In view of Observation 13, we need to show that IB does not include any vertex $A^\bullet$ with $A \in V(n-1,k)$ nor any vertex A° with $A \in V_0(n-1,k)$, and includes all the other vertices of $\mathsf{QR}^\bullet(n,k)$.

Let $A \in V(n-1,k)$ and let L be any intermediate complex obtained in step (B1). Since $A^\bullet$ is an inside vertex of $\mathsf{Reg}(A^\bullet, \mathsf{L})$, Observation 2 implies that it is not contained in the interior boundary of L, and hence also not in IB.

An analogous argument shows that no vertex A° with $A \in V_0(n-1,k)$ is contained in IB. The contractions of the edges in step (B2) do not make a substantial difference, since their only effect on the interior boundary of the complex is to replace A°_* with $A\langle n-1\rangle^\circ$ in each face containing A°_*.

We prove that IB includes the vertices $A\langle n\rangle^\bullet$ and A°, where $A \in V_+(n-1,k)$. Consider a vertex A°. Condition (P2) ensures that A° is not an inside vertex of any complex $\mathsf{Reg}(B^\bullet, \mathsf{QK}(n-1,k))$, so by Observation 2, A° is a vertex of the interior boundary of the complex obtained by adding the cone over $\mathsf{Reg}(B^\bullet, \mathsf{QK}(n-1,k))$. Furthermore, it is not hard to prove by induction that (P2) is preserved when we replace $\mathsf{QK}(n-1,k)$ with the interior boundary of a complex obtained in subsequent applications of rule (B1). In this way, we show that A° is contained in the complex K_1. The argument for rule (B2) is similar, and so is the proof for the vertices of the type $A\langle n\rangle^\bullet$. Summing up, these arguments show that the vertex set of IB is W.

The proof that IB is the induced subcomplex on W is based on property (P3) and induction; we leave it to the reader.

(ii) Assume first that $n = 2k+2$. For $j \in [2k-1]$, let I'_j be the analogue of the independent set I_j used in the definition of $\mathsf{QK}(2k+1,k)$, but defined in C_{2k-1} and of size $k-1$. Thus, $I'_j = \{j, j+2, \ldots, j+2k-4\}$ with arithmetic performed in $[2k-1]$. The assertion follows from the following property of the mapping f, valid for any $j \in [2k + 1]$:

$$f(I_j) = \begin{cases} I'_2 & \text{if } j = 1, \\ I'_1 & \text{if } j = 2, \\ I'_{j-1} & \text{if } 3 \leq j \leq 2k, \\ I'_1 & \text{if } j = 2k + 1. \end{cases}$$

Let us now assume that $n > 2k + 2$. Consider the mapping h from the vertex set of $\mathsf{QK}(n - 1, k)$ to the vertex set of IB, defined as follows:

$$\begin{aligned} h(A^\bullet) &= A\langle n\rangle^\bullet \quad \text{for } A \in V(n-1,k), \\ h(B^\circ) &= B\langle n-1\rangle^\circ \quad \text{for } B \in V_0(n-1,k), \\ h(B^\circ) &= B^\circ \quad \text{for } B \in V_+(n-1,k). \end{aligned}$$

We claim that h is a homomorphism of 2-coloured complexes from $\mathsf{QK}(n - 1, k)$ to IB. We need to show that the image of any face of $\mathsf{QK}(n - 1, k)$ under h is a face of IB.

Thus, let σ be a face of $\mathsf{QK}(n-1, k)$. List the black vertices of σ as $A_1^\bullet, \dots, A_t^\bullet$ in the order their equivalence classes (or rather, the equivalence classes of their labels) were processed in step (B1). (If two of them belong to the same class, we order them arbitrarily.)

When the equivalence class of A_1 is being processed, we add the cone over $\mathsf{Reg}(A_1^\bullet, \mathsf{L})$, where L is the interior boundary of the complex constructed until that point. By Observation 2, the interior boundary L' of the resulting complex will contain the face $\sigma \setminus \{A_1^\bullet\} \cup \{A_1\langle n\rangle^\bullet\}$. In a similar fashion, $A_2^\bullet$ will eventually be replaced by $A_2\langle n\rangle^\bullet$ etc., and in the end we obtain a face σ' of IB in which each vertex $A^\bullet$ of σ is replaced by the corresponding 'apex' vertex $A\langle n\rangle^\bullet$.

The procedure for the white vertices B° of σ' is similar: we replace each such vertex with $B \in V_0(n-1, k)$ by the vertex $B\langle n-1\rangle^\circ$ in one application of step (B2). It follows that h is a homomorphism as claimed.

Consider the exterior boundary $\mathsf{QK}(n - 1, k)$ of $\mathsf{QR}^\bullet(n, k)$. By steps (K1)–(K3) below, $\mathsf{QK}(n - 1, k)$ is obtained from $\mathsf{QB}^\bullet(n - 1, k)$ and $\mathsf{QB}^\circ(n - 1, k)$ by glueing them along their common boundary $\mathsf{QK}(n-2, k)$ (viewed as the equator of $\mathsf{QK}(n-1, k)$). Let X be the set of vertices of $\mathsf{QB}^\bullet(n - 1, k)$; it follows from the construction of $\mathsf{QB}^\bullet(n - 1, k)$ and Lemma 16 below that

$$X = \bigcup_{A \in V(n-2,k)} \{A^\bullet, A^\circ\} \cup \left\{ A^\bullet : A \in \bigcup_{i \geq 1} V_{2i-1}(n-1,k) \right\}$$
$$\cup \left\{ A^\circ : A \in \bigcup_{i \geq 1} V_{2i}(n-1,k) \right\}.$$

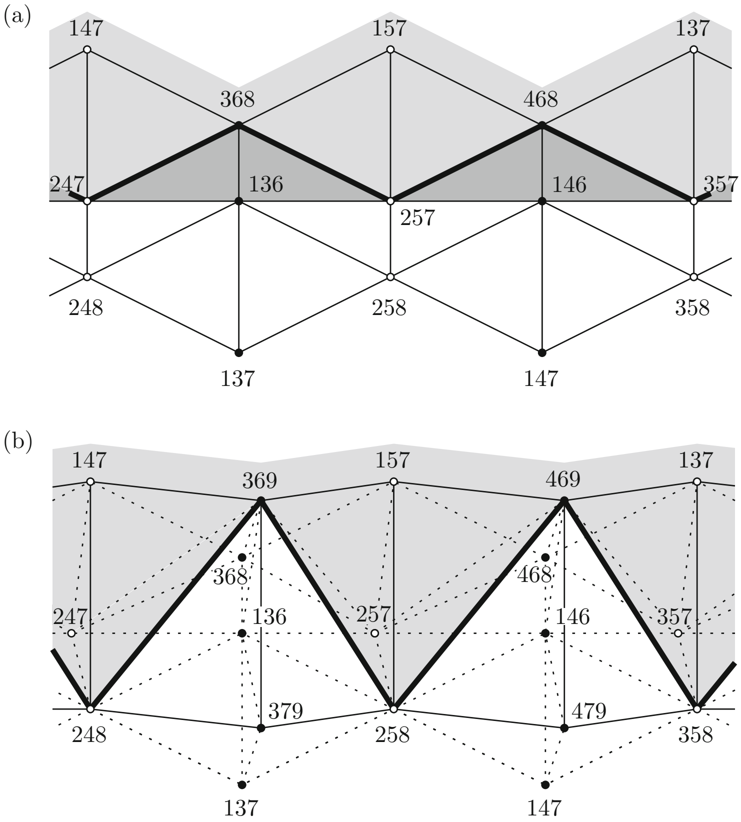


Fig. 7 Complexes defined in the proof of Lemma 14 (for $n = 9, k = 3$). Compare to parts (**a**) and (**f**) of Fig. 6. (**a**) A portion of the complex $\mathsf{QK}(8,3)$. The *thick line* shows the complex K^0. The complex $\mathsf{QR}^\bullet(8,3)$ is shown in *dark gray*, the complex K^+ in *light gray*. (**b**) A portion of the complex $\mathsf{QR}^\bullet(9,3)$. The *thick line* shows the complex L^0. The complex L^+ is shown in *light gray*

In fact, the construction implies that $\mathsf{QB}^\bullet(n-1,k)$ is the induced subcomplex of $\mathsf{QK}(n-1,k)$ on X.

As described in steps (B6)–(B9) below, the complex $\mathsf{QB}^\bullet(n-1,k)$ has been constructed as the union of the complex $\mathsf{QR}^\bullet(n-1,k)$ and a complex, say K^+, isomorphic to $\mathsf{QB}^\circ(n-3,k-1)$. The intersection of $\mathsf{QR}^\bullet(n-1,k)$ and K^+ is the interior boundary K^0 of $\mathsf{QR}^\bullet(n-1,k)$. (See the illustration in Fig. 7a.) By the induction hypothesis, K^0 is isomorphic to $\mathsf{QK}(n-4,k-1)$, and by part (i) of the lemma, it is the induced subcomplex of $\mathsf{QR}^\bullet(n-1,k)$ on vertex set

$$X^0 = \bigcup_{A \in V_+(n-2,k)} \{A\langle n-1\rangle^\bullet, A^\circ\}\,. \tag{2}$$

The said construction of $\mathsf{QB}^\bullet(n-1,k)$ also implies that K^+ is obtained from $\mathsf{QB}^\bullet(n-1,k)$ by removing the set of all the vertices of $\mathsf{QR}^\bullet(n-1,k)$ that are not contained in X^0 (cf. Observation 15 below). Comparing (2) to Observation 13, we find that this set is

$$Y = \{A^\bullet : A \in V_0(n-1,k)\} \cup \{A^\circ : A \in V_0(n-2,k)\}\,.$$

Using Corollary 7 and inspecting the definition of h, we find that the restriction of h to the vertex set of K^+, namely $X \setminus Y$, is one-to-one. Let L^+ be the image of K^+ under h (thus, h maps K^+ isomorphically to L^+), and define L^0 as the image of K^0. (See Fig. 7b.) Furthermore, let K^- be the antipodal copy of K^+, and let L^- be the image of K^- under h. Since h is also one-to-one when restricted to the vertex set of K^-, L^- is isomorphic to $\mathsf{QB}^\bullet(n-3,k-1)$.

It can be shown using the definition of h and Lemma 8 that a vertex $A^\bullet$ or A° of $\mathsf{QK}(n-1,k)$ is mapped by h to L^0 if and only if $A \in V_0(n-1,k) \cup V_1(n-1,k)$. Consequently, K^0 is mapped isomorphically to L^0. Furthermore, it follows that the intersection of L^+ and L^- equals L^0.

We have expressed IB as the union of two complexes, one isomorphic to $\mathsf{QB}^\bullet(n-3,k-1)$ and the other one to $\mathsf{QB}^\circ(n-3,k-1)$, intersecting in a subcomplex isomorphic to $\mathsf{QK}(n-4,k-1)$. In view of steps (K1)–(K3) below, this implies that IB is isomorphic to $\mathsf{QK}(n-3,k-1)$ as claimed. □

We can now finish the construction of $\mathsf{QB}^\bullet(n,k)$ (see Fig. 8 for a concrete example):

(B6) We identify the interior boundary of $\mathsf{QR}^\bullet(n,k)$ with $\mathsf{QK}(n-3,k-1)$ via the isomorphism of Lemma 14(ii).

(B7) Applying the recursion, we extend this embedding of $\mathsf{QK}(n-3,k-1)$ to an embedding of $\mathsf{QB}^\circ(n-2,k-1)$ (note the change of colour).

(B8) We form the complex $\mathsf{QB}^\bullet(n,k)$ as the union of $\mathsf{QR}^\bullet(n,k)$ (constructed above) and $\mathsf{QB}^\circ(n-2,k-1)$.

(B9) We give an explicit rule to relabel the vertices of $\mathsf{QB}^\circ(n-2,k-1)$ with elements of $V(n,k)$ in such a way that the labelling of the boundary matches the original labelling in $\mathsf{QR}^\bullet(n,k)$ and each element of $V(n,k)$ appears as the label of a vertex of $\mathsf{QB}^\bullet(n,k)$ (either a unique non-boundary vertex, or two antipodal boundary vertices).

Observation 15 *Let Y be the set of vertices not contained in the interior boundary IB of $\mathsf{QR}^\bullet(n,k)$. Then the complex $\mathsf{QB}^\bullet(n,k) \setminus Y$, obtained by removing all the vertices in Y, is isomorphic to $\mathsf{QB}^\circ(n-2,k-1)$.*

To relabel the vertices of $\mathsf{QB}^\circ(n-2,k-1)$ so as to accomplish step (B9), we will use the mapping g_n of Sect. 3; recall that for $A \in V(n-2,k-1)$,

$$g_n(A) = \begin{cases} (A+1) \cup \{1\} & \text{if } n-2 \in A,\\ (A+1) \cup \{n\} & \text{otherwise.} \end{cases}$$

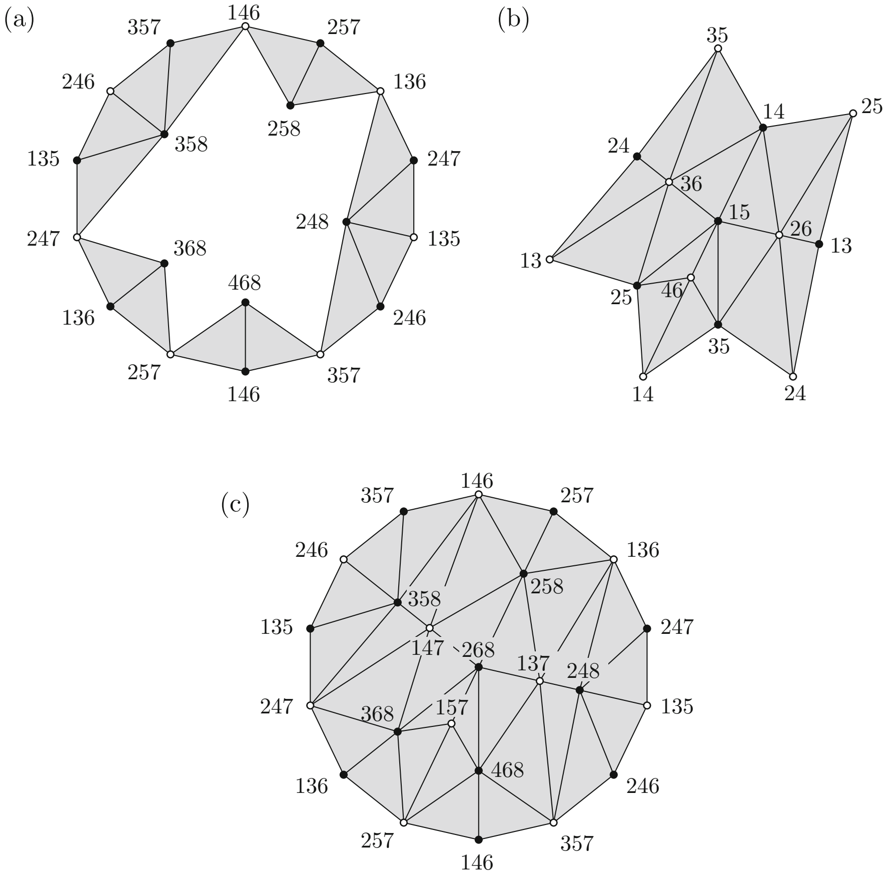


Fig. 8 The construction of $\mathsf{QB}^{\bullet}(8,3)$. (**a**) $\mathsf{QR}^{\bullet}(8,3)$. (**b**) $\mathsf{QB}^{\circ}(6,2)$ (deformed so as to match the interior boundary of $\mathsf{QR}^{\bullet}(8,3)$). (**c**) Filling in $\mathsf{QR}^{\bullet}(8,3)$ using $\mathsf{QB}^{\circ}(6,2)$ produces $\mathsf{QB}^{\bullet}(8,3)$. The labelling of the vertices inside the disk is discussed in rule (B9)

We relabel each black vertex $A^{\bullet}$ of $\mathsf{QB}^{\circ}(n-2,k-1)$ to $g_n(A)^{\bullet}$ (cf. Fig. 8). A white vertex A° is relabelled to

$$\begin{array}{rl} g_n(A)^{\circ} & \text{if } A \in V_+(n-2,k-1), \\ g_{n-1}(A)^{\circ} & \text{otherwise.} \end{array}$$

We need to check that any vertex at the interior boundary of $\mathsf{QR}^{\bullet}(n,k)$ is mapped to itself by $g_n \circ f$ ($g_{n-1} \circ f$, respectively). These are the vertices in the set W defined in Lemma 14(i). Recall that

$$W = \bigcup_{A \in V_+(n-1,k)} \{A\langle n\rangle^{\bullet}, A^{\circ}\}.$$

It follows from Lemma 6 that for $A \in V_+(n-1,k)$, $g_{n-1}(f(A)) = A$ and $g_n(f(A\langle n\rangle)) = A\langle n\rangle$, proving the requested property. Further properties of the labelling will be proved in Lemmas 16 and 18 below.

We finally construct $\mathsf{QK}(n,k)$ as follows:

(K1) We embed a deformed copy of $\mathsf{QB}^\bullet(n,k)$ in $\mathbb{R}^{n-2k+1}$, with its vertices placed in the closed upper hemisphere H^+ of S^{n-2k}, in such a way that the embedded complex is boundary-antisymmetric (thus, the boundary $\mathsf{QK}(n-1,k)$ is necessarily embedded in the 'equator' S^{n-2k-1}).

(K2) Projecting each vertex of $\mathsf{QB}^\bullet(n,k)$ to its antipode in S^{n-2k} and inverting its colour, we obtain a copy of $\mathsf{QB}^\circ(n,k)$ in the closed lower hemisphere H^- that matches the former copy at the boundary.

(K3) $\mathsf{QK}(n,k)$ is the result of glueing the above (deformed) copies of $\mathsf{QB}^\bullet(n,k)$ and $\mathsf{QB}^\circ(n,k)$ together along their boundaries.

To finish the proof of Theorem 10, we need to establish several lemmas that verify the required properties of $\mathsf{QK}(n,k)$.

Lemma 16 *Each element of $V(n,k)$ appears as (the label of) a vertex of $\mathsf{QK}(n,k)$.*

Proof The assertion is easy to check for $n = 2k+1$. If $n > 2k+1$, we inductively assume that it is true for $n-1$. Thus, any set $A \in V(n-1,k)$ is the label of a vertex of $\mathsf{QK}(n-1,k) \subseteq \mathsf{QK}(n,k)$.

By Lemma 4, it is sufficient to consider a set $A \in V_+(n,k)$. Let $B = f(A)$, where $B \in V(n-2,k-1)$. By the induction hypothesis, B is the label of a vertex of $\mathsf{QK}(n-2,k-1)$, and hence of the complex $\mathsf{QB}^\circ(n-2,k-1)$ used in the construction of $\mathsf{QB}^\bullet(n,k)$. We may assume that the vertex is $B^\bullet$ (the argument for B° being symmetric). The vertex was labelled with $g_n(B)$ in $\mathsf{QB}^\bullet(n,k)$; by Lemma 6, $g_n(f(A)) = A$ when $A \in V_+(n,k)$, so A does appear as a vertex label in $\mathsf{QB}^\bullet(n,k)$ and $\mathsf{QK}(n,k)$. □

Lemma 17 *Any edge $A^\bullet B^\circ$ of $\mathsf{QR}^\bullet(n,k)$, where $A, B \in V(n-1,k)$, is an edge of $\mathsf{QK}(n-1,k)$.*

Proof Consider an edge $A^\bullet B^\circ$ of $\mathsf{QR}^\bullet(n,k)$ but not of $\mathsf{QK}(n-1,k)$, where $A, B \in V(n-1,k)$. In the construction of $\mathsf{QR}^\bullet(n,k)$, which starts from $\mathsf{QK}(n-1,k)$, the edge $A^\bullet B^\circ$ was not added in step (B1) as this step consists in adding cones whose apex does not belong to $\mathsf{QK}(n-1,k)$. There is an application of rule (B1) where the equivalence class of A is processed. If L is the interior boundary of the complex obtained at the point of this application, then $A^\bullet$ is clearly an inside vertex of $\mathsf{Reg}(A^\bullet, \mathsf{L})$. By Observation 2, after step (B1) is completed, $A^\bullet$ is not contained in the interior boundary K_1 of the resulting complex. Consequently, step (B2) does not influence the set of edges incident with $A^\bullet$. Thus, there is no step where $A^\bullet B^\circ$ can be added, which is a contradiction. □

Lemma 18 *For any $A, B \in V(n,k)$ such that $A^\bullet$ and B° are adjacent in $\mathsf{QK}(n,k)$, $A \cap B = \emptyset$.*

Proof We proceed by induction on n. The claim is easy to verify for $n = 2k + 1$. Assume that this is not the case; in addition, we may assume that $k > 1$. Let $A^\bullet B^\circ$ be an edge of $\mathsf{QK}(n, k)$.

By the induction hypothesis, it may be assumed that $A^\bullet B^\circ$ is an edge of $\mathsf{QB}^\bullet(n, k)$ but not of $\mathsf{QK}(n-1, k)$. Suppose first that $A^\bullet B^\circ$ is an edge of $\mathsf{QR}^\bullet(n, k)$. By the fact that each white vertex of $\mathsf{QR}^\bullet(n, k)$ is a vertex of $\mathsf{QK}(n - 1, k)$ and by Lemma 17, we find that $A^\bullet$ is not a vertex of $\mathsf{QK}(n - 1, k)$. Inspecting steps (B1)–(B2) of the construction, we observe that there are two possibilities:

- there is a set $C \in V(n - 1, k)$ such that $A = C\langle n\rangle$ and $C^\bullet B^\circ$ is an edge of $\mathsf{QK}(n - 1, k)$, or
- there are sets $C \in V(n-1, k), D \in V_0(n-1, k)$ such that $A = C\langle n\rangle, B = D\langle n-1\rangle$ and $C^\bullet D^\circ$ is an edge of $\mathsf{QK}(n - 1, k)$.

In the first case, $C \cap B = \emptyset$ by the induction hypothesis and $n \notin C$, so $A \cap B = \emptyset$. In the second case, we similarly have $C \cap D = \emptyset$ by the induction hypothesis; since $n - 1 \notin A$ and $n \notin B$, we conclude that $A \cap B = \emptyset$.

It remains to consider the case that the edge $A^\bullet B^\circ$ is not an edge of $\mathsf{QR}^\bullet(n, k)$. By the construction of $\mathsf{QB}^\bullet(n, k)$, $f(A)^\bullet f(B)^\circ$ is an edge of $\mathsf{QB}^\circ(n - 2, k - 1)$. By the induction hypothesis, $f(A) \cap f(B) = \emptyset$. By Lemma 6, since $A, B \in V_+(n, k)$, $A = g_n(f(A))$ and $B = g_n(f(B))$. The definition of g_n shows that $A \cap B = \emptyset$ if one of $f(A), f(B)$ contains $n - 2$. Suppose thus that $n - 2 \notin f(A) \cup f(B)$. Then the $(n - 2)$-level of both $f(A)$ and $f(B)$ is even. Since the vertices $f(A)^\bullet$ and $f(B)^\circ$ have different colours, the $(n - 2)$-levels actually have to be zero by property (P1) of $\mathsf{QK}(n - 2, k - 1)$. Thus, $f(A), f(B) \in V_0(n - 2, k - 1)$, so $f(A)^\bullet f(B)^\circ$ is an edge of the exterior boundary $\mathsf{QK}(n - 3, k - 1)$ of $\mathsf{QB}^\circ(n - 2, k - 1)$—but then $A^\bullet B^\circ$ would be an edge of the exterior boundary of $\mathsf{QR}^\bullet(n, k)$, a contradiction. □

Lemmas 16 and 18 imply part (iii) of Theorem 10. Parts (i) and (iv) follow easily from the construction. Part (ii) is a consequence of the following lemma:

Lemma 19 *The complex* $\mathsf{QK}(n, k)$ *contains no monochromatic maximal faces.*

Proof By induction. The claim is evident for $\mathsf{QK}(2k + 1, k)$. For $n > 2k + 1$, let us assume that $\mathsf{QK}(n - 1, k)$ has no monochromatic maximal faces. The complex $\mathsf{QR}^\bullet(n, k)$ is obtained by two operations: adding cones and contracting 1-dimensional monochromatic faces. None of these operations can create a monochromatic maximal face, so $\mathsf{QR}^\bullet(n, k)$ has no such faces. The rest follows using Lemma 14 and induction. □

It only remains to check properties (P1)–(P3) of $\mathsf{QK}(n, k)$. This verification is left to the reader, which concludes the proof of Theorem 10.

Theorem 1 is a direct consequence of Theorem 10 and the results in [4]. Indeed, let the graph $QG(n, k)$ be obtained from the associated graph of $\mathsf{QK}(n, k)$ by identifying antipodal pairs of vertices (and discarding the colours). By Theorem 10 (i)–(ii) and [4, Lemma 3.2], $QG(n, k)$ is a quadrangulation of the projective space $\mathbb{P}^{n-2k}$. Theorem 10 (iii) implies that the quadrangulation is a spanning subgraph of $SG(n, k)$, while part (iv) implies that $QG(n, k)$ contains the $(2k + 1)$-cycle

$QG(2k+1,k)$ and is therefore non-bipartite. Finally, by [4, Theorem 1.1] and the easy upper bound on $\chi(SG(n,k))$, the chromatic number of $QG(n,k)$ is $n-2k+2$.

5 Conclusion

We conclude this paper with two open problems.

While the proof of Theorem 10 provides a recursive characterisation of the pairs of sets in $V(n,k)$ that are adjacent in the graph $QG(n,k)$, it would be desirable to define this graph directly, without recursion. We have no such definition so far.

Recall that the Schrijver graph $SG(n,k)$ is a vertex-critical subgraph of the Kneser graph $KG(n,k)$ with the same chromatic number, namely $n-2k+2$. By Theorem 1, the spanning subgraph $QG(n,k)$ of $SG(n,k)$ has the same chromatic number, and we conjecture that it is the natural next step in the direction set by Schrijver:

Conjecture 20 *For any $k \geq 1$ and $n \geq 2k+1$, $QG(n,k)$ is edge-critical.*

The conjecture is clearly true for $n = 2k+1$ (odd cycles) and its validity for $n = 2k+2$ can be derived from a result of Gimbel and Thomassen [2].

Acknowledgements We thank two anonymous reviewers for a careful reading of the first versions and a number of suggestions. This project was started while the second author was visiting Fabricio Benevides and Víctor Campos at the Universidade Federal do Ceará.

References

1. D. Attali, A. Lieutier, D. Salinas, Efficient data structure for representing and simplifying simplicial complexes in high dimensions. Int. J. Comput. Geom. Appl. **22**(4), 279–303 (2012)
2. J. Gimbel, C. Thomassen, Coloring graphs with fixed genus and girth. Trans. Am. Math. Soc. **349**, 4555–4564 (1997)
3. A. Hatcher, *Algebraic Topology* (Cambridge University Press, Cambridge, 2002)
4. T. Kaiser, M. Stehlík, Colouring quadrangulations of projective spaces. J. Combin. Theory Ser. B **113**, 1–17 (2015)
5. M. Kneser, Aufgabe 300. Jahresber. Deutsch. Math.-Verein. **58**, 27 (1955)
6. L. Lovász, Kneser's conjecture, chromatic number, and homotopy. J. Combin. Theory Ser. A **25**(3), 319–324 (1978)
7. J. Matoušek, *Using the Borsuk-Ulam Theorem*. Universitext (Springer, Berlin, 2003)
8. J.R. Munkres, *Elements of Algebraic Topology* (Addison–Wesley Publishing Company, Menlo Park, 1984)
9. E. Nevo, Higher minors and Van Kampen's obstruction. Math. Scand. **101**(2), 161–176 (2007)
10. A. Schrijver, Vertex-critical subgraphs of Kneser graphs. Nieuw Arch. Wisk. (3) **26**(3), 454–461 (1978)

Near-Optimal Lower Bounds for ϵ-Nets for Half-Spaces and Low Complexity Set Systems

Andrey Kupavskii, Nabil H. Mustafa, and János Pach

Abstract Following groundbreaking work by Haussler and Welzl (1987), the use of small ϵ-nets has become a standard technique for solving algorithmic and extremal problems in geometry and learning theory. Two significant recent developments are: (i) an upper bound on the size of the smallest ϵ-nets for set systems, as a function of their so-called shallow-cell complexity (Chan, Grant, Könemann, and Sharpe); and (ii) the construction of a set system whose members can be obtained by intersecting a point set in $\mathbb{R}^4$ by a family of half-spaces such that the size of any ϵ-net for them is $\Omega(\frac{1}{\epsilon}\log\frac{1}{\epsilon})$ (Pach and Tardos).

The present paper completes both of these avenues of research. We (i) give a lower bound, matching the result of Chan et al., and (ii) generalize the construction of Pach and Tardos to half-spaces in $\mathbb{R}^d$, for any $d \geq 4$, to show that the general upper bound, $O(\frac{d}{\epsilon}\log\frac{1}{\epsilon})$, of Haussler and Welzl for the size of the smallest ϵ-nets is tight.

The work of Andrey Kupavskii has been supported in part by the Swiss National Science Foundation Grants 200021-137574 and 200020-14453 and by the grant N 15-01-03530 of the Russian Foundation for Basic Research.

The work of Nabil H. Mustafa has been supported by the grant ANR SAGA (JCJC-14-CE25-0016-01).

The work of János Pach and Andrey Kupavskii has been partially supported by Swiss National Science Foundation Grants 200021-165977 and 200020-162884.

A. Kupavskii
Moscow Institute of Physics and Technology, Dolgoprudny, Russia

EPFL, Lausanne, Switzerland
e-mail: kupavskii@yandex.ru

N.H. Mustafa
LIGM, Equipe A3SI, ESIEE Paris, Université Paris-Est, Champs-sur-Marne, France
e-mail: mustafan@esiee.fr

J. Pach (✉)
EPFL, Lausanne, Switzerland

Rényi Institute, Budapest, Hungary
e-mail: pach@cims.nyu.edu

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_21

1 Introduction

Let X be a finite set and let $\mathcal{R}$ be a system of subsets of an underlying set containing X. In computational geometry, the pair $(X, \mathcal{R})$ is usually called a *range space*. A subset $X' \subseteq X$ is called an *ϵ-net* for $(X, \mathcal{R})$ if $X' \cap R \neq \emptyset$ for every $R \in \mathcal{R}$ with $|R \cap X| \geq \epsilon|X|$. The use of small-sized ϵ-nets in geometrically defined range spaces has become a standard technique in discrete and computational geometry, with many combinatorial and algorithmic consequences. In most applications, ϵ-nets precisely and provably capture the most important quantitative and qualitative properties that one would expect from a random sample. Typical applications include the existence of spanning trees and simplicial partitions with low crossing number, upper bounds for discrepancy of set systems, LP rounding, range searching, streaming algorithms; see [13, 18].

For any subset $Y \subseteq X$, define the *projection* of $\mathcal{R}$ on Y to be the set system

$$\mathcal{R}|_Y := \left\{Y \cap R : R \in \mathcal{R}\right\}.$$

The *Vapnik-Chervonenkis dimension* or, in short, the *VC-dimension* of the range space $(X, \mathcal{R})$ is the minimum integer d such that $|\mathcal{R}|_Y| < 2^{|R|}$ for any subset $Y \subseteq X$ with $|Y| > d$. According to the Sauer–Shelah lemma [21, 23] (discovered earlier by Vapnik and Chervonenkis [24]), for any range space $(X, \mathcal{R})$ whose VC-dimension is at most d and for any subset $Y \subseteq X$, we have $|\mathcal{R}|_Y| \leq \sum_{i=0}^{d} \binom{|Y|}{i} = O(|Y|^d)$.

A straightforward sampling argument shows that every range space $(X, \mathcal{R})$ has an ϵ-net of size $O(\frac{1}{\epsilon} \log |\mathcal{R}|_X|)$. The remarkable result of Haussler and Welzl [10], based on the previous work of Vapnik and Chervonenkis [24], shows that much smaller ϵ-nets exist if we assume that our range space has small VC-dimension. Haussler and Welzl [10] showed that if the VC-dimension of a range space $(X, \mathcal{R})$ is at most d, then by picking a random sample of size $\Theta(\frac{d}{\epsilon} \log \frac{d}{\epsilon})$, we obtain an ϵ-net with positive probability. Actually, they only used the weaker assumption that $|\mathcal{R}|_Y| = O(|Y|^d)$ for every $Y \subseteq X$. This bound was later improved to $(1 + o(1))(\frac{d}{\epsilon} \log \frac{1}{\epsilon})$, as $\frac{1}{\epsilon} \to \infty$ and d is large [11]. In the sequel, we will refer to this result as the *ϵ-net theorem*. The key feature of the ϵ-net theorem is that it guarantees the existence of an ϵ-net whose size is *independent* of both $|X|$ and $|\mathcal{R}|_X|$. Furthermore, if one only requires the VC-dimension of $(X, \mathcal{R})$ to be bounded by d, then this bound cannot be improved. It was shown in [11] that given any $\epsilon > 0$ and integer $d \geq 2$, there exist range spaces with VC-dimension at most d, and for which any ϵ-net must have size at least $\left(1 - \frac{2}{d} + \frac{1}{d(d+2)} + o(1)\right)\frac{d}{\epsilon} \log \frac{1}{\epsilon}$.

The effectiveness of ϵ-net theory in geometry derives from the fact that most "geometrically defined" range spaces $(X, \mathcal{R})$ arising in applications have bounded VC-dimension and, hence, satisfy the preconditions of the ϵ-net theorem.

There are two important types of geometric set systems, both involving points and geometric objects in $\mathbb{R}^d$, that are used in such applications. Let $\mathcal{R}$ be a family of possibly unbounded geometric objects in $\mathbb{R}^d$, such as the family of all half-spaces, all balls, all polytopes with a bounded number of facets, or all *semialgebraic sets* of bounded complexity, i.e., subsets of $\mathbb{R}^d$ defined by at most D polynomial equations or inequalities in the d variables, each of degree at most D. Given a finite set of points $X \subset \mathbb{R}^d$, we define the *primal range space* $(X, \mathcal{R})$ as the set system "induced by containment" in the objects from $\mathcal{R}$. Formally, it is a set system with the set of elements X and sets $\{X \cap R : R \in \mathcal{R}\}$. The combinatorial properties of this range space depend on the projection $\mathcal{R}|_X$. Using this terminology, Radon's theorem [13] implies that the primal range space on a ground set X, induced by containment in half-spaces in $\mathbb{R}^d$, has VC-dimension at most $d+1$ [18]. Thus, by the ϵ-net theorem, this range space has an ϵ-net of size $O(\frac{d}{\epsilon}\log\frac{1}{\epsilon})$.

In many applications, it is natural to consider the dual range space, in which the roles of the points and ranges are swapped. As above, let $\mathcal{R}$ be a family of geometric objects (ranges) in $\mathbb{R}^d$. Given a finite set of objects $\mathcal{S} \subseteq \mathcal{R}$, the *dual range space* "induced" by them is defined as the set system (hypergraph) on the ground set $\mathcal{S}$, consisting of the sets $S_x := \{S \in \mathcal{S} : x \in S\}$ for all $x \in \mathbb{R}^d$. It can be shown that if for any $X \subset \mathbb{R}^d$ the VC-dimension of the range space $(X, \mathcal{R})$ is less than d, then the VC-dimension of the dual range space induced by any subset of $\mathcal{R}$ is less than 2^d [13].

Recent progress In many geometric scenarios, however, one can find smaller ϵ-nets than those whose existence is guaranteed by the ϵ-net theorem. It has been known for a long time that this is the case, e.g., for primal set systems induced by containment in balls in $\mathbb{R}^2$ and half-spaces in $\mathbb{R}^2$ and $\mathbb{R}^3$. Over the past two decades, a number of specialized techniques have been developed to show the existence of small-sized ϵ-nets for such set systems [3–7, 9, 11, 12, 14, 15, 20, 25, 26]. Based on these successes, it was generally believed that in most geometric scenarios one should be able to substantially strengthen the ϵ-net theorem, and obtain perhaps even a $O\big(\frac{1}{\epsilon}\big)$ upper bound for the size of the smallest ϵ-nets. In this direction, there have been two significant recent developments: one positive and one negative.

Upper bounds Following the work of Clarkson and Varadarajan [9], it has been gradually realized that if one replaces the condition that the range space $(X, \mathcal{R})$ has bounded VC-dimension by a more refined combinatorial property, one can prove the existence of ϵ-nets of size $o(\frac{1}{\epsilon}\log\frac{1}{\epsilon})$. To formulate this property, we need to introduce some terminology.

Given a function $\varphi : \mathbb{N} \to \mathbb{R}^+$, we say that the range space $(X, \mathcal{R})$ has *shallow-cell complexity* φ if there exists a constant $c = c(\mathcal{R}) > 0$ such that, for every $Y \subseteq X$ and for every positive integer l, the number of at most l-element sets in $\mathcal{R}|_Y$ is $O\big(|Y| \cdot \varphi(|Y|) \cdot l^c\big)$. Note that if the VC-dimension of $(X, \mathcal{R})$ is d, then for every $Y \subseteq X$, the number of elements of the projection of the set system $\mathcal{R}$ to Y satisfies $|\mathcal{R}|_Y| = O(|Y|^d)$. However, the condition that $(X, \mathcal{R})$ has *shallow-cell complexity* φ for some function $\varphi(n) = O(n^{d'}), 0 < d' < d-1$ and some constant $c = c(\mathcal{R})$, implies not only that $|\mathcal{R}|_Y| = O(|Y|^{1+d'+c})$, but it reveals some nontrivial finer details about the distribution of the sizes of the smaller members of $\mathcal{R}|_Y$.

Several of the range spaces mentioned earlier turned out to have low shallow-cell complexity. For instance, the primal range spaces induced by containment of points in disks in $\mathbb{R}^2$ or half-spaces in $\mathbb{R}^3$ have shallow-cell complexity $\varphi(n) = O(1)$. In general, it is known [13] that the primal range space induced by containment of points by half-spaces in $\mathbb{R}^d$ has shallow-cell complexity $\varphi(n) = O\big(n^{\lfloor d/2 \rfloor - 1}\big)$.

Define the *union complexity* of a family of objects $\mathcal{R}$, as the maximum number of *faces* (boundary pieces) of all dimensions that the union of any n members of $\mathcal{R}$ can have; see [1]. Applying a simple probabilistic technique developed by Clarkson and Shor [8], one can find an interesting relationship between the union complexity of a family of objects $\mathcal{R}$ and the shallow-cell complexities of the *dual* range spaces induced by subsets $\mathcal{S} \subset \mathcal{R}$. Suppose that the union complexity of a family $\mathcal{R}$ of objects in the plane is $O\big(n\varphi(n)\big)$, for some "well-behaved" non-decreasing function φ. Then the number of at most l-element subsets in the dual range space induced by any $\mathcal{S} \subset \mathcal{R}$ is $O\big(l^2 \cdot \frac{|\mathcal{S}|}{l}\varphi(\frac{|\mathcal{S}|}{l})\big) = O\big(|\mathcal{S}|\varphi(|\mathcal{S}|)l\big)$ [22]; i.e., the dual range space induced by $\mathcal{S}$ has shallow-cell complexity $O\big(\varphi(n)\big)$. According to the above definitions, this means that for any $\mathcal{S} \subset \mathcal{R}$ and for any positive integer l, the number of subsets $\mathcal{S}' \in \binom{\mathcal{S}}{\leq l}$ for which there is a point $p' \in \mathbb{R}^2$ contained in all elements of $\mathcal{S}'$, but in none of the elements of $\mathcal{S} \setminus \mathcal{S}'$, is at most $O\big(|\mathcal{S}|\varphi(|\mathcal{S}|)l\big)$. For small values of l, the points p' are not heavily covered. Thus, the corresponding *cells* $\bigcap_{S \in \mathcal{S}'} S \setminus \bigcup_{T \in \mathcal{S} \setminus \mathcal{S}'} T$ of the arrangement $\mathcal{S}$ are "shallow," and the number of these shallow cells is bounded from above. This explains the use of the term "shallow-cell complexity".

A series of elegant results [3, 6, 20, 26] illustrate that if the shallow-cell complexity of a set system is $\varphi(n) = o(n)$, then it permits smaller ϵ-nets than what is guaranteed by the ϵ-net theorem. The following theorem represents the current state of the art; see [17] for a simple proof of this statement.

Theorem A *Let $(X, \mathcal{R})$ be a range space with shallow-cell complexity $\varphi(\cdot)$, where $\varphi(n) = O(n^d)$ for some constant d. Then, for every $\epsilon > 0$, it has an ϵ-net of size $O\big(\frac{1}{\epsilon}\log\varphi(\frac{1}{\epsilon})\big)$, where the constant hidden in the O-notation depends on d.*

Proof (Sketch.) The main result in [6] shows the existence of ϵ-nets of size $O\big(\frac{1}{\epsilon}\log\varphi(|X|)\big)$ for any non-decreasing function φ[1]. To get a bound independent of $|X|$, first compute a small $(\epsilon/2)$-approximation $A \subseteq X$ for $(X, \mathcal{R})$ [13]. It is known that there is such an A with $|A| = O\big(\frac{d}{\epsilon^2}\log\frac{1}{\epsilon}\big) = O(\frac{1}{\epsilon^3})$, and for any $R \in \mathcal{R}$, we have $\frac{|R \cap A|}{|A|} \geq \frac{|R|}{|X|} - \frac{\epsilon}{2}$. In particular, any $R \in \mathcal{R}$ with $|R| \geq \epsilon|X|$ contains at least an $\frac{\epsilon}{2}$-fraction of the elements of A. Therefore, an $(\epsilon/2)$-net for $(A, \mathcal{R}|_A)$ is an ϵ-net for $(X, \mathcal{R})$. Computing an $(\epsilon/2)$-net for $(A, \mathcal{R}|_A)$ gives the required set of size $O\big(\frac{2}{\epsilon}\log\varphi(|A|)\big) = O\big(\frac{1}{\epsilon}\log\varphi(\frac{1}{\epsilon^3})\big) = O\big(\frac{1}{\epsilon}\log\varphi(\frac{1}{\epsilon})\big)$. □

Note that in the bounds on the sizes of ϵ-nets based on VC-dimension, we explicitly state the dependence on d. On the other hand, in the bounds based on shallow-cell complexity, we will assume that d is a constant.

Lower bounds It was conjectured for a long time [14] that most geometrically defined range spaces of bounded Vapnik-Chervonenkis dimension have "linear-sized" ϵ-nets, i.e., ϵ-nets of size $O\big(\frac{1}{\epsilon}\big)$. These hopes were shattered by Alon [2],

[1] Their result is in fact for the more general problem of *small weight* ϵ-nets.

who established a superlinear (but barely superlinear!) lower bound on the size of ϵ-nets for the primal range space induced by straight lines in the plane. Shortly after, Pach and Tardos [19] managed to establish a tight lower bound of $\Omega(\frac{1}{\epsilon}\log\frac{1}{\epsilon})$ for the size of ϵ-nets in primal range spaces induced by half-spaces in $\mathbb{R}^4$, and in several other geometric scenarios.

Theorem B ([19]) *Let $\mathcal{F}$ denote the family of half-spaces in $\mathbb{R}^4$. For any $\epsilon > 0$ and any sufficiently large integer n, there exists a set $X \subset \mathbb{R}^4$ of n points such that in the (primal) range spaces $(X, \mathcal{F})$, the size of every ϵ-net is at least $\frac{1}{9\epsilon}\log\frac{1}{\epsilon}$.*

Our contributions The aim of this paper is to complete both avenues of research opened by Theorems A and B. Our first theorem, proved in Sect. 2, generalizes Theorem B to $\mathbb{R}^d$, for $d \geq 4$. It provides an asymptotically tight bound in terms of both ε and d, and hence completely settles the ϵ-net problem for half-spaces.

Theorem 1 *For any integer $d \geq 4$, real $\epsilon > 0$ and any sufficiently large integer $n \geq n_0(\epsilon)$, there exist primal range spaces $(X, \mathcal{F})$ induced by n-element point sets X and collections of half-spaces $\mathcal{F}$ in $\mathbb{R}^d$ such that the size of every ϵ-net for $(X, \mathcal{F})$ is at least $\frac{\lfloor d/4\rfloor}{9\epsilon}\log\frac{1}{\epsilon}$.*

As was mentioned in the first subsection, for any $d \geq 1$, the VC-dimension of any range space induced by points and half-spaces in $\mathbb{R}^d$ is at most $d+1$. Thus, Theorem 1 matches, up to a constant factor independent of d and ϵ, the upper bound implied by the ϵ-net theorem of Haussler and Welzl. Noga Alon pointed out to us that it is very easy to show that for a fixed $\epsilon > 0$, the lower bound for ϵ-nets in range spaces induced by half-spaces in $\mathbb{R}^d$ has to grow at least linearly in d. To see this, suppose that we want to obtain a $\frac{1}{3}$-net, say, for the range space induced by *open* half-spaces on a set X of $3d$ points in general position in $\mathbb{R}^d$. Notice that for this we need at least $d+1$ points. Indeed, any d points of X span a hyperplane, and one of the open half-spaces determined by this hyperplane contains at least $\frac{|X|}{3}$ points.

The key element of the proof of Theorem B [19] was to construct a set $\mathcal{B}$ of $(k+3)2^{k-2}$ axis-parallel rectangles in the plane such that for any subset of them there is a set Q of at most 2^{k-1} points that hit none of the rectangles that belong to this subset and all the rectangles in its complement (the precise statement is given in Sect. 3). We generalize this statement to $\mathbb{R}^d$ by constructing roughly $\frac{d}{2}$ times more axis-parallel boxes[2] than in the planar case, but the size of the set Q remains the same size. In Sect. 3, we prove

Lemma 2 *Let $k, d \geq 2$ be integers. Then there exists a set $\mathcal{B}$ of $\lfloor\frac{d}{2}\rfloor(k+3)2^{k-2}$ axis-parallel boxes in $\mathbb{R}^d$ such that for any subset $\mathcal{S} \subseteq \mathcal{B}$, one can find a 2^{k-1}-element set Q of points with the property that*

(i) $Q \cap B \neq \emptyset$ for any $B \in \mathcal{B} \setminus \mathcal{S}$, and
(ii) $Q \cap B = \emptyset$ for any $B \in \mathcal{S}$.

[2] An *axis-parallel box* in $\mathbb{R}^d$ is the Cartesian product of d intervals. For simplicity, in the sequel, they will be called "boxes".

In the next section we show how this lemma implies the bound of Theorem 1, which is $\lfloor \frac{d}{4} \rfloor$ times better than the bound in Theorem B. The proof of Lemma 2 will be given in Sect. 3.

In Sect. 4, we show that the bound in Theorem A cannot be improved.

Definition 1 A function $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ is called *submultiplicative* if there exists a $x_0 \in \mathbb{R}^+$ such that for every $\alpha, 0 < \alpha < 1$, and $x > x_0$, we have $\varphi^{\alpha}(x^{1/\alpha}) \leq \varphi(x)$.

Some examples of submultiplicative functions are x^c for any positive c, $2^{\sqrt{\log x}}$, $\log x$, $\log\log x$, $\log^* x$, and the inverse Ackermann function.

Theorem 3 *Let d be a fixed positive integer and let $\varphi : \mathbb{R}^+ \to \mathbb{R}^+, \varphi(n) \to \infty$, be a monotonically increasing submultiplicative function that tends to infinity such that $\varphi(n) = O(n^d)$. Then, for any $\epsilon > 0$, there exist range spaces $(X, \mathcal{F})$ that have*

(i) shallow-cell complexity $\varphi(\cdot)$, and for which
(ii) the size of any ϵ-net is $\Omega(\frac{1}{\epsilon} \log \varphi(\frac{1}{\epsilon}))$.

Theorem 3 becomes interesting when $\varphi(n) = o(n)$ and the upper bound $O\big(\frac{1}{\epsilon} \log \varphi(\frac{1}{\epsilon})\big)$ in Theorem A *improves* on the general upper bound $O\big(\frac{1}{\epsilon} \log \frac{1}{\epsilon}\big)$ guaranteed by the ϵ-net theorem. Theorem 3 shows that, even if $\varphi(n) = o(n)$, this improved bound is asymptotically tight.

The best upper and lower bounds for the size of small ϵ-nets in range spaces with a given shallow-cell complexity $\varphi(\cdot)$ are based on purely combinatorial arguments, and they imply directly or indirectly all known results on ϵ-nets in geometrically defined range spaces (see [16] for a detailed discussion). This suggests that the introduction of the notion of shallow-cell complexity provided the right framework for ϵ-net theory.

2 Proof of Theorem 1 Using Lemma 2

Let $\mathcal{B}$ be a set of d-dimensional axis-parallel boxes in $\mathbb{R}^d$. We recall that the *dual range space* induced by $\mathcal{B}$ is the set system (hypergraph) on the ground set $\mathcal{B}$ consisting of the sets $\mathcal{B}_p := \{B \in \mathcal{B} : p \in B\}$ for all $p \in \mathbb{R}^d$.

Lemma 4 *Let $d \geq 1$ be an integer, and consider the dual range space induced by a set of axis-parallel boxes $\mathcal{B}$ in $\mathbb{R}^d$. Then there exists a function $f : \mathcal{B} \to \mathbb{R}^{2d}$ such that for every point $p \in \mathbb{R}^d$, there is a half-space H in $\mathbb{R}^{2d}$ with $\{f(B) : B \in \mathcal{B}_p\} = H \cap \{f(B) : B \in \mathcal{B}\}$.*

Proof By translation, we can assume that all boxes in $\mathcal{B}$ lie in the positive orthant of $\mathbb{R}^d$.

Consider the function $g : \mathcal{B} \to \mathbb{R}^{2d}$ mapping a box $B = [x_1^l, x_1^r] \times [x_2^l, x_2^r] \times \cdots \times [x_d^l, x_d^r]$ to the point $(x_1^l, 1/x_1^r, x_2^l, 1/x_2^r, \ldots, x_d^l, 1/x_d^r)$ lying in the positive orthant of $\mathbb{R}^{2d}$. Furthermore, for any $p = (a_1, a_2, \ldots, a_d) \in \mathbb{R}^d$ in the positive orthant, let C_p denote the box $[0, a_1] \times [0, 1/a_1] \times [0, a_2] \times [0, 1/a_2] \times \cdots \times [0, a_d] \times [0, 1/a_d]$ in $\mathbb{R}^{2d}$. Clearly, a point p lies in a box B in $\mathbb{R}^d$ if and only if $g(B) \in C_p$ in $\mathbb{R}^{2d}$. Thus,

g maps the set of boxes in $\mathcal{B}$ to a set of points in $\mathbb{R}^{2d}$, such that for any point p in the positive orthant of $\mathbb{R}^d$, the set of boxes $\mathcal{B}_p \subset \mathcal{B}$ that contain p are mapped to the set of points that belong to the box C_p. (Note that C_p contains the origin.)

We complete the proof by applying the following simple transformation ([19, Lemma 2.3]) to the set $Q = g(\mathcal{B})$: to each point $q \in Q$ in the positive orthant of $\mathbb{R}^{2d}$, we can assign another point q' in the positive orthant of $\mathbb{R}^{2d}$ such that for each box in $\mathbb{R}^{2d}$ that contains the origin, there is a half-space with the property that q belongs to the box if and only if q' belongs to the corresponding half-space. The mapping $f(B) = (g(B))'$ for every $B \in \mathcal{B}$ meets the requirements of the lemma. □

Lemma 5 *Given any integer $d \geq 2$, a real number $\epsilon > 0$, and a sufficiently large integer $n \geq n_0(\epsilon)$, there exists a set $\mathcal{B}$ of n axis-parallel boxes in $\mathbb{R}^d$ such that the size of any ϵ-net for the dual set system induced by $\mathcal{B}$ is at least $\frac{\lfloor \frac{d}{2} \rfloor}{9\epsilon} \log \frac{1}{\epsilon}$.*

Proof Let $\epsilon = \frac{\alpha}{2^{k-1}}$ with $k \in \mathbb{N}, k \geq 2$, and $\frac{1}{3} \leq \alpha \leq \frac{2}{3}$. Applying Lemma 2, we obtain a set $\mathcal{B}$ of $\lfloor \frac{d}{2} \rfloor (k+3)2^{k-2}$ boxes in $\mathbb{R}^d$. We claim that the dual range space induced by these boxes does not admit an ϵ-net of size $(1-\alpha)|\mathcal{B}|$.

Assume for contradiction that there is an ϵ-net $\mathcal{S} \subseteq \mathcal{B}$ with $|\mathcal{S}| \leq (1-\alpha)|\mathcal{B}|$. According to Lemma 2, there exists a set Q of 2^{k-1} points in $\mathbb{R}^d$ with the property that no box in $\mathcal{S}$ contains any point of Q, but every member of $\mathcal{B} \setminus \mathcal{S}$ does. By the pigeonhole principle, there is a point $p \in Q$ contained in at least

$$\frac{|\mathcal{B} \setminus \mathcal{S}|}{|Q|} \geq \frac{\alpha|\mathcal{B}|}{|Q|} = \frac{\alpha|\mathcal{B}|}{2^{k-1}} = \epsilon|\mathcal{B}|$$

members of $\mathcal{B} \setminus \mathcal{S}$. Thus, none of the at least $\epsilon|\mathcal{B}|$ members of $\mathcal{B}$ hit by p belong to $\mathcal{S}$, contradicting the assumption that $\mathcal{S}$ was an ϵ-net.

Hence, the size of any ϵ-net in the dual range space induced by $\mathcal{B}$ is at least

$$(1-\alpha)|\mathcal{B}| = (1-\alpha)\left\lfloor \frac{d}{2} \right\rfloor (k+3)2^{k-2} = \frac{(1-\alpha)\alpha}{2} \cdot \left\lfloor \frac{d}{2} \right\rfloor \cdot \frac{k+3}{\epsilon} \geq \frac{1}{9} \cdot \left\lfloor \frac{d}{2} \right\rfloor \cdot \frac{1}{\epsilon} \cdot \log \frac{1}{\epsilon}.$$

The system of boxes constructed above has a fixed number of elements, depending on the value of $1/\epsilon$. We can obtain arbitrarily large constructions by replacing each box of $B \in \mathcal{B}$ with several slightly translated copies of B (we refer the reader to [19] for details). □

Now we are in a position to establish Theorem 1. By Lemma 4, any lower bound for the size of ϵ-nets in the dual range space induced by the set $\mathcal{B}$ of boxes in $\mathbb{R}^d$ gives the same lower bound for the size of an ϵ-net in the (primal) range space on the set of points $f(\mathcal{B}) \subset \mathbb{R}^{2d}$ corresponding to these boxes, in which the ranges are half-spaces in $\mathbb{R}^{2d}$. For any integer $d \geq 4$ and any real $\epsilon > 0$, Lemma 5 guarantees the existence of a set $\mathcal{B}$ of n axis-parallel boxes in $\mathbb{R}^{\lfloor d/2 \rfloor}$ such that any ϵ-net for the dual set system induced by $\mathcal{B}$ has size at least $\frac{\lfloor \frac{\lfloor d/2 \rfloor}{2} \rfloor}{9\epsilon} \log \frac{1}{\epsilon} = \frac{\lfloor d/4 \rfloor}{9\epsilon} \log \frac{1}{\epsilon}$. This fact, together with Lemma 4, implies the stated bound. □

3 Proof of Lemma 2

The proof of Lemma 2 is based on the following key statement.

Lemma C ([19]) *Let $k \geq 2$ be an integer. Then there exists a set $\mathcal{R}$ of $(k+3)2^{k-2}$ axis-parallel rectangles in $\mathbb{R}^2$ such that for any $\mathcal{S} \subseteq \mathcal{R}$, there exists a 2^{k-1}-element set Q of points in $\mathbb{R}^2$ with the property that*

(i) $Q \cap R \neq \emptyset$ for any $R \in \mathcal{R} \setminus \mathcal{S}$, and
(ii) $Q \cap R = \emptyset$ for any $R \in \mathcal{S}$.

Denote the x- and y-coordinates of a point $p \in \mathbb{R}^2$ by $x(p)$ and $y(p)$ respectively, and set $m = \lfloor \frac{d}{2} \rfloor$. Let $\mathcal{R} = \{R_1, \ldots, R_t\}$, $t = (k+3)2^{k-2}$, be a set of rectangles satisfying the conditions of Lemma C. By scaling, one can assume that $R \subset [0,1]^2$ for every $R \in \mathcal{R}$.

Given that a box in $\mathbb{R}^d$ is the product of d intervals, the idea of the construction is to 'lift' the rectangles in Lemma C, i.e., the set $\mathcal{R}$, to boxes in $\mathbb{R}^d$. So a rectangle $R \in \mathcal{R}$ can be mapped to a box in $\mathbb{R}^d$ which is the product d intervals: the first two being the intervals defining R, and the other $d-2$ intervals in the product being the full interval $[0,1]$. One can then again lift the same set $\mathcal{R}$ in a 'non-interfering' way by mapping R to a box whose 3-rd and 4-th intervals are the intervals of R and the remaining intervals are $[0,1]$. In this way, by packing intervals of each $R \in \mathcal{R}$ into disjoint coordinates, one can lift $\mathcal{R}$ m times to get a set of $\lfloor \frac{d}{2} \rfloor \cdot |\mathcal{R}|$ boxes in $\mathbb{R}^d$.

Formally, for $i = 1 \ldots m$, define the injective functions f_i that map a point in $\mathbb{R}^2$ to a product of d intervals in $\mathbb{R}^d$, as follows.

$$f_i(p) = \underbrace{[0,1] \times \cdots \times [0,1]}_{2i-2 \text{ intervals}} \times x(p) \times y(p) \times \underbrace{[0,1] \times \cdots \times [0,1]}_{d-2i \text{ intervals}}, \quad p \in \mathbb{R}^2.$$

This mapping lifts each rectangle $R \in \mathcal{R}$ to the box $f_i(R) = \{f_i(p) \ : \ p \in R\}$, and each set of rectangles $\mathcal{R}' \subseteq \mathcal{R}$ to the set of boxes $f_i(\mathcal{R}') = \{f_i(R) \ : \ R \in \mathcal{R}'\}$.

We now show that $\mathcal{B} = \bigcup_{i=1}^m f_i(\mathcal{R})$ is the desired set of $\lfloor \frac{d}{2} \rfloor (k+3)2^{k-2}$ boxes in $\mathbb{R}^d$. Let $\mathcal{S} \subseteq \mathcal{B}$ be a fixed set of boxes. For any index $i \in [1, m]$, set $\mathcal{R}_i \subseteq \mathcal{R}$ to be the set of preimage rectangles under f_i of the boxes in $\mathcal{S} \cap f_i(\mathcal{R})$, i.e., $\mathcal{R}_i$ satisfies $\mathcal{S} \cap f_i(\mathcal{R}) = f_i(\mathcal{R}_i)$. Let $Q_i = \{q_1^i, \ldots, q_{2^{k-1}}^i\} \subset \mathbb{R}^2$ be the set of points hitting all rectangles in $\mathcal{R} \setminus \mathcal{R}_i$ and no rectangle in $\mathcal{R}_i$; such a set exists by Lemma C. Now we argue that the set

$$Q = \begin{cases} \left\{ \left(x(q_j^1), y(q_j^1), \ldots, x(q_j^m), y(q_j^m) \right) \ : \ j \in [1, 2^{k-1}] \right\} & \text{if } d \text{ is even,} \\ \left\{ \left(x(q_j^1), y(q_j^1), \ldots, x(q_j^m), y(q_j^m), 1 \right) \ : \ j \in [1, 2^{k-1}] \right\} & \text{if } d \text{ is odd,} \end{cases}$$

of 2^{k-1} points in $\mathbb{R}^d$ is the required set for $\mathcal{S}$; i.e., Q hits all the boxes in $\mathcal{B} \setminus \mathcal{S}$, and none of the boxes in $B \in \mathcal{S}$. Take any box $B \in \mathcal{B} \setminus \mathcal{S}$; then there exists an index i and a rectangle $R \in \mathcal{R} \setminus \mathcal{R}_i$ such that R is the preimage rectangle of B under f_i. By

Lemma C, R contains a point $q \in Q_i$, and thus $B = f_i(R)$ contains the point $q' \in Q$ with $x(q)$ and $y(q)$ in its $(2i-1)$-th and $2i$-th coordinates, as all the remaining intervals defining B are $[0, 1]$ and so each such interval contains the corresponding coordinate of q'. On the other hand, let $B \in \mathcal{S}$ be a box with the preimage rectangle $R \in \mathcal{R}_i$. By Lemma C, R is not hit by any point of Q_i, and thus for any point $q' \in Q$, the $(2i-1)$-th and $2i$-th coordinates cannot both be contained in the corresponding two intervals defining B. Therefore, q' does not hit B. □

4 Proof of Theorem 3

The goal of this section is to establish lower bounds on the sizes of ϵ-nets in range spaces with given shallow-cell complexity $\varphi(\cdot)$, where $\varphi(\cdot)$ is a submultiplicative function. We will use the following property of submultiplicative functions.

Claim 6 *Let $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ be a submultiplicative function. Then*

- (i) *for all sufficiently large $x, y \in \mathbb{R}^+$, we have $\varphi(xy) \leq \varphi(x)\varphi(y)$, and*
- (ii) *if there exists a sufficiently large $x \in \mathbb{R}^+$ and a constant c such that $\varphi(x) \leq x^c$, then $\varphi(n) \leq n^c$ for every $n \geq x$.*

Proof Both of these properties follow immediately from the submultiplicativity of $\varphi(\cdot)$:

$$(i).\quad \varphi(xy) = \big(\varphi(xy)\big)^{\log_{xy} x} \cdot \big(\varphi(xy)\big)^{\log_{xy} y} \leq \varphi\big((xy)^{\log_{xy} x}\big) \cdot \varphi\big((xy)^{\log_{xy} y}\big) = \varphi(x) \cdot \varphi(y).$$

$$(ii).\quad \varphi^{\log_n x}(n) \leq \varphi(x) \leq x^c \implies \varphi(n) \leq x^{\frac{c}{\log_n x}} = x^{c \log_x n} = n^c.$$

□

Theorem 3 is a consequence of the following more precise statement.

Theorem 7 *Let $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ be a monotonically increasing submultiplicative function which tends to infinity and is bounded from above by a polynomial of constant degree. For any $0 < \delta < \frac{1}{10}$, one can find an $\epsilon_0 > 0$ with the following property: for any $0 < \epsilon < \epsilon_0$, there exists a range space with shallow-cell complexity $\varphi(\cdot)$ on a set of $n = \frac{\log \varphi\left(\frac{1}{\epsilon}\right)}{\epsilon}$ elements, in which the size of any ϵ-net is at least $\frac{\left(\frac{1}{2}-\delta\right)}{\epsilon} \log \varphi\left(\frac{1}{\epsilon}\right)$.*

Proof The parameters of the range space are as follows:

$$n = \frac{\log \varphi\left(\frac{1}{\epsilon}\right)}{\epsilon}, \quad m = \epsilon n = \log \varphi\left(\frac{1}{\epsilon}\right), \quad p = \frac{n\varphi^{1-2\delta}(n)}{\binom{n}{m}}.$$

Let d be the smallest integer such that $\varphi(n) = O(n^d)$. By Claim 6, part (ii), for any large enough n', we have $(n')^{d-1} \leq \varphi(n') \leq c_1 (n')^d$, for a suitable constant $c_1 \geq 1$.

In the most interesting case where $\varphi(n) = o(n)$, we have $d = 1$. For a small enough ϵ, we have $c_1 \le \log \varphi\big(\frac{1}{\epsilon}\big)$, so that

$$m = \log \varphi\left(\frac{1}{\epsilon}\right) \le \log\left(c_1\epsilon^{-d}\right) \le d \log \frac{c_1}{\epsilon} \le d \log n. \tag{1}$$

Consider a range space $([n], \mathcal{F})$ with a ground set $[n] = \{1, 2, \dots, n\}$ and with a system of m-element subsets $\mathcal{F}$, where each m-element subset of $[n]$ is added to $\mathcal{F}$ independently with probability p. The next claim follows by a routine application of the Chernoff bound.

Claim 8 *With high probability,* $|\mathcal{F}| \le 2n\varphi^{1-2\delta}(n)$.

Theorem 7 follows by combining the next two lemmas that show that, with high probability, the range space $([n], \mathcal{F})$

(i) does not admit an ϵ-net of size less than $\frac{\frac{1}{2}-\delta}{\epsilon} \log \varphi(\frac{1}{\epsilon})$, and
(ii) has shallow-cell complexity $\varphi(\cdot)$.

For the proofs, we need to assume that $n = n(\delta, d, \varphi)$ is a sufficiently large constant, or, equivalently, that $\epsilon_0 = \epsilon_0(\delta, d)$ is sufficiently small.

Lemma 9 *With high probability, the range space* $([n], \mathcal{F})$ *has shallow-cell complexity* $\varphi(\cdot)$.

The reason why this lemma holds is that for any $X \subset [n]$ and any k, it is very unlikely that the number of at most k-element sets exceeds the number permitted by the shallow-cell complexity condition. To bound the probability of the union of these events for all X and k, we simply use the union bound.

Proof It is enough to show that for all sufficiently large $x \ge x_0$, every $X \subseteq [n]$, $|X| = x$, and every $l \le m$, the number of sets of size *exactly* l in $\mathcal{F}|_X$ is $O(x\varphi(x))$, as this implies that the number of sets in $\mathcal{F}|_X$ of size at most l is $O\big(x\varphi(x)l\big)$. In the computations below, we will also assume that $l \ge d + 1 \ge 2$; otherwise if $l \le d$, and assuming $x \ge x_0 \ge 2d$, we have

$$\binom{x}{l} \le \binom{x}{d} \le x^d \le x\varphi(x),$$

where the last inequality follows by the assumption on $\varphi(x)$, provided that x is sufficiently large. We distinguish two cases.

Case 1: $\boldsymbol{x > \frac{n}{\varphi^{\delta/d}(x)}}$**.** In this case, we trivially upper-bound $|\mathcal{F}|_X|$ by $|\mathcal{F}|$. By Claim 8, with high probability, we have

$$\begin{aligned} |\mathcal{F}| &\le 2n \cdot \varphi^{1-2\delta}(n) \le 2n \cdot \left(\varphi(x) \cdot \varphi\left(\frac{n}{x}\right)\right)^{1-2\delta} \quad \big(\text{by Claim 6}\big) \\ &\le 2n \cdot \Big(\varphi(x) \cdot \varphi\big(\varphi^{\delta/d}(x)\big)\Big)^{1-2\delta} \qquad \Big(\text{as } \frac{n}{x} \le \varphi^{\delta/d}(x)\Big) \end{aligned}$$

$$\begin{aligned}&\leq 2n\cdot\left(c_1\varphi(x)\varphi^{\delta}(x)\right)^{1-2\delta}\qquad\left(\text{using } \varphi(t)\leq c_1t^d\right)\\&\leq 2c_1'n\varphi(x)^{1-\delta}\leq 2c_1'x\varphi(x)^{1-\delta+\delta/d}=O(x\varphi(x)).\end{aligned}$$

Case 2: $\boldsymbol{x \leq \frac{n}{\varphi^{\delta/d}(x)}}$. Denote the largest integer x that satisfies this inequality by x_1. It is clear that $x_1 = o(n)$ (recall that $\varphi(\cdot)$ is monotonically increasing and tends to infinity). We also denote the system of all l-element subsets of $\mathcal{F}|_X$ by $\mathcal{F}|_X^l$ and the set of all l-element subsets of X by $\binom{X}{l}$. Let E be the event that $\mathcal{F}$ does not have the required $\varphi(\cdot)$-shallow-cell complexity property. Then $\Pr[E] \leq \sum_{l=2}^{m}\Pr[E_l]$, where E_l is the event that for some $X \subset [n]$, $|X| = x$, there are more than $x\varphi(x)$ elements in $\mathcal{F}|_X^l$. Then, for any fixed $l \geq d+1 \geq 2$, we have

$$\Pr[E_l] \leq \sum_{x=x_0}^{x_1}\Pr\left[\exists X \subseteq [n], |X| = x, |\mathcal{F}|_X^l| > x\varphi(x)\right]$$

$$\leq \sum_{x=x_0}^{x_1}\binom{n}{x}\sum_{s=\lceil x\varphi(x)\rceil}^{\binom{x}{l}}\Pr\left[\text{For a fixed } X, |X| = x, |\{S \in \mathcal{F}|_X, |S| = l\}| = s\right]$$

$$\leq \sum_{x=x_0}^{x_1}\binom{n}{x}\sum_{s=\lceil x\varphi(x)\rceil}^{\binom{x}{l}}\binom{\binom{x}{l}}{s}\Pr\left[\text{For a fixed } X, |X| = x, \mathcal{S} \subseteq \binom{X}{l}, |\mathcal{S}| = s,\right.$$

$$\left.\text{we have } \mathcal{F}|_X^l = \mathcal{S}\right]$$

$$\leq \sum_{x=x_0}^{x_1}\binom{n}{x}\sum_{s=\lceil x\varphi(x)\rceil}^{\binom{x}{l}}\binom{\binom{x}{l}}{s}\left(1-(1-p)^{\binom{n-x}{m-l}}\right)^{s}(1-p)^{\binom{n-x}{m-l}\left(\binom{x}{l}-s\right)} \tag{2}$$

$$\leq \sum_{x=x_0}^{x_1}\sum_{s=\lceil x\varphi(x)\rceil}^{\binom{x}{l}}\left(\frac{en}{x}\right)^{x}\left(\frac{e\left(\frac{ex}{l}\right)^{l}}{s}\right)^{s}\left(p\binom{n-x}{m-l}\right)^{s} \tag{3}$$

$$\leq \sum_{x=x_0}^{x_1}\sum_{s=\lceil x\varphi(x)\rceil}^{\binom{x}{l}}\left(\frac{en}{x}\right)^{x}\left(\frac{e^{l+1}x^{l-1}}{l^l\varphi(x)}p\binom{n}{m}\frac{m^l}{(n-x-m)^l}\right)^{s} \tag{4}$$

$$\leq \sum_{x=x_0}^{x_1}\sum_{s=\lceil x\varphi(x)\rceil}^{\binom{x}{l}}\left(\frac{en}{x}\right)^{x}\left(\left(\frac{emx}{n}\right)^{l-1}\frac{e^2m\varphi^{1-2\delta}(n)}{\varphi(x)}\right)^{s} \tag{5}$$

In the transition to the expression (3), we used several times (i) the bound $\binom{a}{b} \leq \left(\frac{ea}{b}\right)^b$ for any $a, b \in \mathbb{N}$; (ii) the inequality $(1-p)^b \geq 1-bp$ for any integer $b \geq 1$ and real $0 \leq p \leq 1$; and (iii) we upper-bounded the last factor of (2) by 1.

In the transition from (3) to (4) we lower-bounded s by $x\varphi(x)$. We also used the estimate $\binom{n-x}{m-l} \leq \binom{n}{m}\frac{m^l}{(n-x-m)^l}$, which can be verified as follows.

$$\begin{aligned}\binom{n-x}{m-l} &= \binom{n-x}{m}\prod_{i=0}^{l-1}\frac{m-i}{n-x-m+(i+1)}\\ &\leq \binom{n-x}{m}\Big(\frac{m}{n-x-m}\Big)^l \leq \binom{n}{m}\frac{m^l}{(n-x-m)^l}.\end{aligned}$$

Finally, to obtain (5), we substituted the formula for p and used the fact that

$$l^l(n-x-m)^l = \big(l\cdot(n-x-m)\big)^l \geq \Big(l\cdot\frac{n}{2}\Big)^l \geq n^l,$$

as $x \leq x_1 = o(n)$, $m = \epsilon n \leq \frac{n}{4}$ for $\epsilon < \epsilon_0 \leq 1/4$ and $l \geq 2$.

Denote $x_2 = \lceil n^{1-\delta}\rceil$. We split the expression (5) into two sums Σ_1 and Σ_2. Let

$$\Sigma_1 := \sum_{x=x_0}^{x_2-1}\sum_{s=\lceil x\varphi(x)\rceil}^{\binom{x}{l}}\Big(\frac{en}{x}\Big)^x\left(\Big(\frac{emx}{n}\Big)^{l-1}\frac{e^2m\varphi^{1-2\delta}(n)}{\varphi(x)}\right)^s$$

$$\Sigma_2 := \sum_{x=x_2}^{x_1}\sum_{s=\lceil x\varphi(x)\rceil}^{\binom{x}{l}}\Big(\frac{en}{x}\Big)^x\left(\Big(\frac{emx}{n}\Big)^{l-1}\frac{e^2m\varphi^{1-2\delta}(n)}{\varphi(x)}\right)^s$$

These two sums will be bounded separately. We have

$$\Sigma_1 \leq \sum_{x=x_0}^{x_2-1}\sum_{s=\lceil x\varphi(x)\rceil}^{\binom{x}{l}}\Big(\frac{en}{x}\Big)^x\left(\Big(\frac{emx}{n}\Big)^{l-1}\frac{c_1^{1-2\delta}e^2mn^{d-2d\delta}}{x^{d-2d\delta}\varphi^{2\delta}(x)}\right)^s \tag{6}$$

$$\leq \sum_{x=x_0}^{x_2-1}\sum_{s=\lceil x\varphi(x)\rceil}^{\binom{x}{l}}\Big(\frac{en}{x}\Big)^x\left(\Big(\frac{emx}{n}\Big)^{l-1-d+2d\delta}Cm^{d+1-2d\delta}\right)^s \quad \text{(for some } C>0\text{)}$$

$$\leq \sum_{x=x_0}^{x_2-1}\sum_{s=\lceil x\varphi(x)\rceil}^{\binom{x}{l}}\Big(\frac{en}{x}\Big)^x\left(\big(n^{-\delta/2}\big)^{l-1-d+2d\delta}Cm^{d+1}\right)^s \tag{7}$$

$$\leq \sum_{x=x_0}^{x_2-1}x^l\Big(\frac{en}{x}\Big)^x\left(n^{-\frac{\delta}{2}\cdot 2d\delta}n^{\frac{\delta^2}{2}}\right)^{x\varphi(x)} \leq \sum_{x=x_0}^{x_2-1}x^l\Big(\frac{en}{x}\Big)^x n^{-\frac{x\varphi(x)d\delta^2}{2}} \tag{8}$$

$$\leq \sum_{x=x_0}^{x_2-1}n^{2x-\frac{x\varphi(x)d\delta^2}{2}} \leq \sum_{x=x_0}^{x_2-1}n^{-2x} \leq \frac{n}{n^{2x_0}} = o\left(\frac{1}{m}\right). \tag{9}$$

To obtain (6), we used the property that $\varphi(n) \le \varphi(x)\varphi\big(\frac{n}{x}\big) \le c_1\varphi(x)\big(\frac{n}{x}\big)^d$, provided that $n, x, \frac{n}{x}$ are sufficiently large. To establish (7), we used the fact that $x \le x_2 = n^{1-\delta}$ and that $em \le ed\log n \le n^{\delta/2}$. In the transition to (8), we needed that $l \ge d+1,\ d \ge 1$ and that $Cm^{d+1} \le C(d\log n)^{d+1} = o\big(n^{\delta^2/2}\big)$. Then we lower-bounded s by $x\varphi(x)$. To arrive at (9), we used that $l \le x$. The last inequality follows from the fact that x_0 is large enough, so that $\varphi(x) \ge \varphi(x_0) \ge 8/\big(d\delta^2\big)$ and that $m = o(n)$.

Next, we turn to bounding Σ_2. First, observe that

$$\varphi^{1-2\delta}(n) \le \varphi^{\frac{1-2\delta}{1-\delta}}(n^{1-\delta}) \le \varphi^{\frac{1-2\delta}{1-\delta}}(x) \le \varphi^{1-\delta}(x),$$

where we used the submultiplicativity and monotonicity of the function $\varphi(n)$ and the fact that $x \ge x_2 = \lceil n^{1-\delta} \rceil$. Second, note by that restricting ϵ to be small enough, by the submultiplicativity of $\varphi(\cdot)$, we have that for any $x \ge x_2$,

$$m \le \log\varphi\left(\frac{1}{\epsilon}\right) \le \varphi^{\delta/4}\left(\frac{1}{\epsilon}\right) \le \varphi^{\delta/3}(x_2) \le \varphi^{\delta/3}(x). \tag{10}$$

Substituting the bound for $\varphi^{1-2\delta}(n)$ in Σ_2, setting $C = e^2$, and by (10), we obtain

$$\begin{aligned}
\Sigma_2 &\le \sum_{x=x_2}^{x_1} \sum_{s=\lceil x\varphi(x)\rceil}^{\binom{x}{l}} \left(\frac{en}{x}\right)^x \left(\left(\frac{emx}{n}\right)^{l-1} C\varphi^{-2\delta/3}(x)\right)^s \\
&\le \sum_{x=x_2}^{x_1} x^l \left(\frac{en}{x}\right)^x \left(\frac{emx}{n} C\varphi^{-2\delta/3}(x)\right)^{x\varphi(x)} \qquad (11)\\
&\le \sum_{x=x_2}^{x_1} \left(\frac{n}{x}\right)^{x-x\varphi(x)} \left(e^{1+x/(x\varphi(x))} m x^{l/(x\varphi(x))} C\varphi^{-2\delta/3}(x)\right)^{x\varphi(x)} \\
&\le \sum_{x=x_2}^{x_1} \left(\frac{n}{x}\right)^{x-x\varphi(x)} \left(C'\varphi^{-\delta/3}(x)\right)^{x\varphi(x)} \quad \text{(for some constant } C' > 0) \qquad (12)\\
&\le n\left(\frac{n}{x_1}\right)^{x_2-x_2\varphi(x_2)} \left(C\varphi^{-\delta/3}(x_2)\right)^{x_2\varphi(x_2)} \le \left(\frac{n}{x_1}\right)^{x_2-x_2\varphi(x_2)} \qquad (13)\\
&= \left(\frac{x_1}{n}\right)^{x_2\varphi(x_2)-x_2} = o\left(\frac{1}{m}\right).
\end{aligned}$$

In the transition to (11), we used that $l \ge 2$ and $n^{1-\delta} \le x_1 \le n/\varphi^{\delta/d}(x_1)$, which implies that $x \le x_1 \le n/\varphi^{\delta/2d}(n)$ and, therefore,

$$emx < e\log\varphi(1/\epsilon)\frac{n}{\varphi^{\delta/2d}(n)} \le e\log\varphi(n)\frac{n}{\varphi^{\delta/2d}(n)} \le \frac{n}{\varphi^{\delta/3d}(n)} \le n.$$

To get (12), we used that for some constant $c > 1$ we have $x^{l/(x\varphi(x))} \le c^{m/\varphi(x)} \le c^{\log\varphi(x)/\varphi(x)} = O(1)$ and that $m \le \varphi^{\delta/3}(x)$ for $x \ge x_2$ by (10). To obtain (13), observe that $n^{1/(x_2\varphi(x_2))} = O(1)$. At the last equation, we used that $x_1 = o(n)$, $ne/x_1 \to \infty$ as $n \to \infty$ and that $x_2\varphi(x_2) - x_2 \ge x_2 = \Omega(n^{1-\delta})$.

We have shown that for every $l = 2, \dots, m$, we have $\Pr[E_l] = o(1/m)$, as m tends to infinity. Thus, we can conclude that $\Pr[E] \le \sum_{l=2}^{m} \Pr[E_l] = o(1)$ and, hence, with high probability the range space $([n], \mathcal{F})$ has shallow-cell complexity $\varphi(.)$. □

Now we are in a position to prove that with high probability the range space $([n], \mathcal{F})$ does not admit a small ϵ-net.

Lemma 10 *With high probability, the size of any ϵ-net of the range space $([n], \mathcal{F})$ is at least* $\frac{\left(\frac{1}{2}-\delta\right)}{\epsilon}\log\varphi\left(\frac{1}{\epsilon}\right)$.

Proof Denote by μ the probability that the range space has an ϵ-net of size $t = \frac{\left(\frac{1}{2}-\delta\right)}{\epsilon}\log\varphi\left(\frac{1}{\epsilon}\right) = \left(\frac{1}{2}-\delta\right)n$. Then we have

$$\mu \le \sum_{\substack{X\subseteq[n]\\|X|=t}} \Pr\left[X \text{ is an } \epsilon\text{-net for } \mathcal{F}\right] \le \binom{n}{t}(1-p)^{\binom{n-t}{m}} \le \binom{n}{t} e^{-p\binom{n-t}{m}} \le 2^n e^{-n\varphi^{\delta/2}(n)} = o(1). \tag{14}$$

To verify the last two inequalities, notice that, since $1 - ax > e^{-bx}$ for $b > a$, $0 < x < \frac{1}{a} - \frac{1}{b}$, we have

$$\begin{aligned}
p\binom{n-t}{m} &\ge p\binom{n}{m}\left(\frac{n-m-t}{n-t}\right)^t \ge n\varphi^{1-2\delta}(n)\left(1-\frac{m}{n-t}\right)^t \\
&= n\varphi^{1-2\delta}(n)\left(1-\frac{m}{(\frac{1}{2}+\delta)n}\right)^t \ge n\varphi^{1-2\delta}(n)e^{-\frac{mt}{\frac{1}{2}(1+\delta)n}} \\
&= n\varphi^{1-2\delta}(n)e^{-\frac{1-2\delta}{1+\delta}\log\varphi(\frac{1}{\epsilon})} = n\varphi^{1-2\delta}(n)\varphi^{-\frac{1-2\delta}{1+\delta}}\Big(\frac{1}{\epsilon}\Big) \ge n\varphi^{1-2\delta}(n)\varphi^{-\frac{1-2\delta}{1+\delta}}\big(n\big) \ge n\varphi^{\delta/2}(n).
\end{aligned}$$

Here the last but one inequality follows from $n \ge \frac{1}{\epsilon}$ and from the monotonicity of $\varphi(\cdot)$. The last inequality holds, because $\delta \le 1/10$. □

Thus, Lemmas 9 and 10 imply that with high probability the range space $([n], \mathcal{F})$ has shallow-cell complexity $\varphi(\cdot)$ and it admits no ϵ-net of size less than $\frac{\left(\frac{1}{2}-\delta\right)}{\epsilon}\log\varphi\left(\frac{1}{\epsilon}\right)$. This completes the proof of the theorem. □

Acknowledgements We thank the anonymous referees for carefully reading our manuscript and for pointing out several problems in the presentation, including a mistake in the proof of Lemma 10. A preliminary version of this paper was accepted in *SoCG 2016* (Symposium on Computational Geometry). We also thank the anonymous reviewers of the conference version for their feedback and valuable suggestions.

References

1. P.K. Agarwal, J. Pach, M. Sharir, State of the union (of geometric objects): a review, in *Computational Geometry: Twenty Years Later*, ed. by J. Goodman, J. Pach, R. Pollack (American Mathematical Society, 2008), pp. 9–48
2. N. Alon, A non-linear lower bound for planar epsilon-nets. Discrete Comput. Geom. **47**(2), 235–244 (2012)
3. B. Aronov, E. Ezra, M. Sharir, Small-size ϵ-nets for axis-parallel rectangles and boxes. SIAM J. Comput. **39**(7), 3248–3282 (2010)
4. P. Ashok, U. Azmi, S. Govindarajan, Small strong epsilon nets. Comput. Geom. **47**(9), 899–909 (2014)
5. N. Bus, S. Garg, N.H. Mustafa, S. Ray, Tighter estimates for epsilon-nets for disks. Comput. Geom. **53**, 27–35 (2016)
6. T.M. Chan, E. Grant, J. Könemann, M. Sharpe, Weighted capacitated, priority, and geometric set cover via improved quasi-uniform sampling, in *Proceedings of Symposium on Discrete Algorithms (SODA)* (2012), pp. 1576–1585
7. B. Chazelle, J. Friedman, A deterministic view of random sampling and its use in geometry. Combinatorica **10**(3), 229–249 (1990)
8. K.L. Clarkson, P.W. Shor, Application of random sampling in computational geometry, II. Discrete Comput. Geom. **4**, 387–421 (1989)
9. K. Clarkson, K. Varadarajan, Improved approximation algorithms for geometric set cover. Discrete Comput. Geom. **37**, 43–58 (2007)
10. D. Haussler, E. Welzl, Epsilon-nets and simplex range queries. Discrete Comput. Geom. **2**, 127–151 (1987)
11. J. Komlós, J. Pach, G.J. Woeginger, Almost tight bounds for epsilon-nets. Discrete Comput. Geom. **7**, 163–173 (1992)
12. J. Matoušek, On constants for cuttings in the plane. Discrete Comput. Geom. **20**(4), 427–448 (1998)
13. J. Matoušek, *Lectures in Discrete Geometry* (Springer, New York, 2002)
14. J. Matoušek, R. Seidel, E. Welzl, How to net a lot with little: small epsilon-nets for disks and halfspaces, in *Proceedings of Symposium on Computational Geometry* (1990), pp. 16–22
15. N.H. Mustafa, S. Ray, Near-optimal generalisations of a theorem of Macbeath, in *Proceedings of the Symposium on Theoretical Aspects of Computer Science (STACS)* (2014), pp. 578–589
16. N.H. Mustafa, K. Varadarajan, Epsilon-approximations and epsilon-nets, in *Handbook of Discrete and Computational Geometry*, ed. by J.E. Goodman, J. O'Rourke, C.D. Tóth (CRC Press LLC, 2016, to appear)
17. N.H. Mustafa, K. Dutta, A. Ghosh, A simple proof of optimal epsilon-nets. Combinatorica (to appear)
18. J. Pach, P.K. Agarwal, *Combinatorial Geometry* (Wiley, New York, 1995)
19. J. Pach, G. Tardos, Tight lower bounds for the size of epsilon-nets. J. AMS **26**, 645–658 (2013)
20. E. Pyrga, S. Ray, New existence proofs for epsilon-nets, in *Proceedings of the Symposium on Computational Geometry (SoCG)* (2008), pp. 199–207
21. N. Sauer, On the density of families of sets. J. Combin. Theory Ser. A **13**, 145–147 (1972)
22. M. Sharir, On k-sets in arrangement of curves and surfaces. Discrete Comput. Geom. **6**, 593–613 (1991)
23. S. Shelah, A combinatorial problem; stability and order for models and theories in infinitary languages. Pac. J. Math. **41**, 247–261 (1972)
24. V.N. Vapnik, A.Ya. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities. Theory Probab. Appl. **16**(2), 264–280 (1971)
25. K. Varadarajan, Epsilon nets and union complexity, in *Proceedings of the Symposium on Computational Geometry (SoCG)* (2009), pp. 11–16
26. K. Varadarajan, Weighted geometric set cover via quasi uniform sampling, in *Proceedings of the Symposium on Theory of Computing (STOC)* (ACM, New York, 2010), pp. 641–648

Random Simplicial Complexes: Around the Phase Transition

Nathan Linial and Yuval Peled

Abstract This article surveys some of the work done in recent years on random simplicial complexes. We mostly consider higher-dimensional analogs of the well known phase transition in $G(n,p)$ theory that occurs at $p = \frac{1}{n}$. Our main objective is to provide a more streamlined and unified perspective of some of the papers in this area.

1 Introduction

There are at least two different perspectives from which our subject can be viewed. We survey some recent developments in the emerging field of *high-dimensional combinatorics*. However, these results can be viewed as well as part of an ongoing effort to apply *the probabilistic method in topology*. The systematic study of random graphs was started by Erdős and Rényi in the early 1960s and had a major impact on discrete mathematics, computer science and engineering. Since graphs are one-dimensional simplicial complexes, why not develop an analogous theory of d-dimensional random simplicial complexes for all $d \geq 1$? To this end, an analog of Erdős and Rényi's $G(n,p)$ model, called $Y_d(n,p)$, was introduced in [12]. Such a simplicial complex Y is d-dimensional, it has n vertices and has a full $(d-1)$-dimensional skeleton. Each d-face is placed in Y independently with probability $p = p(n)$. Note that $Y_1(n,p)$ is identical with $G(n,p)$. Throughout the paper we consider a fixed dimension $d > 1$ and investigate the asymptotic topological properties of $Y_d(n,p)$. We say that a property holds asymptotically-almost-surely (a.a.s.) if its probability tends to 1 as n tends to infinity.

One of the most natural questions to ask in any model of random graphs concerns graph *connectivity*. As Erdős and Rényi famously showed, the threshold for graph

Supported by ERC grant 339096 "High-dimensional combinatorics".

Yuval Peled is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship.

N. Linial • Y. Peled (✉)
Department of Computer Science, Hebrew University of Jerusalem, 91904 Jerusalem, Israel
e-mail: nati@cs.huji.ac.il; yuvalp@cs.huji.ac.il

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_22

connectivity in $G(n,p)$ is $p = \frac{\ln n}{n}$. To draw the analogy from a topological perspective, one should seek the threshold for the vanishing of the $(d-1)$-st homology. This indeed was the motivating problem in [12]. As that paper showed, and together with subsequent work [17] this threshold in $Y_d(n,p)$ is $p = \frac{d \ln n}{n}$. Here the coefficients can come from any fixed finite abelian group. The same question for *integral* $(d-1)$-st homology has attracted considerable attention and the answer is believed to be the same. This was recently confirmed for $d = 2$ [15], and is not yet fully resolved for higher dimensions (but see [10]). The threshold for the vanishing of the fundamental group of $Y_2(n,p)$ is fairly well (but still not perfectly) understood [7, 11].

Since we tend to work by analogy with the $G(n,p)$ theory, it is a very challenging problem to seek a high-dimensional counterpart to the *phase transition* that occurs at $p = \frac{1}{n}$. It is here that the random graph a.a.s. acquires cycles. Namely, for every $0 < c < 1$ there is a $0 < q = q(c) < 1$ such that a graph in $G(n, \frac{c}{n})$ is a forest with probability $q + o_n(1)$, but for $p \geq \frac{1}{n}$, a $G(n,p)$ graph has, a.a.s. at least one cycle. These notions have natural analogs in higher-dimensional complexes that suggest what is being sought. However, even more famously, a *giant* connected component with $\Omega(n)$ vertices emerges at $p = \frac{1}{n}$. Since there is no natural notion of connected components at dimensions $d > 1$, it is not even clear what to ask. Finding the correct framework for asking this question and discovering the answer is indeed one of the main accomplishments of the research that we survey here.

Another reason that makes the high-dimensional scenario more complicated than the graph-theoretic picture is that there are several natural analogs for acyclicity. A $(d-1)$-face τ in a d-dimensional complex Y is *free* if it is contained in exactly one d-dimensional face σ of Y. In the corresponding *elementary collapse* step, which is a special case of homotopy equivalence, τ and σ are removed from Y. We say that Y is *d-collapsible* if it is possible to eliminate all its d-faces by a series of elementary collapses. Otherwise, the maximal subcomplex of Y in which all $(d-1)$-faces are contained in at least two d-faces is called the *core* of Y. Note that a graph (i.e., a 1-dimensional complex) is 1-collapsible if and only if it is acyclic, i.e., a forest.

A d-dimensional complex Y is said to be *d-acyclic* if its d-th homology group vanishes. Namely, if the d-dimensional boundary matrix $\partial_d(Y)$ has a trivial right kernel. Unless otherwise stated, we consider this matrix over the reals. The real d-Betti number of Y is $\beta_d(Y;\mathbb{R}) := \dim H_d(Y;\mathbb{R}) = \dim(\ker \partial_d(Y))$.

Whereas acyclicity and 1-collapsibility are equivalent for graphs, this is no longer the case for d-dimensional complexes. Clearly, a d-collapsible simplicial complex has a trivial d-th homology, but the reverse implication does not hold in dimension $d \geq 2$.

In this view, there are now two potentially separate thresholds to determine in $Y_d(n,p)$: For d-collapsibility and for the vanishing of the d-th homology. These questions were answered and the respective thresholds were determined in a series of four papers. A lower bound on the threshold for d-collapsibility was found in [6] and a matching upper bound was proved in [4]. An upper bound on the threshold for the vanishing of the d-th homology was found in [5], with a recent matching

Table 1 The critical constants γ_d and c_d

d	2	3	4	5	10	100	1000
γ_d	2.455	3.089	3.509	3.822	4.749	7.555	10.175
c_d	2.754	3.907	4.962	5.984	$11 - 10^{-3.73}$	$101 - 10^{-41.8}$	$1001 - 10^{-431.7}$

lower bound for real homology [13]. We conjecture that the same bound holds for all coefficient rings, but this remains open at present. In the present article we note an error in the proof that is presented in [4], and we indicate how to overcome it. The main theorem in the paper is correct, and here we also derive some additional information on the face numbers of the core.

The purpose of this paper is to survey these results and present the main ingredients of the proofs. In particular, we highlight the key role of the local structure of random complexes in all these proofs.

Both thresholds are of the form $p = \frac{c}{n}$. Namely there is a constant $c = \gamma_d$ corresponding to the d-collapsibility threshold and $c = c_d$ for acyclicity. As functions of the dimension d, the constants γ_d and c_d differ substantially. Our results allow us to numerically compute them to desirable accuracy (See Table 1).

We briefly refer to a d-dimensional complexes as a d-complex. Before stating the theorems, a small technical remark is in order. An obvious obstacle for a d-complex Y to be either d-collapsible or d-acyclic is that it contains the boundary of a $(d+1)$-simplex $\partial\Delta_{d+1}$, i.e. all the $d+2$ d-faces that are spanned by some $d+2$ vertices. In the random complex $Y_d\left(n, \frac{c}{n}\right)$, these objects appear with probability bounded away from both zero and one, and it is easy to see that their number is Poisson distributed with a constant expectation. In particular, $Y_d\left(n, \frac{c}{n}\right)$ is $\partial\Delta_{d+1}$-free with positive probability. There are several ways to go around this technical difficulty. In [6] a model of random complexes conditioned on being $\partial\Delta_{d+1}$-free was considered, which allowed a cleaner form for the theorems. Here we work with the simple binomial model, and consequently must mention these simplices.

We turn to the main theorem. Let $d \geq 2$ be an integer, $c > 0$ real and denote the core of $Y = Y_d\left(n, \frac{c}{n}\right)$ by $\tilde{Y}$. We define γ_d as the minimum of the function $\psi(x) := -\frac{\ln x}{(1-x)^d}$, $0 < x < 1$. Furthermore, we let x_* be the unique root in $(0, 1)$ of $(d+1)(1-x) + (1+dx)\ln x = 0$, and $c_d := \psi(x_*)$.

In addition, for an integer k and $\lambda > 0$ real, we let $\Psi_k(\lambda) := \Pr[\mathrm{Poi}(\lambda) \geq k]$, and $t = t(c, d)$ be the smallest positive root of

$$t = e^{-c(1-t)^d}, \tag{1}$$

or equivalently $1 - t = \Psi_1(c(1-t)^d)$.

Theorem 1.1 *Let $d \geq 2$ be an integer, $c > 0$ real, and $Y = Y_d\left(n, \frac{c}{n}\right)$.*

(I) The collapsible regime: If $c < \gamma_d$ then a.a.s. either Y is d-collapsible or its core is comprised of $O_d(1)$ vertex disjoint $\partial\Delta_{d+1}$'s.

(II) The intermediate regime: If $\gamma_d < c < c_d$ then a.a.s.

(a) Y is not d-collapsible. Moreover, its core $\tilde{Y}$ contains a constant fraction of the $(d-1)$-faces:

$$f_{d-1}(\tilde{Y}) = \Psi_2(c(1-t)^d)\binom{n}{d}(1+o(1)), \;\; f_d(\tilde{Y}) = \frac{c}{n}\binom{n}{d+1}(1-t)^{d+1}(1+o(1)). \tag{2}$$

In particular, $f_d(\tilde{Y}) < f_{d-1}(\tilde{Y})$.

(b) Either Y is d-acyclic or $H_d(Y;\mathbb{R})$ is generated by $O_d(1)$ vertex disjoint $\partial\Delta_{d+1}$'s.

(III) The cyclic regime: If $c > c_d$ then a.a.s. $H_d(Y;\mathbb{R})$ is non-trivial. Furthermore, $f_{d-1}(\tilde{Y})$ and $f_d(\tilde{Y})$ still satisfy equation (2), but in this regime $f_{d-1}(\tilde{Y}) < f_d(\tilde{Y})$ and

$$\beta_d(Y) = \left(\frac{c}{d+1}(1-t)^{d+1} - (1-t) + ct(1-t)^d\right)\binom{n}{d}(1+o(1))$$
$$= \big(f_d(\tilde{Y}) - f_{d-1}(\tilde{Y})\big)(1+o(1)).$$

See an illustration of Theorem 1.1 for $d = 2$ in Fig. 1. It is not hard to determine the asymptotic behaviour (in d) of these expressions, namely,

$$c_d = (d+1)(1-e^{-(d+1)}) + O_d(d^3e^{-2d})$$

and

$$\gamma_d = (1+o_d(1))\ln d.$$

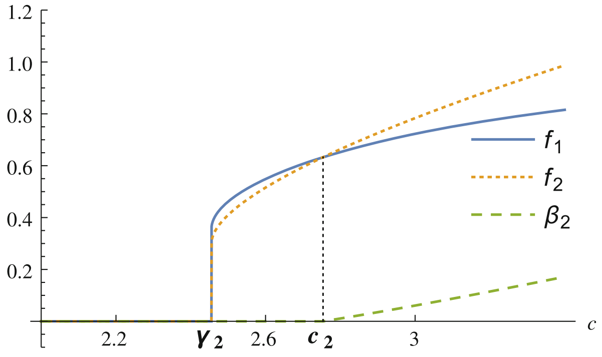


Fig. 1 Illustration of Theorem 1.1 for $d = 2$. Here f_1, f_2 are the face numbers of $\tilde{Y}$ and β_2 is the second Betti number. The functions are normalized by $\binom{n}{2}$, and $n \to \infty$

Consequently, there is a wide range of the parameter $p = p(d, n)$ for which almost all the complexes in $Y_d(n, p)$ are acyclic and non-collapsible.

Note that $f_d(\tilde{Y}) - f_{d-1}(\tilde{Y}) > 0$ implies that $H_d(Y; \mathbb{R}) \neq 0$. Moreover, this difference of the face numbers is a lower bound for the d-th Betti number. Theorem 1.1 shows that for all $0 \leq p \leq 1$, and up to the appearance of $\partial\Delta_{d+1}$'s these two conditions are typically equivalent and the lower bound is asymptotically tight. This is clearly a probabilistic statement which does not hold in general.

We turn to deal with the emergence of the giant component, a subject on which there exists an extensive body of literature. As mentioned above, there is no obvious high-dimensional counterpart to the notion of connected components, and we need a conceptual idea in order to even get started. The notion of *shadows*, introduced in [14], offers a way around this difficulty. The idea is to tie connected components with cycles, which do have natural high dimensional counterparts. The shadow of a graph G is the set of those edges that are not in G, whose addition to G creates a new cycle. It turns out that the giant component emerges exactly when the shadow of the evolving random graph acquires positive density. In particular, for $c > 1$ the shadow of $G(n, \frac{c}{n})$ has density $(1-t)^2 + o(1)$, where t is the unique root in $(0, 1)$ of $t = e^{-c(1-t)}$ (See Fig. 2a).

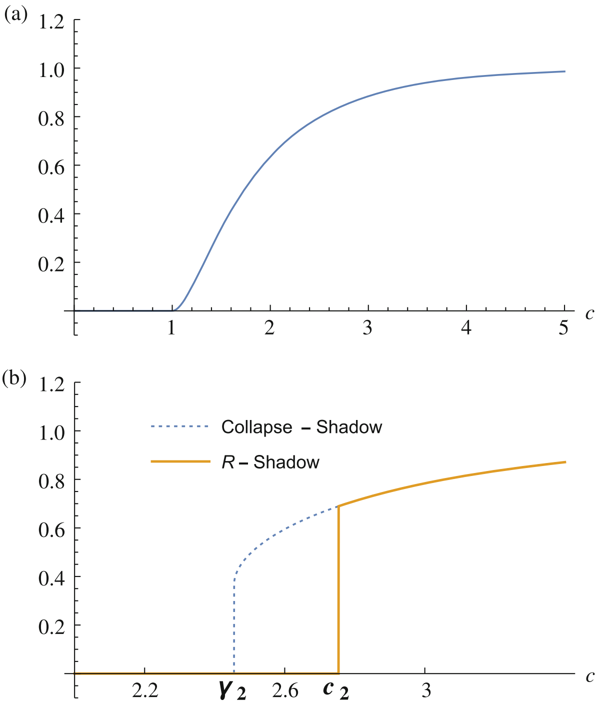


Fig. 2 Illustration of Theorem 1.2 for $d = 2$, and comparison to the density of the shadow of a random graph. (**a**) Density of the shadow of $G(n, \frac{c}{n})$. (**b**) Density of the C-shadow and $\mathbb{R}$-shadow of $Y_2\left(n, \frac{c}{n}\right)$

This suggests how we should define $\mathrm{SH}_{\mathbb{R}}(Y)$, the shadow of Y, a d-dimensional complex with full $(d-1)$-skeleton. Namely, it is the following set of d-faces:

$$\mathrm{SH}_{\mathbb{R}}(Y) = \{\sigma \notin Y : H_d(Y;\mathbb{R}) \text{ is a proper subspace of } H_d(Y \cup \{\sigma\};\mathbb{R})\}.$$

In words, a d-face belongs to $\mathrm{SH}_{\mathbb{R}}(Y)$ if it is not in Y and its addition to Y creates a new d-cycle.

We are considering throughout the vanishing of the d-th homology and d-collapsibility. These two notions capture acyclicity from an algebraic, respectively, combinatorial, perspective. For $d = 1$ the two coincide, but they differ widely for $d \geq 2$. This dual perspective carries over to two notions of shadows. A d-face σ that does not belong to a d-complex Y is in Y's $\mathbb{R}$-shadow if its addition to Y increases the d-homology. It is in the *C-shadow* of Y if its addition to Y increases the core. Again, these notions coincide for $d = 1$, and the $\mathbb{R}$-shadow is always contained in the C-shadow, but for $d \geq 2$ they may differ.

The notion of shadows lets us compare the phase transitions of random graphs and random complexes of higher dimensions, and a substantial qualitative difference reveals itself. While the density of the shadow of $G(n,p)$ undergoes a smooth transition around $p = 1/n$, when $d \geq 2$ both the C-shadow and the $\mathbb{R}$-shadow of $Y_d\left(n, \frac{c}{n}\right)$ undergo *discontinuous* first-order phase transitions at the critical points γ_d and c_d respectively.

Theorem 1.2 *Let $d \geq 2$ be an integer, $c > 0$ real, and $Y = Y_d\left(n, \frac{c}{n}\right)$.*

(I) The collapsible regime: If $c < \gamma_d$ then a.a.s. $|SH_{\mathbb{R}}(Y)| \leq |SH_C(Y)| = \Theta(n)$.
(II) The intermediate regime: If $\gamma_d < c < c_d$ then a.a.s. $|SH_{\mathbb{R}}(Y)| = \Theta(n)$, and

$$|SH_C(Y)| = \binom{n}{d+1}((1-t)^{d+1} + o(1)).$$

(III) The cyclic regime: If $c > c_d$ then a.a.s. the size of both $SH_{\mathbb{R}}(Y)$ and $SH_C(Y)$ is $\binom{n}{d+1}((1-t)^{d+1} + o(1))$.

An essential idea that is common to all these results is that in the range $p = \Theta(\frac{1}{n})$ many of the interesting properties of $Y_d(n,p)$ can be revealed by studying its *local* structure. Initially, this seemed as merely a useful tool in studying the threshold for d-collapsibility, and in establishing an upper bound on the threshold of the vanishing of the d-th homology. However, in obtaining a lower bound on this threshold, it became apparent that this idea should be viewed in the wider context of *local weak limits*. This framework was introduced by Benjamini and Schramm [8] and Aldous and Steele [3]. In recent years, this approach was used in deriving new asymptotic results in various fields of mathematics (e.g. [1, 16]).

The study of d-collapsibility in random complexes was significantly influenced by work on k-cores in random hypergraphs and specifically the works by Molloy [18] and Riordan [20]. Also, the proof of Theorem 1.1 makes substantial use of tools from the paper of Bordenave, Lelarge and Salez [9] on the rank of the adjacency matrix of random graphs.

The rest of the paper is organized as follows. Section 2 gives some necessary background material about simplicial complexes. In Sect. 3 we introduce the concept of a Poisson d-tree which is the local weak limit of random simplicial complexes. The main ingredients of the proofs of the main theorems are presented in Sects. 4.1 and 5.1, that respectively addressing the subjects of collapsibility and acyclicity. Concluding remarks and open questions are presented in Sect. 6.

2 Preliminaries

A simplicial complex Y is a collection of subsets of its *vertex set* V that is closed under taking subsets. Namely, $\sigma \in Y$ and $\tau \subseteq \sigma$ imply that $\tau \in Y$ as well. Members of Y are called *faces* or *simplices*. The *dimension* of the simplex $\sigma \in Y$ is defined as $|\sigma| - 1$, and $\dim(Y)$ is defined as $\max \dim(A)$ over all faces $A \in Y$. A d-dimensional simplex is also called a d-simplex or a d-face, and a d-dimensional simplicial complex is also referred to as a d-complex. The set of j-faces in Y is denoted by Y_j, and the face numbers by $f_j(Y) := |Y_j|$. For $t < \dim(Y)$, the *t-skeleton* of Y is the simplicial complex that consists of all faces of dimension $\leq t$ in Y, and Y is said to have a *full* t-dimensional skeleton if its t-skeleton contains all the t-faces of V. In this paper, the *degree* $d_Y(\tau)$ of a face τ in a complex Y is the number of $\dim(Y)$-faces that contain it. A face of degree zero is said to be *exposed*. Although we are directly interested only in finite complexes, infinite ones do play a role here, but we consider only *locally-finite* complexes in which every face has a finite degree. We occasionally use the bipartite incidence graph between $(d-1)$-faces and d-faces of a d-complex Y. This allows us, in particular, to speak about *distances* among such faces.

The permutations on the vertices of a face σ are split in two *orientations*, according to the permutation's sign. The *boundary operator* $\partial = \partial_d$ maps an oriented d-simplex $\sigma = (v_0, \ldots, v_d)$ to the formal sum $\sum_{i=0}^{d}(-1)^i(\sigma^i)$, where $\sigma^i = (v_0, \ldots v_{i-1}, v_{i+1}, \ldots, v_d)$ is an oriented $(d-1)$-simplex. We fix some commutative ring R and linearly extend the boundary operator to free R-sums of simplices. We denote by $\partial_d(Y)$ the d-dimensional boundary operator of a d-complex Y.

When Y is finite, we consider the $f_{d-1}(Y) \times f_d(Y)$ matrix form of ∂_d by choosing arbitrary orientations for $(d-1)$-simplices and d-simplices. Note that changing the orientation of a d-simplex (resp. $d-1$-simplex) results in multiplying the corresponding column (resp. row) by -1.

The d-th homology group $H_d(Y; R)$ (or vector space if R is a field) of a d-complex Y is the (right) kernel of its boundary operator ∂_d. Most of the homology groups in this paper are considered over $\mathbb{R}$. An element in $H_d(Y; \mathbb{R})$ is called a *d-cycle*, and the whole group is called the d-cycle space of Y. The *d-th Betti number* $\beta_d(Y; \mathbb{R})$ of a complex Y is defined to be the dimension of $H_d(Y; \mathbb{R})$.

Recall the concept of elementary collapse as defined in the introduction. A d-collapse phase is a procedure in which all the possible elementary d-collapses take place at once. In case more than one $(d-1)$-face can collapse some d-face, one of them is chosen by some predetermined arbitrary criteria. Given a d-complex Y, the complex $R_k(Y)$ is the complex that is obtained from Y after k phases of d-collapse. Similarly, $R_\infty(Y)$ is obtained after all possible d-collapse steps are carried out. A *d-core* (or core, for brevity) is a d-complex in which all the $(d-1)$-faces are of degree ≥ 2. The *d-core of Y* is the maximal d-core subcomplex of Y. Note that the d-core of Y is obtained from $R_\infty(Y)$ by removing the exposed $(d-1)$-faces.

3 Poisson *d*-Tree

The concept of a Poisson d-tree process was introduced in [6] and turned out to be extremely useful in the study of random simplicial complexes. It can be viewed as a high-dimensional counterpart of the Poisson Galton-Watson process which plays a key role in the study of the giant component in $G(n,p)$ graphs.

A *rooted d-tree* is a pair (T,o) where T is a d-complex and o is some $(d-1)$-face of T. A d-tree is generated by the following process. Initially the complex consists of the $(d-1)$-face o. At every step $k \geq 0$, every $(d-1)$-face τ of distance k from o picks a non-negative number $m = m_\tau$ of new vertices $v_1, \ldots, v_m$, and adds the d-faces $v_1\tau, \ldots, v_m\tau$ to T.

We use some self-explanatory terminology in our study of d-trees. A *leaf* is a $(d-1)$-face with no *descendant* d-faces. If (T,o) is a rooted d-tree and T' is a subtree of T which contains the root o we refer to (T',o) as a *rooted subtree*. If τ is a $(d-1)$-face of T, the *branch of T rooted at τ* is the subtree that contains τ and all its descendants. A $(d-1)$-face τ is an *ancestor* of a $(d-1)$-face τ' if τ' belongs to the *branch rooted at τ*. The *depth* of a $(d-1)$-face is its distance from the root, and the depth of the d-tree is the maximal depth of any of its $(d-1)$-faces.

A *Poisson d-tree with parameter c*, denoted by $T_d(c)$, is a rooted d-tree in which all the numbers m_τ throughout this generative process are i.i.d. *Poi*(c)-distributed. The rooted subtree of $T_d(c)$ that consists of the first k generations of this process is denoted $T_{d,k}(c)$.

The most important fact about $T_d(c)$ in this context is that it approximates the local neighborhood of a $(d-1)$-face in $Y_d\left(n, \frac{c}{n}\right)$. This fact is well-known and very useful in the Erdős-Rényi random graphs.

Lemma 3.1 ([6]) *For every fixed integer $k > 0$, the k-neighborhood of a fixed $(d-1)$-face τ in $Y = Y_d\left(n, \frac{c}{n}\right)$ converges in distribution to the k-neighborhood of the root of $T_d(c)$ as $n \to \infty$.*

Proof First, we observe that a.a.s. the k-neighborhood of a fixed $(d-1)$-face τ is a d-tree. Indeed, the violation of this statement requires that at least $k+1$ d-faces are spanned in Y by τ's fixed vertices and k additional ones. The fact that this event is negligible follows from a first moment argument. In addition, conditioned on the

k-neighborhood being a d-tree, the number m_η of new vertices that a $(d-1)$-face η adds to this d-tree is Bin $\left(n-o(n), \frac{c}{n}\right)$-distributed, which tends to Poisson with parameter c as $n \to \infty$. □

This lemma easily implies convergence of $Y_d\left(n, \frac{c}{n}\right) \to T_d(c)$ in the sense of local weak convergence introduced by Benjamini and Schramm [8] and Aldous and Steele [3]. Here is a brief explanation of this concept. A rooted d-complex is a pair (Y, τ) of a d-complex and some $(d-1)$-face in it. We denote by $(Y, \tau)_k$ the τ-rooted subcomplex of (Y, τ) comprised of all the d-faces of distance at most k from τ and their subfaces. Let $\mathcal{Y}_d$ be the set of all (isomorphism types of) rooted d-complexes, equipped with the metric

$$\operatorname{dist}((Y, \tau), (Y', \tau')) = \inf\left\{\frac{1}{t+1} \;:\; (Y, \tau)_k \cong (Y', \tau')_k\right\}.$$

It can be easily verified that $(\mathcal{Y}_d, \operatorname{dist})$ is a separable and complete metric space, which comes, as usual, equipped with its Borel σ-algebra (See [2]). The fact that $Y_d\left(n, \frac{c}{n}\right)$ converges to $T_d(c)$ means that for every bounded and continuous function $f : \mathcal{Y}_d \to \mathbb{R}$,

$$\mathbb{E}_{Y=Y_d\left(n,\frac{c}{n}\right)}[f(Y, \tau)] \xrightarrow[n\to\infty]{} \mathbb{E}_{T=T_d(c)}[f(T, o)],$$

for every fixed $(d-1)$-face τ and the root o of T.

As we explain below, this fact will be applied directly to a function of particular interest in this context, namely, the degree of the root τ after k phases of τ-rooted collapse. In addition, it will be used in combination with the spectral theorem to bound the Betti numbers of $Y_d\left(n, \frac{c}{n}\right)$ with the spectral measure of the Poisson d-tree.

3.1 Rooted Collapse

Let Y be a d-complex and τ some $(d-1)$-face of Y. A *τ-rooted collapse* of Y is a d-collapse process in which we forbid to collapse τ. Let k be a non-negative integer. The complex obtained from Y after k phases in the τ-rooted collapse process is denoted by $R_k(Y, \tau)$.

In the case $Y = Y_d\left(n, \frac{c}{n}\right)$, the degree $d_{R_k(Y,\tau)}(\tau)$ turns out to be relevant to several different questions. We approximate it using $\delta_k := d_{R_k(T,o)}(o)$, where $T = T_d(c)$, the Poisson d-tree with root o.

Lemma 3.2 *With the above notations*

$$\mathbb{E}[d_{R_k(Y,\tau)}(\tau)] \xrightarrow[n\to\infty]{} \mathbb{E}[\delta_k].$$

Note that this is not a direct corollary of the local weak convergence. Even though the function $d_{R_k(Y,\tau)}(\tau)$ is continuous, being dependent only on some fixed neighborhood of the root, it is not bounded. Nevertheless, we allow ourselves to omit the proof, since this difficulty can be bypassed by a simple calculus trick. Namely, by considering the function $\min\{d_{R_k(Y,\tau)}(\tau), A\}$, where A is a sufficiently large constant.

Lemma 3.3 *Let $c > 0$ and $(t_k)_{k \geq -1}$ a sequence of real numbers defined by*

$$t_{-1} = 0 \ , \quad t_{k+1} = e^{-c(1-t_k)^d}, \ \forall k \geq 0.$$

Then, δ_k is Poisson distributed with parameter $c(1 - t_{k-1})^d$, for every $k \geq 0$.
We refer throughout the paper to the sequences t_k of real numbers and δ_k of random variables that are defined here without denoting the underlying parameter $c > 0$ that is clear from the context.

Proof By induction on k. The case $k = 0$ is trivial since δ_0 is Poisson distributed with parameter c. For the induction step, let us consider the distribution of δ_k. A d-face σ that contains the root o survives k phases of rooted collapse if and only if each of its $(d-1)$-faces τ other than the root (there are d such τ's) is contained in a d-face other then σ after $k-1$ phases. This occurs if and only if τ has a positive degree in the branch of T rooted at τ after $k-1$ phases of τ-rooted collapse. Since this subtree is also a Poisson d-tree with parameter c, this occurs, by the induction hypothesis, with probability $\Pr[\delta_{k-1} > 0] = 1 - t_{k-1}$. Moreover, different branches of the tree are independent, so these events for different τ's and σ's are independent. Namely, the distribution of δ_k is a Binomial distribution with $\delta_0 = \text{Poi(c)}$ trials and success probability $(1-t_{k-1})^d$. By a standard computation in probability theory, this implies that δ_k is Poisson distributed with parameter $c(1 - t_{k-1})^d$. □

We say that *a rooted d-tree is collapsible* if its root gets exposed in the rooted collapse process. For instance, the previous lemma shows that the probability that $T_d(c)$ is collapsed after k phases is t_k, and the probability that $T_d(c)$ is collapsible is $t = t(c, d)$.

4 *d*-Collapsibility

The behavior of the sequence (t_k) of Lemma 3.3 changes quite substantially when $c = \gamma_d$. A simple calculus exercise tells us that $\lim_{k\to\infty} t_k = t(c, d)$, which is defined in (1). In addition, $t = t(c, d)$ equals to 1 if $c < \gamma_d$, and if $c > \gamma_d$ then t is strictly smaller then 1.

In other words, if $c < \gamma_d$ then the root of $T_d(c)$ gets exposed after k collapse phases with probability $1-o_k(1)$. Moreover, the expected degree of the root is $o_k(1)$. On the other hand, if $c > \gamma_d$ then, for arbitrarily large k, with probability bounded away from zero some d-faces that contain the root will survive the collapse process.

How do these facts reflect on the behavior of the random simplicial complex $Y = Y_d\left(n, \frac{c}{n}\right)$ under d-collapse phases? Many parameters of $R_k(Y)$ can be understood almost directly from the τ-rooted collapse of Y, where τ is a typical $(d-1)$-face. Moreover, the Poisson d-tree plays a key role here since k phases of τ-rooted collapse depend only on the k-neighborhood of τ. Consequently, as k grows, almost all $(d-1)$-faces of Y will either collapse or become exposed in $R_k(Y)$ if $c < \gamma_d$. On the other hand, if $c > \gamma_d$, a constant fraction of the $(d-1)$-faces survive k phases of d-collapse, but only very few of them remain with degree 1, giving the collapse process a slim chance to continue much further. In fact, the fraction of the $(d-1)$-faces that are contained in Y's core is asymptotically approximated by the probability that the root of $T_d(c)$ has degree ≥ 2 after infinitely many collapse phases.

While the transition from the Poisson d-tree to the random simplicial complex is straightforward in the subcritical regime, in the supercritical regime we follow an involved argument of Riordan [20] for k-cores of random graphs. The reader is encouraged to read the introduction of Riordan's paper for an intuitive discussion of the proof method.

4.1 The Collapsible Regime: Theorem 1.1 (I)

Let $Y = Y_d\left(n, \frac{c}{n}\right)$, $c < \gamma_d$, and let τ be some $(d-1)$-face in Y.

$$\begin{aligned}
\mathbb{E}[f_d(R_\infty(Y))] &\leq \mathbb{E}[f_d(R_k(Y))] \\
&= \frac{1}{d+1}\mathbb{E}\left[\sum_{\tau\in Y_{d-1}} \mathbf{1}_{\tau\in R_k(Y)}\cdot d_{R_k(Y)}(\tau)\right] \\
&= \frac{1}{d+1}\binom{n}{d}\mathbb{E}\left[\mathbf{1}_{\tau\in R_k(Y)}\cdot d_{R_k(Y)}(\tau)\right] \qquad (3)\\
&\leq \frac{1}{d+1}\binom{n}{d}\mathbb{E}\left[d_{R_k(Y,\tau)}(\tau)\right] \qquad (4)\\
&= \frac{1}{d+1}\binom{n}{d}(1+o(1))\mathbb{E}[\delta_k] \\
&= \frac{1}{d+1}\binom{n}{d}(1+o(1))c(1-t_{k-1})^d.
\end{aligned}$$

Identity (3) is obtained by considering some fixed $(d-1)$-face τ, using linearity of expectation and symmetry. The subsequent inequality (4) is due to the fact that in the τ-rooted collapse process fewer collapses occur than in d-collapse phases, whence an inequality $d_{R_k(Y)}(\tau) \leq d_{R_k(Y,\tau)}(\tau)$ between τ's degrees after k phases in either d-collapse processes. The following equations are straightforward applications of the lemmas in Sect. 3.

Consequently,$f_d(R_\infty(Y)) = o(n^d)$ a.a.s. for every $c < \gamma_d$.

The argument which completes the proof says that for every $c > 0$, the complex $Y_d\left(n, \frac{c}{n}\right)$ has no core subcomplex with $o(n^d)$ d-faces, other than vertex disjoint $\partial\Delta_{d+1}$'s. This is proved in Theorem 4.1 of [6], concerning inclusion-minimal core complexes. It turns out that a slight modification of that proof yields a more general conclusion.

A d-complex whose d-faces are comprised of a vertex-disjoint union of boundaries of $(d+1)$-simplices is called here a *d-gravel.*

Lemma 4.1 *For every integer $d \geq 2$ and real $c > 0$ real there is $\alpha > 0$ such that a.a.s. the following holds. Let $Y = Y_d\left(n, \frac{c}{n}\right)$, then either $R_\infty(Y)$ is a d-gravel or $f_d(R_\infty(Y)) > \alpha n^d$.*

Proof Let $m_1 := (d^3 \log n)^d$. Our first goal is to show that every core d-subcomplex C of Y with $f_d(C) \leq m_1$ is a d-gravel. A simple first moment argument yields that Y cannot contain two intersecting copies of $\partial\Delta_{d+1}$, nor can it contain more than $\log\log n$ copies of $\partial\Delta_{d+1}$. A core d-complex C that is comprised of exactly l vertex disjoint $\partial\Delta_{d+1}$'s and m additional d-faces is said to have *type* (l, m).

Let C have type (l, m). We partition its vertex set into $S \dot{\cup} T$, where S is the set of the vertices in some $\partial\Delta_{d+1}$ of C. Let $T' \subseteq T$ be those vertices in T of degree $d+1$ in C (i.e., such a vertex is in exactly $d+1$ of C's d-faces). Since C is a core, the degree of every vertex in C is at least $d+1$.

In addition, every d-face of C with two or more vertices of degree $d+1$ is included in a $\partial\Delta_{d+1}$. Recall the notion of a *link of a vertex* v in a simplicial complex Y. Namely, $\mathrm{lk}_Y(v) = \{\tau \in Y : v \dot{\cup} \tau \in Y\}$. In particular, the link of a vertex in a d-core is a $(d-1)$-core. Therefore, if a vertex has degree $d+1$ in C then its link is a $\partial\Delta_d$. Suppose that the vertices $v, u \in C$ have degree $d+1$ in C and are contained in a common d-face $\sigma = (u, v, x_1, \ldots, x_{d-1})$. Their links are $\partial\Delta_d$'s so there exist vertices $u', v' \notin \sigma$ such that $\mathrm{lk}_C(u) = \partial(u', v, x_1, \ldots, x_{d-1})$ and $\mathrm{lk}_C(v) = \partial(u, v', x_1, \ldots, x_{d-1})$. This can occur only if $u' = v'$ and the claim follows.

As a result, every non-gravel d-face contains at most one vertex of T', so that $m \geq |T'|(d+1)$. Counting incidences of vertices and d-faces in C yields

$$\big(m + l(d+2)\big) \cdot (d+1) \geq (|S| + |T|)(d+1) + (|T| - |T'|).$$

But $|S| = l(d+2)$, and a simple manipulation of these inequalities gives $|T| \leq m \cdot \frac{d+3}{d+4}$. We can assume w.l.o.g. that $m \leq m_1,\ l \leq \log\log n$ and derive the following upper bound on the number of type (l, m) core d-complexes with at most n vertices

$$n^{l(d+2)+\frac{d+3}{d+4}m} \cdot (|S| + |T|)^{(d+1)m} = n^{l(d+2)} \cdot \left[n^{\frac{d+3}{d+4}} \cdot O(\log^{d(d+1)} n)\right]^m.$$

The first term counts the choices for S, T, and the second the choice of non-gravel d-faces. We conclude that a.a.s. Y contains no core of type (l, m) with $l < \log\log n$ and $0 < m \leq m_1$. This is because any subcomplex of type (l, m) appears in Y with probability $(c/n)^{l(d+2)+m}$.

The proof of Theorem 4.1 in [6] yields a constant $\alpha = \alpha(c, d)$ such that a.a.s. $Y_d\left(n, \frac{c}{n}\right)$ has no inclusion-minimal subcomplex that is a core with $m_1 \leq m \leq \alpha n^d$ d-faces. In fact, the argument presented at [6] only uses the fact that a minimal core C is *connected* in the sense that between every two $(d-1)$-faces τ, τ' in C there is a path alternating between $(d-1)$-faces and d-faces of C with an inclusion relation. However, since every core is a union of connected cores, this means that there are no cores of size m in Y. It follows that the only possible cores that Y can contain have type $(l, 0)$, i.e., it is a d-gravel. □

4.2 The Core of $Y_d\left(n, \frac{c}{n}\right)$: Theorem 1.1 (II.a)

The proof of this theorem closely follows the argument of Riordan [20]. We fix the dimension d and refer to $r(c) := 1 - t(c, d)$ as a function of c. For brevity we denote $r^+(c) := \Psi_2(cr(c)^d)$. Note that both $r(c)$ and $r^+(c)$ are continuous, bounded away from 0 and increasing when $c > \gamma_d$. Our main goal is to show that for every $\tilde{c} > \gamma_d$ and $\varepsilon > 0, f_{d-1}(\tilde{Y}) > (r^+(\tilde{c}) - \varepsilon)\binom{n}{d}$, where $\tilde{Y}$ is the d-core of $Y_d\left(n, \frac{\tilde{c}}{n}\right)$.

This is motivated by the fact that $r^+(\tilde{c})$ is the probability that the root's degree is ≥ 2 after every finite number of rooted collapse phases in $T_d(\tilde{c})$. In other words, this is the probability that the root survives the non-rooted collapse process. Although this argument is simple and appealing, the actual proof is substantially more involved. Our strategy is to define some carefully crafted property $\mathcal{A}$ of d-trees of depth $\log\log n \ll S = S(n) \ll \log n$ such that the following two statements hold a.a.s. First, the subset $A \subset Y_{d-1}$ of $(d-1)$-faces τ such that Y contains a τ-rooted d-tree with property $\mathcal{A}$ is of density at least $(r^+(\tilde{c}) - \varepsilon)$. Second, for every $\tau \in A$ there exists a d-tree $T_\tau \subset Y$ in which τ's degree is at least 2 and every leaf also belongs to A. Consequently, no $(d-1)$-face in A can be collapsed.

We refer throughout the proof to certain properties of rooted d-trees (T, o), and occasionally write $T \in \mathcal{P}$ to say that T has property $\mathcal{P}$. Every property $\mathcal{P}$ of rooted d-trees induces a property of $(d-1)$-faces in a d-complex Y. Namely, we say that the $(d-1)$-face τ *has property* $\mathcal{P}$ if Y contains a d-tree rooted on τ that has property $\mathcal{P}$.

Here are some relevant properties: $\mathcal{D}_{\leq L}$ means T has depth $\leq L$, and $\mathcal{D}_{<\infty}$ means that it has finite depth. Let $\mathcal{P}_1$ and $\mathcal{P}$ be properties of finite (resp. general) d-trees. A d-tree (T, o) has property $\mathcal{P}_1 \circ \mathcal{P}$ when: (i) T has a finite rooted subtree T' with property $\mathcal{P}_1$, and (ii) For every leaf τ of T', the branch of T rooted at τ has property $\mathcal{P}$. For example, we consider the property that T has depth $k + 1$ and it does not collapse in k phases, i.e., $\mathcal{B}_k := \{T \in \mathcal{D}_{\leq k+1} \mid \delta_k(T) > 0\}$ and note that $\mathcal{B}_k = \mathcal{B}_0 \circ \mathcal{B}_{k-1}$. Property $\mathcal{B}$ means that T does not collapse at finite time.

We also define for $k \geq 0$, the properties $\mathcal{R}_k$ which are stronger than $\mathcal{B}_k$ as follows

$$\mathcal{R}_0 = \{T \in \mathcal{D}_{\leq 1} \mid \delta_0(T) \geq 2\}\,, \quad \mathcal{R}_k = \mathcal{R}_0 \circ \mathcal{B}_{k-1} \cup \mathcal{B}_0 \circ \mathcal{R}_{k-1},\ k > 0.$$

The difference between $\mathcal{B}_k$ and $\mathcal{R}_k$ is this: $T \in \mathcal{B}_k$ means that T has depth $k+1$ and every non-leaf $(d-1)$-face has at least one descendant d-face. In defining $\mathcal{R}_k$ we add the requirement that along every root-to-leaf path we encounter at least one $(d-1)$-face with ≥ 2 descendant d-faces.

Finally, we introduce a stochastic version of $\mathcal{P}$, a property of finite rooted d-trees (T, o). For some $0 \leq p \leq 1$ we *mark* each leaf of T independently with probability p, and remove every d-face that contains any unmarked leaf. We say that the event $\mathcal{M}(\mathcal{P}, p, T)$ holds if the remaining d-tree has property $\mathcal{P}$. Marking is a convenient way of capturing the following phenomenon: We let each leaf in a finite d-tree grow a Poisson d-tree and we only ask whether or not this "tail" has some desired property. This is expressed in the simple identity

$$\Pr[\mathcal{M}(\mathcal{P}_1, p, T_{d,k}(c))] = \Pr[T_d(c) \in \mathcal{P}_1 \circ \mathcal{P}] \tag{5}$$

where d, k are integers, $c > 0$, $\mathcal{P}_1$ is a property of depth-k trees and $\mathcal{P}$ is a property of probability p for Poisson d-trees with parameter c. For instance,

$$\Pr[\mathcal{M}(\mathcal{B}_k, r(c), T_{d,k+1}(c))] = \Pr[T_d(c) \in \mathcal{B}] = r(c)$$

The following lemma can be viewed as a variation on this identity. It shows that although property $\mathcal{R}_k$ is stronger than $\mathcal{B}_k$, the two are almost equally likely in a Poisson d-tree.

Lemma 4.2 *For every $c > c_1 > \gamma_d$ there is a sufficiently large k such that*

$$\Pr[\mathcal{M}(\mathcal{R}_k, r(c_1), T_{d,k+1}(c))] > r(c_1).$$

Proof Consider the following probabilistic experiment where we randomize thrice. Initially we generate the first $k+1$ generations of $T_d(c)$. Then we do the random marking that yields the d-tree T. Finally we remove each d-face of T independently with probability c_1/c. We denote the component of the root by T'. Note that T' is distributed like a $T_{d,k+1}(c_1)$ to which random $r(c_1)$-marking is applied. In particular, $\Pr[T' \in \mathcal{B}_k] = r(c_1)$, hence we need to prove that $\Pr[T \in \mathcal{R}_k] > \Pr[T' \in \mathcal{B}_k]$. Since $T' \in \mathcal{B}_k$ implies that $T \in \mathcal{B}_k$ it suffices to show that $\Pr[W] > \Pr[L]$, where

$$W = [T \in \mathcal{R}_k,\ T' \notin \mathcal{B}_k] \text{ and } L = [T \in \mathcal{B}_k \setminus \mathcal{R}_k,\ T' \in \mathcal{B}_k].$$

To this end we show that $\Pr[L] \to 0$ as $k \to \infty$ whereas $\Pr[W]$ stays bounded away from 0.

Indeed, if $T \in \mathcal{B}_k \setminus \mathcal{R}_k$, then there exists some d-face of depth k whose removal violates property $\mathcal{B}_k$. This d-face survives in T' with probability $(c_1/c)^k$, so that

$\Pr[L] < (c_1/c)^k$. On the other hand W contains the event that $\delta_0(T) = 2$, $T \in \mathcal{R}_0 \circ \mathcal{B}_{k-1}$ and $\delta_0(T') = 0$ whose probability is positive and independent of k. □

A d-tree T of depth $L+1$ is *(p, η)-rigid* if $\Pr[\mathcal{M}(\mathcal{R}_L, p, T)] > 1 - \eta$.

Lemma 4.3 *For every $c > c_1 > \gamma_d$ and $\eta > 0$ and for a large enough integer L there holds*

$$\Pr[T_{d,L+1}(c) \textit{ is } (r(c), \eta)\textit{-rigid}] \geq r(c_1).$$

Proof Below we assume that k is large enough, as required in Lemma 4.2. We claim that

$$\Pr[T_d(c) \in \mathcal{R}_k \circ \mathcal{B}] = \Pr[\mathcal{M}(\mathcal{R}_k, r(c), T_{d,k+1}(c))] \geq \Pr[\mathcal{M}(\mathcal{R}_k, r(c_1), T_{d,k+1}(c))] > r(c_1).$$

The equality is a special case of identity (5). The first inequality follows by a simple monotonicity consideration and the fact that $r(c) > r(c_1)$. The last inequality comes from Lemma 4.2. The condition of (p, η)-rigidity is stronger the smaller η is, so we fix it to satisfy

$$\Pr[T_d(c) \in \mathcal{R}_k \circ \mathcal{B}] - r(c_1) > \eta.$$

For fixed k the conditions $T_{d,k+l+2}(c) \in \mathcal{R}_k \circ \mathcal{B}_l$ become more strict as l grows and their conjunction over all $l \geq 0$ is exactly the condition $T_d(c) \in \mathcal{R}_k \circ \mathcal{B}$. Therefore, we can and will choose l large enough so that

$$\Pr[T_{d,k+l+2}(c) \in \mathcal{R}_k \circ \mathcal{B}_l] \leq \Pr[T_d(c) \in \mathcal{R}_k \circ \mathcal{B}] + \eta^2.$$

For a d-tree T of depth $L+1 = k+l+2$, let $\phi(T) := \Pr[\mathcal{M}(\mathcal{R}_k \circ \mathcal{B}_l, r(c), T)]$. We denote by $\mathcal{L}$ the property that T is $(r(c), \eta)$-rigid. The expectation of $\phi(T)$, where T is $T_{d,L+1}(c)$ distributed, equals the probability $\Pr[T_d(c) \in \mathcal{R}_k \circ \mathcal{B}]$. In addition, $\phi(T) = 0$ if $T \notin \mathcal{R}_k \circ \mathcal{B}_l$ and $\phi(T) \leq 1 - \eta$ if $T \notin \mathcal{L}$ since $\mathcal{R}_k \circ \mathcal{B}_l$ implies $\mathcal{R}_L$. Therefore

$$\Pr[T_d(c) \in \mathcal{R}_k \circ \mathcal{B}] \leq \Pr[T_{d,L+1}(c) \in (\mathcal{R}_k \circ \mathcal{B}_l) \cap \mathcal{L}] + (1-\eta) \Pr[T_{d,L+1}(c) \in (\mathcal{R}_k \circ \mathcal{B}_l) \setminus \mathcal{L}],$$

whence

$$\eta \cdot \Pr[T_{d,L+1}(c) \in \mathcal{R}_k \circ \mathcal{B}_l \setminus \mathcal{L}] \leq \Pr[T_{d,L+1}(c) \in \mathcal{R}_k \circ \mathcal{B}_l] - \Pr[T_d(c) \in \mathcal{R}_k \circ \mathcal{B}].$$

Putting everything together we conclude that $\Pr[T_{d,L+1}(c) \in \mathcal{L}] \geq \Pr[T_{d,L+1}(c) \in \mathcal{R}_k \circ \mathcal{B}_l] - \eta > r(c_1)$, as stated. □

We set all the parameters that appear in the discussion below. Recall that our goal is to show that for every $\tilde{c} > \gamma_d$ and $\varepsilon > 0, f_{d-1}(\tilde{Y}) > (r^+(\tilde{c}) - \varepsilon)\binom{n}{d}$. Let $\gamma_d < c_1 < c < \tilde{c}$ such that $r^+(c_1) \geq r^+(\tilde{c}) - \varepsilon/2$ and $\Psi_1(\tilde{c}r(c_1)^d) > r(c)$. Again we choose k large enough to make Lemma 4.2 hold, and we fix some $0 < \eta < d^{-2(k+1)}/8$. Also

L is so large that Lemma 4.3 holds. Recall that $\mathcal{L}$ denotes the property $(r(c), \eta)$-rigidity of d-trees of depth $L + 1$.

Consider an integer $S = S(n)$ such that $\log\log n \ll S \ll \log n$, and we define for $0 \le s \le S$, the properties $\mathcal{A}_s$ by

$$\mathcal{A}_0 = \mathcal{L}\ , \quad \mathcal{A}_s = \mathcal{R}_k \circ \mathcal{A}_{s-1},\ 0 < s \le S.$$

Note that $\mathcal{A}_S$ depends on the first $Q := (k + 1)S + L + 1$ generations. Finally, we define two key properties:

$$\mathcal{A} = \mathcal{R}_0 \circ \mathcal{D}_{\le L} \circ \mathcal{A}_S\ , \quad \mathcal{P} = \mathcal{R}_0 \circ \mathcal{D}_{<\infty} \circ \mathcal{A}.$$

In words, $T \in \mathcal{A}$ means that T has a rooted subtree T' of depth $\le L + 1$ in which the root's degree is ≥ 2 and the branch of T rooted at every leaf of T' has property $\mathcal{A}_S$. $T \in \mathcal{P}$ means that T has a finite rooted subtree T' in which the root's degree is ≥ 2 and the branch of T rooted at every leaf of T' has property $\mathcal{A}$.

It follows by induction that $\Pr[T_{d,(k+1)s+L+1} \in \mathcal{A}_s] > r(c_1)$ for every $s \ge 0$. Indeed, Lemma 4.3 yields the case $s = 0$, and the inductive step follows from Lemma 4.2. Also, since $\mathcal{R}_0 \circ \mathcal{A}_S$ implies $\mathcal{A}$, the probability that $T_d(c)$ has a rooted subtree that satisfies $\mathcal{A}$ is at least

$$\Pr[T_{d,Q+1}(c) \in \mathcal{R}_0 \circ \mathcal{A}_S] \ge \Psi_2(cr(c_1)^d) \ge \Psi_2(c_1 r(c_1)^d) = r^+(c_1). \tag{6}$$

Indeed, the probability that in a d-face which contains the root all the branches rooted at a $(d - 1)$-face of depth 1 has $\mathcal{A}_S$ is at least $r(c_1)^d$. Hence, the number of such d-faces is $\mathrm{Poi}(cr(c_1)^d)$ distributed.

We come to the most significant step in the proof: We show, by a first-moment argument, that a.a.s. every $(d - 1)$-face in $Y = Y_d\left(n, \frac{\tilde{c}}{n}\right)$ that has property $\mathcal{A}$ also has $\mathcal{P}$. This implies that every such $(d-1)$-face τ is contained in a d-tree $T_\tau \subset Y$ in which (i) τ is of degree ≥ 2, and (ii) every leaf of T_τ has property $\mathcal{A}$. Consequently, the union of these d-trees $\{T_\tau \mid \tau \text{ has property } \mathcal{A}\}$ is contained in the core $\tilde{Y}$.

Claim 4.4 *For every $s \ge 0$, and every $T \in \mathcal{A}_s$,*

$$\Pr[\mathcal{M}(\mathcal{B}_{(k+1)s} \circ \mathcal{R}_L, r(c), T)] \ge 1 - \frac{2^{-2^s}}{4d^{2(k+1)}}.$$

Proof We proceed by induction on s. Our definition of $\mathcal{L}$ and the choice of η yield the case $s = 0$. Let $s \ge 0$ be an integer, and $T \in \mathcal{A}_{s+1} = \mathcal{R}_k \circ \mathcal{A}_s$. Let $T' \subset T$ be a *minimal* d-tree of depth $k + 1$ such that $T' \in \mathcal{R}_k$ and every branch of T rooted at a leaf of T' has the property $\mathcal{A}_s$. By a straightforward computation, T' has exactly $2d^{k+1}$ leaves. By induction, after the marking process the branch rooted at every leaf of T' fails to have property $\mathcal{B}_{(k+1)s} \circ \mathcal{R}_L$ independently with probability at most $\frac{2^{-2^s}}{4d^{2(k+1)}}$. Let us refer to such a leaf as *bad*, and remove from T' every d-face that contains a bad leaf. Since initially T' had the property $\mathcal{R}_k$, it now has property $\mathcal{B}_k$

unless at least 2 d-faces were removed. But this can only occur if at least 2 leaves are bad, an event of probability at most

$$\binom{2d^{k+1}}{2}\left(\frac{2^{-2^s}}{4d^{2(k+1)}}\right)^2 \leq \frac{2^{-2^{s+1}}}{4d^{2(k+1)}}.$$

Namely, the tree T has the property $\mathcal{B}_k \circ \mathcal{B}_{(k+1)s} \circ \mathcal{R}_L = \mathcal{B}_{(k+1)(s+1)} \circ \mathcal{R}_L$ with the desired probability. □

This leads to the following key lemma.

Lemma 4.5 *For every fixed $(d-1)$-face τ in $Y = Y_d\left(n, \frac{\tilde{c}}{n}\right)$, the probability that τ has property $\mathcal{A}$ but does not have $\mathcal{P}$ is $o(n^{-d})$.*

Proof It is easy to show that with probability $o(n^{-d})$ the $(1+L+2Q)$-neighborhood of τ consists of at most $n^{1/3}$ vertices, and we condition on this event. Suppose that τ has the property $\mathcal{A}$, and consider $\tau \in T \subset Y$ such that $T \in \mathcal{A}$. In particular, there exists a d-tree $T' \subset T$ rooted at τ of depth at most $L+1$ such that $d_{T'}(\tau) = 2$ and every branch $T''_\pi \subset T$ rooted at a leaf π of T' has property $\mathcal{A}_S$. Denote by $X \subset Y_{d-1}$ the union of the leaves of T''_π over all the leaves π of T'.

We now expose an additional subset of the $(Q+1)$-neighborhoods of the $(d-1)$-faces of X with the following precaution. When we reach some $(d-1)$-face ρ and query whether a d-face that contains it belongs to Y, we only consider d-faces of the form $v\rho$ where v is a vertex that does not belong to T nor did it appear in the exposing process upto the current query. In this manner, every $\rho \in X$ is the root of a d-tree $\tilde{T}_\rho \subset Y$ in which every $(d-1)$-face has at least $\mathrm{Bin}(n - n^{1/3}, \frac{\tilde{c}}{n})$ descendants. Therefore, the probability that $\tilde{T}_\rho \in \mathcal{B}_0 \circ \mathcal{A}_S$ is at least

$$\Pr[T_{d,Q}(\tilde{c}) \in \mathcal{B}_0 \circ \mathcal{A}_S] - o(1) \geq \Psi_1(\tilde{c}r(c_1)^d) - o(1) > r(c),$$

and these events are independent over $\rho \in X$. Therefore we can consider the event $\tilde{T}_\rho \in \mathcal{B}_0 \circ \mathcal{A}_S$ as an alternative for marking the leaves of d-trees T''_π and plug it in Claim 4.4. Since the number of leaves of T' is bounded, the claim implies that with probability $1 - O\left(2^{-2^S}\right) = 1 - o(n^{-d})$, after the described exposure of the additional neighborhoods, all the subtrees rooted in these leaves have the property $\mathcal{B}_{(k+1)S} \circ \mathcal{R}_L \circ \mathcal{B}_0 \circ \mathcal{A}_S$. But recall that $\mathcal{R}_L$ means that in every path from the root of the d-tree to a $(d-1)$-face of depth $L+1$, there is a $(d-1)$-face with at least two descendants. In other words, $\mathcal{R}_L \circ \mathcal{B}_0 \circ \mathcal{A}_S$ implies $\mathcal{D}_{\leq L} \circ \mathcal{R}_0 \circ \mathcal{D}_{\leq L} \circ \mathcal{A}_S = \mathcal{D}_{\leq L} \circ \mathcal{A}$. In particular, all the leaves of T' have the property $\mathcal{D}_{<\infty} \circ \mathcal{A}$, and since $d_{T'}(\tau) = 2$, it follows that τ has the property $\mathcal{P}$. □

We are now ready to prove the theorem.

Proof of Theorem 1.1 (II.a) Let $N_{\mathcal{A}}$ denote the number of $(d-1)$-faces in $Y = Y_d\left(n, \frac{\tilde{c}}{n}\right)$ that have property $\mathcal{A}$. By the previous discussion, we know that $f_{d-1}(\tilde{Y}) \geq N_{\mathcal{A}}$. We approximate the expectation of $N_{\mathcal{A}}$ by $\Pr_{\tilde{c}}[\mathcal{A}]$, the probability that $T_d(\tilde{c})$ has a rooted subtree that satisfies $\mathcal{A}$,

$$\mathbb{E}[N_{\mathcal{A}}] = (\Pr_{\tilde{c}}[\mathcal{A}] + o(1))\binom{n}{d} \geq r^+(c_1)\binom{n}{d} \geq (r^+(\tilde{c}) - \varepsilon/2)\binom{n}{d}$$

by Eq. (6). Since $\mathcal{A}$ depends only on the $O(S)$-neighborhood of the $(d-1)$-face, and two $(d-1)$-faces have non-disjoint neighborhoods with negligible probability, it follows that $\mathbb{E}[N_{\mathcal{A}}^2] = \mathbb{E}[N_{\mathcal{A}}]^2(1 + o(1))$. By the second moment method $\Pr[N_{\mathcal{A}} < (r^+(\tilde{c}) - \varepsilon)\binom{n}{d}] = o(1)$. The upper bound is much simpler. Let N_k denote the number of $(d-1)$-faces that survive (= did not collapse nor become free) the first k phases of collapse. Clearly, $f_{d-1}(\tilde{Y}) \leq N_k$ for every k. Similarly to $N_{\mathcal{A}}$, this property depends on the k-neighborhood of a $(d-1)$-face and by the same argument as before, N_k is concentrated around its expectation. The expectation of N_k can be bounded by the Poisson d-tree as follows.

$$\mathbb{E}[N_k] \leq (\Pr_{\tilde{c}}[\delta_{k-1} \geq 2] + o(1))\binom{n}{d} = (\Psi_2(\tilde{c}(1 - t_{k-2})^d) + o(1))\binom{n}{d},$$

and since $t_k \to t$, we obtain that $\Psi_2(\tilde{c}(1 - t_{k-2})^d)$ tends to $r^+(\tilde{c})$ as $k \to \infty$.

We turn to prove that a.a.s. the number of d-faces in the core $f_d(\tilde{Y}) = r(\tilde{c})^{d+1}\frac{\tilde{c}}{n}\binom{n}{d+1}(1 + o(1))$. Let $M_{\mathcal{A}}$ denote the number of d-faces in Y all of whose $(d-1)$-faces have property $\mathcal{A}$. Clearly, $f_d(\tilde{Y}) \geq M_{\mathcal{A}}$ since no $(d-1)$-face with property $\mathcal{A}$ is collapsed. In addition, since this is a local property it suffices, as before, to compute the expectation of $M_{\mathcal{A}}$. The probability that a d-simplex σ belongs to Y is $\frac{\tilde{c}}{n}$, and we can expose a subset T of its neighborhood in the same careful fashion as done in the proof of Lemma 4.5. The probability that all the d-trees growing from σ's $(d-1)$-faces have property $\mathcal{A}_S$ is at least $\Pr_{\tilde{c}}[\mathcal{A}_S]^{d+1} - o(1) > r(c_1)^{d+1}$. If this occurs, then every $(d-1)$-face $\tau \subset \sigma$ has property $\mathcal{A}$ by letting τ be the root of the d-tree $T \cup \{\sigma\}$. Consequently $\mathbb{E}[M_{\mathcal{A}}] \geq r(c_1)^{d+1}\frac{\tilde{c}}{n}\binom{n}{d+1}$. The upper bound is proved similarly, by showing that the probability that a d-face survives the first k collapse phases tends to $r(\tilde{c})^{d+1}$ as k grows. □

4.3 C-Shadow of $Y_d\left(n, \frac{c}{n}\right)$

Here we prove the parts of Theorem 1.2 that deal with the C-shadow of $Y_d\left(n, \frac{c}{n}\right)$. Namely, we show that for $c < \gamma_d$, the C-shadow of $Y = Y_d\left(n, \frac{c}{n}\right)$ has size $\Theta(n)$, and for $c > \gamma_d$ its size is

$$|\mathrm{SH}_C(Y)| = \binom{n}{d+1}((1-t)^{d+1} + o(1)).$$

Both statements follow directly from the previous proofs. Regarding the range $c < \gamma_d$, a simple second moment calculation shows that a.a.s. there are $\Theta(n)$ sets of $d+2$ vertices in Y that span all but one of the d-faces in the boundary of a $(d+1)$-simplex. The missing d-face in every such configuration is obviously in the C-shadow. On the other hand, if the C-shadow is large, viz., $|\mathrm{SH}_C(Y)| \gg n$, then for every $c < c' < \gamma_d$, with probability bounded away from zero, the core of $Y_d\left(n, \frac{c'}{n}\right)$ contains a complex that is not the boundary of $(d+1)$-simplex. But this contradicts Theorem 1.1(I).

We prove the supercritical case $c > \gamma_d$ in much the same way that we calculated the number of d-faces in the core. Namely, for the lower bound we count d-simplices not in Y all of whose $(d-1)$-faces have property $\mathcal{A}$. For the upper bound we count d-simplices that if added into Y do not survive k phases of collapse. As before, both properties are local and by a second moment argument are concentrated around their means, which are computed by Poisson d-tree approximations.

5 *d*-Acyclicity

In the previous section we saw that the threshold γ_d for d-collapsibility in $Y_d\left(n, \frac{c}{n}\right)$ coincides with the threshold in which rooted collapsibility in $T_d(c)$ almost surely eliminates all the d-faces containing the root. In the case of d-acyclicity, the correspondence is similar but more intricate. In fact, the threshold c_d for d-acyclicity coincides with *two* seemingly separate thresholds of $T_d(c)$'s parameters. These will used to bound the d-acyclicity threshold from below and above respectively. Since both occur at $c = c_d$ it follows that these bounds are tight. Furthermore, if $c > c_d$, these two parameters yield upper and lower bounds for $\beta_d(Y, \mathbb{R})$ which are tight upto small order error terms. Finally, the tight estimation for $\beta_d(Y, \mathbb{R})$ allows us to compute the density of the *shadow* of Y.

If a d-complex Y has more d-faces than $(d-1)$-faces, then $\beta_d(Y, \mathbb{R}) \geq f_d(Y) - f_{d-1}(Y) > 0$. For $Y = Y_d\left(n, \frac{c}{n}\right)$, this happens only when $c > d+1$, but we can say a bit more. Even though Y and its core $\tilde{Y}$ have the same d-th Betti number, it turns out that there is a wider range of the parameter c for which $\tilde{Y}$ has more d-faces than $(d-1)$-faces. In fact, one can show that $f_d(\tilde{Y}) > f_{d-1}(\tilde{Y})$ if and only if $c > c_d$, using the expressions for these face numbers in Theorem 1.1(II.a). However, it is significantly easier to prove the same lower bound on $\beta_d(Y)$ by analyzing $T_d(c)$ as follows. Let $S_k(Y)$ be obtained by removing all exposed $(d-1)$-faces in $R_k(Y)$. The average degree of the $(d-1)$ faces in $S_k(Y)$ is approximated using the conditional expectation

$$\mathbb{E}[\delta_k \mid \delta_k > 0 \ \wedge \ \delta_{k-1} > 1].$$

In words, this is the expected degree of the root of $T_d(c)$ after k phases of rooted collapses, conditioned on the fact that its degree remains strictly greater than 1 throughout the collapse process. We claim that this captures the average degree of

$(d-1)$-faces in $S_k(Y)$. A $(d-1)$-face τ of Y belongs to $S_k(Y)$ if and only if its degree after $k-1$ phases of τ-rooted collapse is larger than 1 and stays positive after one more phase. Indeed, as long as $d_\tau > 1$ the τ-rooted collapse and non-rooted collapse are identical. A difference occurs when $d_\tau = 1$, at which point the rooted collapse continues as usual, but the non-rooted collapse eliminates τ. If the average degree of $(d-1)$-faces exceeds $d+1$, this yields, via a simple double-counting argument, a positive lower bound for $\beta_d(Y)$. A simple calculus exercise then shows that this condition holds if and only if $c > c_d$. Namely,

$$\lim_{k\to\infty} \mathbb{E}[\delta_k \mid \delta_k > 0 \ \wedge\ \delta_{k-1} > 1] > d+1 \quad \Longleftrightarrow \quad c > c_d.$$

The most substantial role of local weak convergence is in proving the lower bound on the d-acyclicity threshold. We analyze $T = T_d(c)$ using tools from spectral theory and functional analysis. For this reason we are still unable to resolve this question over finite fields of coefficients. As further detailed below, we define $x_T := \pi_{L(T),e_o}(\{0\})$ to be the measure of the atom $\{0\}$ according to the *spectral measure* π of the *Laplacian* $L(T)$ of T with respect to the characteristic vector e_o of the root o. Local weak convergence implies that x_T *is an upper bound on the normalized dimension* of the left kernel Z of $\partial_d(Y)$. Note that if $Y_d\left(n, \frac{c}{n}\right)$ is d-acyclic then dim$Z$ equals $(1+o(1))\binom{n}{d}\left(1-\frac{c}{d+1}\right)$, and otherwise it is greater. Indeed, the proof shows that

$$\mathbb{E}_{T=T_d(c)}[x_T] = 1 - \frac{c}{d+1} \quad \Longleftrightarrow \quad c < c_d,$$

and if $c > c_d$, this expectation is greater than $1 - \frac{c}{d+1}$.

5.1 Acyclicity Beyond Collapsibility: Theorem 1.1(II.b)

To prove that a random complex is d-acyclic beyond the d-collapsibility threshold, we cannot restrict ourselves to purely combinatorial arguments. It is not a-priori clear that the local weak limit of a random complex holds enough information to prove such a statement. Surprisingly, perhaps, this is the case when we work over $\mathbb{R}$. In this section we describe the main ingredients of this method, which appears in [13], where complete proofs can be found.

Let $Y = Y_d\left(n, \frac{c}{n}\right)$. The primary goal in the proof is to find a tight upper bound for $\lim_{n\to\infty} \frac{1}{\binom{n}{d}}\mathbb{E}[\beta_d(Y)]$. It turns out more useful to work with the corresponding Laplace operator $L(Y) = \partial_d(Y)\partial_d(Y)^*$. We consider its kernel Z which coincides with the left kernel of $\partial_d(Y)$. Let $P_Z : \mathbb{R}^{Y_{d-1}} \to Z$ be the orthogonal projection to the space Z. By linear algebra,

$$\dim Z = \sum_{\tau \in Y_{d-1}} \|P_Z(e_\tau)\|^2,$$

where e_τ is the unit vector of τ.

The *spectral theorem* from functional analysis offers a new perspective of $\|P_Z(e_\tau)\|^2$. Associated with every self-adjoint operator L on a Hilbert space $\mathcal{H}$, and a vector $\psi \in \mathcal{H}$ is the *spectral measure of L with respect to* ψ. It is a real measure denoted $\pi_{L,\psi}$ which satisfies

$$\langle F(L)\psi, \psi\rangle = \int_{\mathbb{R}} F(x) d\pi_{L,\psi}(x),$$

for every measurable function $F : \mathbb{R} \to \mathbb{C}$. The operator $F(L)$ is uniquely defined by extending the action of polynomials on the operator L.

If $\mathcal{H}$ is finite-dimensional, $\pi_{L,\psi}$ is a discrete measure supported on the spectrum of L and $\pi_{L,\psi}(\lambda) = \|P_\lambda \psi\|^2$, where P_λ is the orthogonal projection to the λ-eigenspace.

We use this theorem with the measure $\pi_{L(Y),e_\tau}$. Here Y is a d-complex and $\mathcal{H} = \ell^2(Y_{d-1})$. The self adjoint operator is the Laplacian $L(Y)$, and e_τ is the characteristic vector of some $(d-1)$-face of Y.

In particular, with $Y = Y_d\left(n, \frac{c}{n}\right)$ and Z as before, $\|P_Z(e_\tau)\|^2$ is simply the measure of the atom $\{0\}$ according to the spectral measure $\pi_{L(Y),e_\tau}$.

The difficulty with applying the spectral theorem to the Poisson d-tree is that the degrees in this tree may be unbounded. We must, therefore, consider the subtleties of the theory of unbounded operators [19]. Briefly, the Laplacian $L(T)$ of an infinite d-tree T is a symmetric operator, directly defined on the dense subset of finitely supported vectors of $\mathcal{H} = \ell^2(T_{d-1})$. The symmetric densely-defined operator $L(T)$ has a unique extension to $\mathcal{H}$. This extension need not be self-adjoint, and when it does we say that the tree T is self-adjoint. In such cases the spectral theorem can be applied on $L(T)$. It can be shown that a Poisson d-tree is, almost surely, self-adjoint.

We employ the useful property that spectral measures are continuous with respect to local weak convergence. Since $Y = Y_d\left(n, \frac{c}{n}\right)$ converges in local weak convergence to the Poisson d-tree $T = T_d(c)$, which is almost-surely self-adjoint, we conclude that the expected measure $\mathbb{E}_Y[\pi_{L(Y),e_\tau}]$ weakly converges to the expected measure $\mathbb{E}_T[\pi_{L(T),e_o}]$, where o is the root of T. In particular, by measuring the closed set $\{0\}$,

$$\limsup_{n\to\infty} \mathbb{E}\left[\|P_Z(e_\tau)\|^2\right] \leq \mathbb{E}[x_T],$$

where $x_T := \pi_{L(T),e_o}(\{0\})$. Consequently,

$$\mathbb{E}[\dim Z] \leq (1 + o_n(1))\binom{n}{d}\mathbb{E}[x_T].$$

By the Rank-Nullity Theorem from linear algebra,

$$\dim Z - \beta_d(Y) = f_{d-1}(Y) - f_d(Y),$$

and we conclude that

$$\frac{1}{\binom{n}{d}}\mathbb{E}[\beta_d(Y)] \leq \mathbb{E}[x_T] - 1 + \frac{c}{d+1} + o_n(1).$$

There remains the problem of bounding the expectation $\mathbb{E}_T\left[x_T\right]$ without directly computing the operator's kernel. This difficulty is bypassed using the recursive structure of d-trees to derive a simple recursion formulas on these spectral measures, as in the following lemma.

Let T be a self-adjoint d-tree with root o, and let $\sigma_1, \ldots, \sigma_m$ be the d-faces that contain the root. For $1 \leq j \leq m$ and $1 \leq r \leq d$ we denote by $\tau_{j,r}$ the $(d-1)$-faces of σ_j other than the root. Also, $T_{j,r}$ denotes the branch rooted at $\tau_{j,r}$.

Lemma 5.1 *$x_T = 0$ if there exists some $1 \leq j \leq m$ such that $x_{T_{j,1}} = \ldots = x_{T_{j,d}} = 0$. Otherwise,*

$$x_T = \left(1 + \sum_{j=1}^{m}\left(\sum_{r=1}^{d} x_{T_{j,r}}\right)^{-1}\right)^{-1}.$$

Proof sketch Consider the following bounded family of measurable functions $\{H_s : \mathbb{R} \to \mathbb{C} \mid s \in \mathbb{R} \setminus \{0\}\}$,

$$H_s(x) = \frac{is}{x + is}.$$

Note that H_s approaches the Kronecker delta function $\delta_{x,0}$ as $s \to 0$. Given a self-adjoint d-tree T with root o, we define

$$h_T(s) := \int_{\mathbb{R}} H_s(x) d\pi_{L(T),e_o}(x).$$

In particular, $x_T = \lim_{s\to 0} h_T(s)$. The proof is concluded by showing that the functions $h_{T_{j,r}}$'s and h_T satisfy the formula

$$h_T(s)\left(1 + \sum_{j=1}^{m}\left(is + \sum_{r=1}^{d} h_{T_{j,r}}(s)\right)^{-1}\right) = 1, \tag{7}$$

and letting $s \to 0$.

We turn to describe the derivation of Eq. (7). Denote $L := L(T)$, $\tilde{L} := \bigoplus_{j,r} L(T_{j,r})$ and M the Laplacian of the subcomplex of T which contains $\sigma_1, \ldots, \sigma_m$ and their subfaces. In particular, $L = M \oplus \tilde{L}$. The operators $H_s(L)$ and $H_s(\tilde{L})$ are scalar multiples of the *resolvents* $R := (L + is \cdot I)^{-1}$ and $\tilde{R} := (\tilde{L} + is \cdot I)^{-1}$. In particular, $h_T(s) = is \cdot \langle Re_0, e_0 \rangle$. Simple observations about the d-tree structure yields that (i) $\tilde{R}e_o = \frac{1}{is} e_o$ and (ii) $\left\langle \tilde{R}e_{\tau_{j,r}}, e_{\tau_{j',r'}} \right\rangle = 0$ when $(j,r) \neq (j',r')$. In addition, we use the Second Resolvent Identity which states that $RM\tilde{R} = \tilde{R} - R$. Here the operator M has a very concrete and usable form, being the Laplacian of a d-complex which consists of m distinct d-faces with a common $(d-1)$-subface o. Equation (7) is obtained using simple algebraic manipulations by comparing terms of the form $\langle (RM\tilde{R})e_\tau, e_{\tau'} \rangle = \langle (\tilde{R} - R)e_\tau, e_{\tau'} \rangle$, when τ, τ' are $(d-1)$-faces of T of distance at most 1 from the root o. □

The remaining step of the argument is an application of the recursive formula in Lemma 5.1 to the Poisson d-tree.

Lemma 5.2 *Let $T = T_d(c)$ be a rooted Poisson d-tree with parameter c. Then,*

$$\mathbb{E}[x_T] \leq \max \left\{ t + ct(1-t)^d - \frac{c}{d+1} \left(1 - (1-t)^{d+1} \right) \mid t \in [0,1],\ t = e^{-c(1-t)^d} \right\}.$$

Note that this maximum is taken over a finite set, due to the condition $t = e^{-c(1-t)^d}$.

Proof Let T be a Poisson d-tree with root degree m and $\{T_{j,r} \mid 1 \leq j \leq m, 1 \leq r \leq d\}$ its subtrees as above. The parameters x_T, $\{x_{T_{j,r}}\}$ can be considered as random variables when T is $T_d(c)$-distributed. The random variables $\{x_{T_{j,r}}\}$ are i.i.d. and are distributed like x_T since all the subtrees $T_{j,r}$ are independent Poisson d-trees. In addition, these variables satisfy the equation of Lemma 5.1.

These observations suggest the following equivalent description of $\mathcal{D}$, the distribution of the random variable x_T. First sample a $Poi(c)$-distributed integer m, and $x_{T_{j,r}} = \mathcal{D}$ i.i.d for every $1 \leq j \leq m$ and $1 \leq r \leq d$. Given these samples, the value of x_T is determined by Lemma 5.1.

In particular, if we let $t := \Pr(x_T > 0)$, then t satisfies the equation

$$t = \sum_{m=0}^{\infty} \frac{e^{-c}c^m}{m!} \left(1 - (1-t)^d \right)^m = e^{-c(1-t)^d}. \tag{8}$$

Let X be a $\mathcal{D}$-distributed random variable,

$$\mathbb{E}[X] = \mathbb{E} \left[\frac{\mathbf{1}_{\{\forall j \in [m],\ S_j > 0\}}}{1 + \sum_{j=1}^{m} S_j^{-1}} \right]$$

Here $S_1, S_2, \ldots, S_m$ are random variables whose distribution is that of a sum of d i.i.d. $\mathcal{D}$-distributed variables. By expressing the probability $\Pr[\mathbf{1}_{\{\forall j \in [m],\ S_j > 0\}}]$ as t, exploiting the symmetry between the different S_j's and using basic properties of the

Poisson distribution, we are able to express this expectation in terms of t,

$$\mathbb{E}[X] = t + ct(1-t)^d - \frac{c}{d+1}\left(1-(1-t)^{d+1}\right),$$

as was claimed. □

It requires only basic calculus to conclude the following.

1. For $c < c_d$, the maximum of

$$t + ct(1-t)^d - \frac{c}{d+1}\left(1-(1-t)^{d+1}\right), \quad \text{s.t.} \quad t = e^{-c(1-t)^d}$$

   is attained at $t = 1$. Consequently, $\mathbb{E}[\beta_d(Y)] \leq o(n^d)$.
2. For $c > c_d$, the maximum is attained at $t = t(c, d)$. Consequently,

$$\mathbb{E}[\beta_d(Y)] \leq (1 + o_n(1))\binom{n}{d}\left(\frac{c(1-t)^{d+1}}{d+1} - 1 + t + ct(1-t)^d\right). \tag{9}$$

The proof of Theorem 1.1 (II.b) is concluded by the following standard probabilistic argument. Let $c < c_d$ and $\varepsilon > 0$. The absence of small non-collapsible subcomplexes in Y (Lemma 4.1) implies that it has no small d-cycles except $\partial\Delta_{d+1}$'s. The computation above shows that the dimension of the cycle space is $o(n^d)$. Therefore, removing Bin $\left(\binom{n}{d+1}, \varepsilon/n\right) = \Theta(n^d)$ random d-faces eliminates all large (non $\partial\Delta_{d+1}$) d-cycles with high probability, hence $Y_d\left(n, \frac{c-\varepsilon}{n}\right)$ is a.a.s. d-acyclic except $\partial\Delta_{d+1}$'s.

5.2 The Cyclic Regime: Theorem 1.1(III)

The key idea of [5] for the computation of a matching upper bound for the d-acyclicity threshold uses an analysis of the collapse process. Let Y be a d-complex and k some positive integer. Recall that $R_k(Y)$ is the simplicial complex obtained from Y by k phases of d-collapse and $S_k(Y)$ is its subcomplex obtained by removing the exposed $(d-1)$-faces. Clearly, $\beta_d(Y) = \beta_d(S_k(Y))$ since d-collapsing and removing exposed faces does not affect the right kernel of ∂_d. The final ingredient of the strategy is the observation that $\beta_d(S_k(Y)) \geq f_d(S_k(Y)) - f_{d-1}(S_k(Y))$, and the fact that the parameter $f_d(S_k(Y)) - f_{d-1}(S_k(Y))$ can be studied from the local weak limit.

Let $Y = Y_d\left(n, \frac{c}{n}\right)$ where $c > c_d$, and $Y' := S_k(Y)$ for a sufficiently large integer k.

$$\begin{aligned}\mathbb{E}[\beta_d(Y)] &\geq \mathbb{E}[f_d(Y') - f_{d-1}(Y')]\\ &= \mathbb{E}\left[\sum_{\tau\in Y'_{d-1}}\left(\frac{d_{Y'}(\tau)}{d+1} - 1\right)\right]\\ &= \binom{n}{d}\Pr[\tau \in Y']\left(\frac{\mathbb{E}\left[d_{Y'}(\tau)|\tau \in Y'\right]}{d+1} - 1\right). \qquad (10)\end{aligned}$$

For the last equation we can, due to symmetry, consider a fixed τ and apply the Law of Total Expectation to the event $\{\tau \in Y'\}$. As mentioned above, the $(d-1)$-face τ of Y belongs to Y' if and only if in the τ-rooted collapse process of Y, the degree of τ is greater than 1 after $(k-1)$ phases and positive after k steps. By approximating the τ-rooted collapse process of Y with the rooted collapse process on $T_d(c)$ we obtain that up to $1 + o_n(1)$ factor,

$$\Pr[\tau \in Y'] \geq \Pr[\delta_k > 1] = 1 - t_k - c(1 - t_{k-2})^d t_{k-1}. \qquad (11)$$

Furthermore, upto $1 + o_n(1)$ factor,

$$\begin{aligned}\mathbb{E}\left[d_{Y'}(\tau)|\tau \in Y'\right] &= \mathbb{E}[\delta_k \mid \delta_k > 0 \ \wedge \ \delta_{k-1} > 1]\\ &\geq \sum_{j=2}^{\infty} j \cdot \Pr[\delta_k = j \mid \delta_k > 0 \ \wedge \ \delta_{k-1} > 1]\\ &\geq \sum_{j=2}^{\infty} j \cdot \frac{\Pr[\delta_k = j]}{\Pr[\delta_{k-1} > 1]}\\ &= \frac{c(1-t_{k-1})^d(1-t_k)}{1 - t_{k-1} - c(1-t_{k-2})^d t_{k-1}}. \qquad (12)\end{aligned}$$

By combining Inequalities (10), (11) and (12), and letting $k \to \infty$, we obtain that

$$\mathbb{E}[\beta_d(Y)] \geq (1 + o_n(1))\binom{n}{d}\left(\frac{c(1-t)^{d+1}}{d+1} - 1 + t + ct(1-t)^d\right).$$

This bound starts to be meaningful for $c > c_d$, where this expression matches the upper bound from (9). Since β_d is 1-Lipschitz, a straightforward application of Azuma's inequality yields that a.a.s. β_d deviates from its expectation by only $o(n^d)$. In particular, this shows a matching upper bound for the threshold of d-acyclicity.

5.3 $\mathbb{R}$-*Shadow of* $Y_d\left(n, \frac{c}{n}\right)$

The behavior of the $\mathbb{R}$-shadow when $c < c_d$ is studied similarly to the C-shadow in the collapsible regime (See Sect. 4.3).

We turn to the range $c > c_d$. Here we do not give a proof, but only a general intuitive explanation. An accurate analysis of the measure concentration can be found in [13]. Recall that

$$\frac{1}{\binom{n}{d}}\mathbb{E}[\beta_d(Y;\mathbb{R})] \xrightarrow[n\to\infty]{} g_d(c) := \frac{c}{d+1}(1-t)^{d+1} - (1-t) + ct(1-t)^d.$$

A simple technical claim shows that for every $c > c_d$, the limit function $g_d(c)$ is differentiable with respect to the variable c and its derivative equals to $\frac{1}{d+1}(1-t)^{d+1}$.

It turns out to be more convenient to work here with a d-dimensional analog of the so-called *evolution of random graphs*. Let $Y_d(n,m)$ be a random simplicial complex with n vertices, a complete $(d-1)$-skeleton and m uniformly random d-faces. $Y' = Y_d(n, m+1)$ can be sampled by the following procedure. First sample $Y = Y_d(n,m)$ and then add a random d-face which does not belong to Y. Therefore, the following equation holds in expectation,

$$\beta_d(Y';\mathbb{R}) - \beta_d(Y;\mathbb{R}) = \frac{1}{\binom{n}{d+1}}\left|\mathrm{SH}_{\mathbb{R}}(Y)\right|.$$

Letting $m = \mathrm{Bin}\left(\binom{n}{d+1}, \frac{c}{n}\right)$ yields that $Y = Y_d\left(n, \frac{c}{n}\right)$, and its real d-Betti number is $\binom{n}{d} \cdot g_d(c) + o(n^d)$. The previous equation suggests, at least intuitively, the following relation between the growth of g_d and the $\mathbb{R}$-shadow. Namely,

$$\binom{n}{d}\left(g_d\left(c + \frac{d+1}{\binom{n}{d}}\right) - g_d(c)\right) \approx \frac{1}{\binom{n}{d+1}}\left|\mathrm{SH}_{\mathbb{R}}\left(Y\right)\right|,$$

and by letting $n \to \infty$,

$$\frac{1}{\binom{n}{d+1}}\left|\mathrm{SH}_{\mathbb{R}}\left(Y\right)\right| \approx (d+1)g_d'(c) = (1-t)^{d+1}.$$

This argument can be made rigorous by incrementing Y with εn^d random d-faces at a time rather than one by one, and applying standard measure concentration inequalities.

6 Concluding Remarks

Although this article is mostly a review of previous work, it does contain several new results, e.g. Theorem 1.1(I) that deals with the collapsible regime is a little stronger than the original result of [6]. Other notable new results concern the asymptotic densities of the core and the C-shadow of $Y_d\left(n, \frac{c}{n}\right)$ for $c > \gamma_d$, improving the main theorem of [4] which says that $Y_d\left(n, \frac{c}{n}\right)$ is a.a.s. non-collapsible for $c > \gamma_d$. As mentioned above, although the main result in that paper is correct, there is an error in the proof, which we are able to remedy here using the techniques of Riordan [20].

The results surveyed here can be viewed from several perspectives which suggest different problems for future research.

From the combinatorial perspective, the phase transition in the density of the shadow of $Y_d\left(n, \frac{c}{n}\right)$ is of great interest. We conjecture that the $\mathbb{R}$-shadow grows from linear in n to a giant (order $\Theta(n^{d+1})$) in a single step in a random evolution of simplicial complexes. This starkly contrasts with the gradual growth of the giant component in random graphs. It is of particular interest to understand the structure of the critical complex. Numerical experiments suggest that its $(d-1)$-homology group has torsion of size $\exp(\Theta(n^d))$, but we do not know a proof of this yet.

On the topological side, it would be very interesting to better understand the giant d-cycles which appear in $Y_d\left(n, \frac{c}{n}\right)$ when $c > c_d$. Provably, they consist of $\Theta(n^d)$ d-faces, but numerical experiments suggest that they lie in an unknown territory in the realm of homological d-cycles. Unlike closed manifolds, in which the degree of all $(d-1)$-faces equals to 2, it seems that in these d-cycles the average degree approaches $d+1$, which is the largest possible for a minimal d-cycle. Namely, these d-cycles are in some sense the opposite of manifolds.

In addition, the random complexes $Y_d\left(n, \frac{c}{n}\right)$, where $\gamma_d < c < c_d$ have the nice property of being d-acyclic but not d-collapsible. What other interesting topological or combinatorial properties do they have?

There is much more to study about random simplicial complexes in the regime $p = c/n$. In particular, the question regarding the vanishing of the top homology over finite fields is still open and presently out of reach. It is interesting to resolve whether homology thresholds over other fields can also be read off some parameter of the Poisson d-tree.

References

1. D. Aldous, The $\zeta(2)$ limit in the random assignment problem. Random Struct. Algorithms **18**(4), 381–418 (2001)
2. D. Aldous, R. Lyons, Processes on unimodular random networks. Electron. J. Probab **12**(54), 1454–1508 (2007)
3. D. Aldous, M. Steele, The objective method: probabilistic combinatorial optimization and local weak convergence, in *Probability on Discrete Structures*, ed. by H. Kesten (Springer, Berlin, 2004), pp. 1–72

4. L. Aronshtam, N. Linial, The threshold for collapsibility in random complexes (2013). arXiv preprint arXiv:1307.2684
5. L. Aronshtam, N. Linial, When does the top homology of a random simplicial complex vanish? Random Struct. Algorithms **46**, 26–35 (2013)
6. L. Aronshtam, N. Linial, T. Łuczak, R. Meshulam, Collapsibility and vanishing of top homology in random simplicial complexes. Discret. Comput. Geom. **49**(2), 317–334 (2013)
7. E. Babson, C. Hoffman, M. Kahle, The fundamental group of random 2-complexes. J. Am. Math. Soc. **24**(1), 1–28 (2011)
8. I. Benjamini, O. Schramm, Recurrence of distributional limits of finite planar graphs, in *Selected Works of Oded Schramm*, ed. by I. Benjamini, O. Häggström (Springer, New York, 2011), pp. 533–545
9. C. Bordenave, M. Lelarge, J. Salez, The rank of diluted random graphs. Ann. Probab. **39**(3), 1097–1121 (2011)
10. C. Hoffman, M. Kahle, E. Paquette, The threshold for integer homology in random d-complexes (2013). arXiv preprint arXiv:1308.6232
11. D. Korándi, Y. Peled, B. Sudakov, A random triadic process. SIAM J. Discret. Math. **30**(1), 1–19 (2016)
12. N. Linial, R. Meshulam, Homological connectivity of random 2-complexes. Combinatorica **26**(4), 475–487 (2006)
13. N. Linial, Y. Peled, On the phase transition in random simplicial complexes. Ann. Math. **184**, 745–773 (2016)
14. N. Linial, I. Newman, Y. Peled, Y. Rabinovich, Extremal problems on shadows and hypercuts in simplicial complexes (2014). arXiv preprint arXiv:1408.0602
15. T. Łuczak, Y. Peled, Integral homology of random simplicial complexes (2016). arXiv preprint, arXiv:1607.06985
16. R. Lyons, Asymptotic enumeration of spanning trees. Combin. Probab. Comput. **14**(04), 491–522 (2005)
17. R. Meshulam, N. Wallach, Homological connectivity of random k-dimensional complexes. Random Struct. Algorithms **34**(3), 408–417 (2009)
18. M. Molloy, Cores in random hypergraphs and boolean formulas. Random Struct. Algorithms **27**(1), 124–135 (2005)
19. M. Reed, B. Simon, *Methods of Modern Mathematical Physics. I: Functional Analysis*, 2nd edn. (Academic Press [Harcourt Brace Jovanovich, Publishers], New York, 1980). MR 0751959. Zbl 0459.46001
20. O. Riordan, The k-core and branching processes. Combin. Probab. Comput. **17**(1), 111–136 (2008)

Nullspace Embeddings for Outerplanar Graphs

László Lovász and Alexander Schrijver

Dedicated to the memory of Jiří Matoušek

Abstract We study relations between geometric embeddings of graphs and the spectrum of associated matrices, focusing on outerplanar embeddings of graphs. For a simple connected graph $G = (V, E)$, we define a "good" G-matrix as a $V \times V$ matrix with negative entries corresponding to adjacent nodes, zero entries corresponding to distinct nonadjacent nodes, and exactly one negative eigenvalue. We give an algorithmic proof of the fact that if G is a 2-connected graph, then either the nullspace representation defined by any "good" G-matrix with corank 2 is an outerplanar embedding of G, or else there exists a "good" G-matrix with corank 3.

1 Introduction

We study relations between geometric embeddings of graphs, the spectrum of associated matrices and their signature, and topological properties of associated cell complexes. We focus in particular on 1-dimensional and 2-dimensional embeddings of graphs, in the hope that the techniques can be extended to higher dimensions.

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007–2013)/ERC grant agreement no. 227701.

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007–2013)/ERC grant agreement no. 339109.

L. Lovász
Eötvös Loránd University, Budapest, Hungary

A. Schrijver (✉)
University of Amsterdam and CWI, Science Park 107, 1098 XG Amsterdam, The Netherlands
e-mail: lex@cwi.nl

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_23

Spectral parameters of graphs The basic connection between graphs, matrices, and geometric embeddings considered in this paper can be described as follows. We define a *G-matrix* for an undirected graph $G = (V, E)$ as a symmetric real-valued $V \times V$ matrix M with $M_{ij} = 0$ if i and j are distinct nonadjacent nodes. The matrix is *well-signed* if $M_{ij} < 0$ for adjacent nodes i and j. (There is no condition on the diagonal entries.) If, in addition, M has exactly one negative eigenvalue, then let us call it *good* (for the purposes of this introduction). Let $\kappa(G)$ denote the largest d for which there exists a good G-matrix with corank d. (The corank is the dimension of the nullspace.)

The parameter κ is closely tied to certain topological properties of the graph. Combining results of [1, 5, 7, 9] and [8], one gets the following facts:

If G is connected, then $\kappa(G) \leq 1 \Leftrightarrow G$ is a path,
If G is 2-connected, then $\kappa(G) \leq 2 \Leftrightarrow G$ is outerplanar,
If G is 3-connected, then $\kappa(G) \leq 3 \Leftrightarrow G$ is planar,
If G is 4-connected, then $\kappa(G) \leq 4 \Leftrightarrow G$ is linklessly embeddable in $\mathbb{R}^3$.

We study algorithmic aspects of the first two facts. Let us discuss here the second, which says that if G is a 2-connected graph, then either it has an embedding in the plane as an outerplanar map, or else there exists a good G-matrix with corank 3 (and so the graph is not outerplanar). To construct an outerplanar embedding, we use the nullspace of any good G-matrix with corank 2.

Nullspace representations To describe this construction, suppose that a G-matrix M has corank d. Let $U \in \mathbb{R}^{d\times n}$ be a matrix whose rows form a basis of the nullspace of M. This matrix satisfies the equation $UM = 0$. Let u_i be the column of U corresponding to node $i \in V$. The mapping $u : V \to \mathbb{R}^d$ is called the *nullspace representation of V defined by M*. It is unique up to linear transformations of $\mathbb{R}^d$. (For the purist: the map $V \to \ker(M)^*$ is canonically defined; choosing the basis in $\ker(M)$ just identifies $\ker(M)^*$ with $\mathbb{R}^d$.)

If $G = (V, E)$ is a graph and $u : V \to \mathbb{R}^d$ is any map, we can extend it to the edges by mapping the edge ij to the straight line segment between u_i and u_j. If u is the nullspace representation of V defined by M, then this extension gives the *nullspace representation of G defined by M*.

In this paper we give algorithmic proofs of two facts:

1. If G is a connected graph with $\kappa(G) = 1$, then the nullspace representation defined by any good G-matrix with corank 1 yields an embedding of G in the line.
2. If G is 2-connected and $\kappa(G) = 2$, then the nullspace representation defined by any good G-matrix with corank 2 yields an outerplanar embedding of G.

(The word "yields" above hides some issues concerning normalization, to be discussed later.) The proofs are algorithmic in the sense that (say, in the case of (2)) for every 2-connected graph we either construct an outerplanar embedding or a good G-matrix with corank 3 in polynomial time. The alternative proof that can be derived from the results of [6] uses the minor-monotonicity of the Colin de Verdière parameter (see below), and this way it involves repeated reference to the Implicit Function Theorem, and does not seem to be implementable in polynomial time. Our algorithms use exact real arithmetic and a subroutine for finding roots of one-variable polynomials, which are steps that can be easily turned into polynomial-time algorithms (say, in binary arithmetic).

Suppose that the input to our algorithm is a 3-connected planar graph. Then the algorithm outputs a good G-matrix with corank at least 3. Paper [6] also contains the analogous result for planar graphs, which was extended in [4]:

3. If G is 3-connected and $\kappa(G) = 3$, then the nullspace representation defined by any good G-matrix with corank 3 yields a representation of G as the skeleton of a convex 3-polytope.

Thus computing the nullspace representation defined by the matrix M, and performing node-scaling as described in [4], we get a representation of G as the skeleton of a 3-polytope.

Unfortunately, the proof of (3) uses the minor-monotonicity of the Colin de Verdière parameter and the Implicit Function Theorem, and hence it does not yield an efficient algorithm: if the input is not a planar graph, then it does *not* provide a polynomial-time algorithm to compute a good G-matrix with corank at least 4. It would be interesting to see whether our approach can be extended to the case $\kappa \geq 3$. (While we focus on the case $\kappa, \leq 2$, some of our results do bear upon higher dimensions, in particular the results in Sect. 2.2 below.)

A further extension to dimension 4 would be particularly interesting, since for 4-connected graphs G, linkless embeddability is characterized by the property that $\kappa(G) \leq 4$, but it is not known whether the nullspace representation obtained from a good G-matrix of corank 4 yields a linkless embedding of the graph.

The Strong Arnold Hypothesis and the Colin de Verdière number We conclude this introduction with a discussion of the connection between the parameter $\kappa(G)$ and the graph parameter $\mu(G)$ introduced by Colin de Verdière (cf. [11]). This latter is defined similarly to κ as the maximum corank of a good G-matrix M, where it is required, in addition, that M has a nondegeneracy property called the *Strong Arnold Property*. There are several equivalent forms of this property; let us formulate one that is related to our considerations in the sense that it uses any nullspace representation u defined by M: if a symmetric $d \times d$ matrix N satisfies $u_i^{\mathsf{T}} N u_i = 0$ for all $i \in V$ and $u_i^{\mathsf{T}} N u_j = 0$ for each edge ij of G, then $N = 0$. In more geometric terms this means that the nullspace representation of the graph defined by M is not contained in any nontrivial homogeneous quadric.

The relationship between μ and κ is not completely clarified. Trivially $\mu(G) \leq \kappa(G)$. Equality does not hold in general: consider the graph $G_{l,m}$ made from an $(l + m)$-clique by removing the edges of an m-clique. If $l \geq 1$ and $m \geq 3$, then $\mu(G_{l,m}) = l+1$ whereas $\kappa(G_{l,m}) = l+m-2$. (Note that $G_{l,m}$ is not $l+1$-connected.)

Colin de Verdière's parameter has several advantages over κ. First, it is minor-monotone, while $\kappa(G)$ is not minor-monotone, not even subgraph-monotone: any path P satisfies $\kappa(P) \leq 1$, but a disjoint union of paths can have arbitrarily large

$\kappa(G)$. Furthermore, the connection with topological properties of graphs holds for μ without connectivity conditions:

$\mu(G) \leq 1 \Leftrightarrow G$ is a disjoint union of paths,
$\mu(G) \leq 2 \Leftrightarrow G$ is outerplanar,
$\mu(G) \leq 3 \Leftrightarrow G$ is planar,
$\mu(G) \leq 4 \Leftrightarrow G$ is linklessly embeddable in $\mathbb{R}^3$.

Our use of κ is motivated by its easier definition and by the (slightly) stronger, algorithmic results.

We see from the facts above that by requiring that G is $\mu(G)$-connected, we have $\mu(G) = \kappa(G)$ for $\mu(G) \leq 4$. In fact, it was shown by Van der Holst [10] that if G is 2-connected outerplanar or 3-connected planar, then *every* good G-matrix has the Strong Arnold Property. This also holds true for 4-connected linklessly embeddable graphs [8]. One may wonder whether this remains true for $\mu(G)$-connected graphs with larger $\mu(G)$. This would imply that $\mu(G) = \kappa(G)$ for every $\mu(G)$-connected graph.

Remark 1 Our setup is related to rigidity theory of bar-and-joint structures. To formulate just one connection, let G be a graph, M a well-signed G-matrix, and $u : V(G) \to \mathbb{R}^d$ a nullspace representation, considered as specifying a position for each node. Replace the edges by rubber bands of strength M_{ij} (i.e., stretching an edge to length t results in a force of $-M_{ij}t$ pulling the endpoints together). Add "braces" (rigid bars) from the origin to each node; these braces can carry an arbitrary force, as long as it is parallel to the brace. Then the equation $UM = 0$ just says that the structure is in equilibrium (where, as before, U is the matrix with columns u_i). The matrix M is called a (braced) *stress matrix* on the structure (G, u).

Other conditions like the rank of the matrix M, its signature and its Strong Arnold Property also play a significant role in rigidity theory; see [2, 3].

2 *G*-Matrices

2.1 Nullspace Representations

Let us fix a connected graph $G = (V, E)$ on node set $V = [n]$, and an integer $d \geq 1$. We denote by $\mathcal{W}$ the set of well-signed G-matrices with corank at least d, and by $\mathcal{W}^=$, the set of well-signed G-matrices with corank exactly d. We denote by $\mathcal{W}^1$ the set of G-matrices in $\mathcal{W}$ with exactly one negative eigenvalue (counted with multiplicity).

Suppose that we are also given a vector labeling $u : V \to \mathbb{R}^d$, which we can encode as a $d \times V$ matrix U, whose column corresponding to $i \in V$ is the vector u_i. For $p \in \mathbb{R}^d$, let us write $u - p$ for the representation $(u_1 - p, \ldots, u_n - p)$. We denote by $\mathcal{M}_u$ the linear space of G-matrices M with $UM = 0$, by $\mathcal{W}_u$, the set of well-signed G-matrices in $\mathcal{M}_u$, by $\mathcal{W}_u^1$, the set of matrices in $\mathcal{W}_u$ with exactly one negative eigenvalue, and by $\mathcal{W}_u^2$, the set of matrices in $\mathcal{W}_u$ with at least two negative eigenvalues.

We can always perform a linear transformation of $\mathbb{R}^d$, i.e., replace U by AU, where A is any nonsingular $d \times d$ matrix. In the case when $\operatorname{corank}(M) = d$ (which will be the important case for us), the matrix U is determined by M up to such a linear transformation of $\mathbb{R}^d$.

Another simple transformation we use is "node scaling": replacing U by $U' = UD$ and M by $M' = D^{-1}MD^{-1}$, where D is a nonsingular diagonal matrix with positive diagonal. Then M' is a G-matrix and $U'M' = 0$. Moreover, it maintains well-signedness of M. Through this transformation, we may assume that every nonzero vector u_i has unit length. We call such a representation *normalized*.

One of our main tools will be to describe more explicit solutions of the basic equation $UM = 0$ in dimensions 1 and 2. More precisely, given a graph $G = (V, E)$ and a vector labeling $u : V \to \mathbb{R}^2$, our goal is to describe all G-matrices M with $UM = 0$. Note that if the vector labels are nonzero, then it suffices to find the off-diagonal entries: if M_{ij} is given for $ij \in E$ in such a way that $\sum_{j \in N(i)} M_{ij} u_j$ is a scalar multiple of u_i for every node i, then there is a unique choice of diagonal entries M_{ii} that gives a matrix with $UM = 0$:

$$M_{ii} = -\sum_j M_{ij} \frac{u_j^\mathsf{T} u_i}{u_i^\mathsf{T} u_i}. \tag{1}$$

2.2 G-Matrices and Eigenvalues

In this section we consider eigenvalues of well-signed G-matrices; we consider the connected graph G and the dimension parameter d fixed. We start with a couple of simple observations.

Lemma 2 *Let M be a well-signed G-matrix and let $U \in \mathbb{R}^{d \times n}$ such that $UM = 0$ and* $\operatorname{rank}(U) = d$.

(a) *If M is positive semidefinite, then* $\operatorname{corank}(M) = d = 1$, *and all entries of U are nonzero and have the same sign.*
(b) *If M has a negative eigenvalue, then the origin is an interior point of the convex hull of the columns of U.*

Proof Let λ be the smallest eigenvalue of M. As G is connected, λ has multiplicity one by the Perron–Frobenius theorem, and M has a positive eigenvector v belonging to λ. If $\lambda = 0$, then this multiplicity is $d = 1$, and U consists of a single row parallel to v. If $\lambda < 0$, then every row of U, being in the nullspace of M, is orthogonal to v. Thus the entries of v provide a representation of 0 as a convex combination of the columns of U with positive coefficients. □

Lemma 3 *If $d \geq 2$, then the set $\mathcal{W}^1$ is relatively closed in $\mathcal{W}$, and $\mathcal{W}^1 \cap \mathcal{W}^=$ is relatively open in $\mathcal{W}$.*

Proof Let $\lambda_i(M)$ denote the i-th smallest eigenvalue of the matrix M. We claim that for any $M \in \mathcal{W}$,

$$M \in \mathcal{W}^1 \;\Leftrightarrow\; \lambda_2(M) \geq 0. \tag{2}$$

Indeed, if $M \in \mathcal{W}^1$, then trivially $\lambda_2(M) \geq 0$. Conversely, if $\lambda_2(M) \geq 0$, then M has at most one negative eigenvalue. By Lemma 2(a), it has exactly one, that is, $M \in \mathcal{W}^1$. This proves (2). Since $\lambda_2(M)$ is a continuous function of M, the first assertion of the lemma follows.

We claim that if $d \geq 2$, for any $M \in \mathcal{W}$,

$$M \in \mathcal{W}^1 \cap \mathcal{W}^= \;\Leftrightarrow\; \lambda_{d+2}(M) > 0. \tag{3}$$

Indeed, if $M \in \mathcal{W}^1 \cap \mathcal{W}^=$, then M has one negative eigenvalue and exactly d zero eigenvalues, and so $\lambda_{d+2}(M) > 0$. Conversely, assume that $\lambda_{d+2}(M) > 0$. Since M has at least d zero eigenvalues and at least one negative eigenvalue (by Lemma 2(a)), we must have equality in both bounds, which means that $M \in \mathcal{W}^1 \cap \mathcal{W}^=$. This proves (3). Continuity of $\lambda_{d+2}(M)$ implies the second assertion. □

This last lemma implies that each nonempty connected subset of $\mathcal{W}^=$ is either contained in $\mathcal{W}^1$ or is disjoint from $\mathcal{W}^1$. We formulate several consequences of this fact.

Lemma 4 *Suppose that G is 2-connected, and let M be a well-signed G-matrix with one negative eigenvalue and with corank $d = \kappa(G)$. Let u be the nullspace representation defined by M, let $v \in \mathbb{R}^d$, and let $J := \{i : u_i = v\}$. If $|J| \geq 2$, then the origin 0 belongs to the convex hull of $u(V \setminus J)$.*

Proof For $i \in V$, let e_i be the i-th unit basis vector, and for $i, j \in V$, let D^{ij} be the matrix $(e_i - e_j)(e_i - e_j)^\mathsf{T}$. Define

$$M^\alpha := M + \alpha \sum_{\substack{ij \in E \\ i,j \in J}} M_{ij} D^{ij} \qquad (\alpha \in [0,1]).$$

The definition of J implies that $\ker(M) \subseteq \ker(D^{ij})$ for all $i, j \in J$, and hence $\ker(M) \subseteq \ker(M^\alpha)$ for each $\alpha \in [0,1]$. So $\operatorname{corank}(M^\alpha) \geq \operatorname{corank}(M) = \kappa(G)$ for each $\alpha \in [0,1]$. Moreover, M^α is a well-signed G-matrix for each $\alpha \in [0,1)$. So $M^\alpha \in \mathcal{W}$ for each $\alpha \in [0,1)$. As $\kappa(G) = d$, we know $\mathcal{W}^1 \subseteq \mathcal{W}^=$, hence $\mathcal{W}^1 \cap \mathcal{W}^= = \mathcal{W}^1$. So by Lemma 3, $\mathcal{W}^1$ is relatively open and closed in $\mathcal{W}$. Since $M = M^0 \in \mathcal{W}^1$, this implies that $M^\alpha \in \mathcal{W}^1$ for each $\alpha \in [0,1)$. By the continuity of eigenvalues, M^1 has at most one negative eigenvalue. Note that $M^1_{ij} = 0$ for any two distinct $i, j \in J$.

Assume that 0 does not belong to the convex hull of $\{u_i : i \notin J\}$. Then there exists $c \in \mathbb{R}^{\kappa(G)}$ such that $u_i^\mathsf{T} c < 0$ for each $i \notin J$. As 0 belongs to interior of the convex hull of $u(V)$ by Lemma 2(b), this implies that $u_i^\mathsf{T} c = v^\mathsf{T} c > 0$ for each $i \in J$.

As $|J| \geq 2$, the 2-connectivity of G implies that J contains two distinct nodes, say nodes 1 and 2, that have neighbors outside J. Since $\ker(M) \subseteq \ker(M^1)$, we have $\sum_j M^1_{1j} u_j = 0$, and hence

$$M^1_{11} u_1^\mathsf{T} c = -\sum_{j \neq 1} M^1_{1j} u_j^\mathsf{T} c = -\sum_{j \notin J} M^1_{1j} u_j^\mathsf{T} c.$$

As $u_1^\mathsf{T} c > 0$ and $u_j^\mathsf{T} c < 0$ for $j \notin J$, and as $M^1_{1j} \leq 0$ for all $j \notin J$, and $M^1_{1j} < 0$ for at least one $j \notin J$, this implies $M^1_{11} < 0$. Similarly, $M^1_{22} < 0$. As $M^1_{12} = 0$, the first two rows and columns of M^1 induce a negative definite 2×2 submatrix of M^1. This contradicts the fact that M^1 has at most one negative eigenvalue. □

For the next step we need a simple lemma from linear algebra.

Lemma 5 *Let A and M be symmetric $n \times n$ matrices. Assume that A is 0 outside a $k \times k$ principal submatrix, and let M_0 be the complementary $(n-k) \times (n-k)$ principal submatrix of M. Let a and b denote the number of negative eigenvalues of A and M_0, respectively. Then for some $s > 0$, the matrix $sM + A$ has at least $a + b$ negative eigenvalues.*

Proof We may assume $A = \begin{pmatrix} A_0 & 0 \\ 0 & 0 \end{pmatrix}$ and $M = \begin{pmatrix} M_1 & M_2^\mathsf{T} \\ M_2 & M_0 \end{pmatrix}$, with A_0 and M_1 having order $k \times k$. By scaling the last $n-k$ rows and columns of $sM + A$ by $1/\sqrt{s}$, we get the matrix $\begin{pmatrix} sM_1 + A_0 & \sqrt{s}M_2^\mathsf{T} \\ \sqrt{s}M_2 & M_0 \end{pmatrix}$. Letting $s \to 0$, this tends to $B = \begin{pmatrix} A_0 & 0 \\ 0 & M_0 \end{pmatrix}$. Clearly, B has $a + b$ negative eigenvalues, and by the continuity of eigenvalues, the lemma follows. □

Lemma 6 *Let M be a well-signed G-matrix with one negative eigenvalue and with corank $d = \kappa(G)$, let u be the nullspace representation defined by M, and let C be a clique in G of size at most $\kappa(G)$ such that the origin belongs to the convex hull of $u(C)$. Then $G - C$ is disconnected.*

Proof Since the origin belongs to the convex hull of $u(C)$, we can write $0 = \sum_i a_i u_i$ with $a_i \geq 0$, $\sum_i a_i = 1$, and $a_i = 0$ if $i \notin C$. Let A be the matrix $-aa^\mathsf{T}$. Since a is nonzero, A has a negative eigenvalue.

Since $\sum_i a_i u_i = 0$, we have $\ker(M) \subseteq \ker(M + sA)$ for each s. This implies that $\operatorname{corank}(M + sA) \geq \operatorname{corank}(M)$ for each s. Moreover, $M + sA$ is a well-signed G-matrix for $s \geq 0$. So $M + sA \in \mathcal{W}$ for each $s \geq 0$. Hence, as $M \in \mathcal{W}^1$ and as $\mathcal{W}^1 \subseteq \mathcal{W}^=$ (since $d = \kappa(G)$), we know by Lemma 3 that $M + sA \in \mathcal{W}^1$ for every $s \geq 0$. In other words, $M + sA$ has one negative eigenvalue for every $s \geq 0$.

Let M_0 be the matrix obtained from M by deleting the rows and columns with index in C. Note that M_0 has no negative eigenvalue: otherwise by Lemma 5, $M + sA$ has at least two negative eigenvalues for some $s > 0$, a contradiction.

Now suppose that $G - C$ is connected. As $u(C)$ is linearly dependent and $|C| \leq \text{corank}(M)$, $\ker(M)$ contains a nonzero vector x with $x_i = 0$ for all $i \in C$. Then by the Perron–Frobenius theorem, $\text{corank}(M_0) = 1$ and $\ker(M_0)$ is spanned by a positive vector y. As G is connected, x is orthogonal to the positive eigenvector belonging to the negative eigenvalue of M. So x has both positive and negative entries. On the other hand, $x|_{V\setminus C} \in \ker(M_0)$, and so $x|_{V\setminus C}$ must be a multiple of y, a contradiction.

□

Taking C a singleton, we derive:

Corollary 7 *Let G be a 2-connected graph, let $M \in \mathcal{W}^1$ have corank $\kappa(G)$, and let u be the nullspace representation defined by M. Then $u_i \neq 0$ for every node i.*

Equivalently, the nullspace representation defined by M can be normalized by node scaling.

2.3 Auxiliary Algorithms

Now we turn to the algorithmic part, starting with some auxiliary algorithms. The following general argument will be needed repeatedly.

Algorithm 1 (Interpolation)

Input: a continuous family of full-row-rank matrices $U(t) \in \mathbb{R}^{d\times n}$, and a continuous family of symmetric matrices $M(t) \in \mathbb{R}^{n\times n}$ $(0 \leq t \leq 1)$ such that $U(t)M(t) = 0$, $M(0)$ has exactly one negative eigenvalue and $M(1)$ has at least two negative eigenvalues.

Output: a value $t \in [0, 1]$ for which $M(t)$ has at most one negative eigenvalue and at least $d + 1$ zero eigenvalues.

Let $X := \{t \mid \lambda_2(M(t)) \geq 0\}$ and $Y := \{t \mid \lambda_{d+2}(M(t)) \leq 0\}$. Since $U(t)M(t) = 0$ and $U(t)$ has full row rank, every matrix $M(t)$ has at least d zero eigenvalues. Hence $X \cup Y = [0, 1]$. Therefore, as X and Y are closed, and as X is a nonempty proper subset of $[0, 1]$ (since $0 \in X$, $1 \notin X$), we have $X \cap Y \neq \emptyset$, that is, $t \in X \cap Y$ for some t.

How to compute such a value of t? By binary search, we can compute it with arbitrary precision. In our applications, we can do better, since the entries of the families $U(t)$ and $M(t)$ will be (very simple) rational functions of t. We can find those values of t for which $M(t)$ has corank at least $d + 1$ by considering any nonsingular $(n - d) \times (n - d)$ submatrix of $M(0)$, and finding the roots of $\det(B(t)) = 0$, where $B(t)$ is the corresponding submatrix of $M(t)$. Then every value of t with $\text{corank}(M(t)) > d$ is one of these roots. The smallest such value of t will give a matrix $M(t)$ with corank at least $d+1$. Since the matrices $M(s)$ with $s < t$ have at most one negative eigenvalue (as otherwise $[0, t) \cap Y \neq \emptyset$ (since $X \cup Y = [0, 1]$), hence $[0, t) \cap X \cap Y \neq \emptyset$ (as $0 \in X$ and X and Y are closed), so $\text{corank}(M(s)) > d$ for some $s < t$), the matrix $M(t)$ has at most one.

We describe two simple applications of this general method.

Algorithm 2 (Double zero node)

Input: a connected graph $G = (V, E)$, a full-dimensional vector labeling u in $\mathbb{R}^d$, two nodes i and j with $u_i = u_j = 0$, and a matrix $M \in \mathcal{W}_u^1$.

Output: a matrix $M' \in \mathcal{W}_u^1$ with $\text{corank}(M') \geq d + 1$.

Subtract $t > 0$ from both diagonal entries M_{ii} and M_{jj}, to get a matrix $M(t)$. Trivially $M(t) \in \mathcal{W}_u$. Furthermore, if $t > 2\max\{|M_{ii}|, |M_{jj}|, |M_{ij}|\}$, then the submatrix of $M(t)$ formed by rows and columns i and j has negative trace and positive determinant, and so it has two negative eigenvalues. This implies by Interlacing Eigenvalues that $M(t)$ has at least two negative eigenvalues. Calling Algorithm 1, we get a $0 \leq s \leq t$ such that $M(s)$ has at most one negative eigenvalue and at least $d+1$ zeroes. Lemma 2 implies that $M(s)$ cannot be positive semidefinite, so $M(s) \in \mathcal{W}_u^1$.

Algorithm 3 (Zero node)

Input: a 2-connected graph $G = (V, E)$, a full-dimensional vector labeling u in $\mathbb{R}^d$, a node i with $u_i = 0$, and a matrix $M \in \mathcal{W}_u^1$.

Output: a matrix $M'' \in \mathcal{W}_u^1$ with $\text{corank}(M'') \geq d + 1$.

We may assume $i = 1$. Let N be the matrix obtained from M by deleting row and column 1. Any coordinate of the vectors u_j $(j \neq 1)$ is in the nullspace of N. Since $G \setminus 1$ is connected, the Perron–Frobenius Theorem implies that N is not positive semidefinite (otherwise $d = 1$ by Lemma 2(a), and then $(u_i \mid i \neq 1)$ would be the eigenvector of N belonging to the smallest eigenvalue 0, while this vector is not constant in sign). So N has a negative eigenvalue λ, with eigenvector y $(|y| = 1)$. Replacing M_{11} by a sufficiently small negative number s, we get a matrix $M' \in \mathcal{W}_u$ with two negative eigenvalues. Simple linear algebra shows that $s < \left(e_1^{\mathsf{T}} M \binom{0}{y}\right)^2 / \lambda$ suffices. We conclude by calling Algorithm 1 as before.

3 1-Dimensional Nullspace Representations

As a warmup, let us settle the case $d = 1$. For every connected graph $G = (V, E)$, it is easy to construct a singular G-matrix with exactly one negative eigenvalue: start with any G-matrix, and subtract an appropriate constant from the main diagonal. Our goal is to show that unless the graph is a path and the nullspace representation is a monotone embedding in the line, we can modify the matrix to get a G-matrix with one negative eigenvalue and with corank at least 2.

3.1 Nullspace and Neighborhoods

We start with noticing that given a vector $u \in \mathbb{R}^V$, it is easy to describe the matrices in $\mathcal{W}_u$. Indeed, consider any matrix $M \in \mathcal{M}_u$. Then for every node i with $u_i = 0$, we have

$$\sum_{j\in N(i)} M_{ij}u_j = \sum_j M_{ij}u_j = 0. \tag{4}$$

Furthermore, for every node i with $u_i \neq 0$, we have

$$M_{ii} = -\frac{1}{u_i}\sum_{j\in N(i)} M_{ij}u_j. \tag{5}$$

Conversely, if we specify the off-diagonal entries of a G-matrix M so that (4) is satisfied for each i with $u_i = 0$, then we can define M_{ii} for nodes $i \in \operatorname{supp}(u)$ according to (5), and for nodes i with $u_i = 0$ arbitrarily, we get a matrix in $\mathcal{M}_u$.

As an application of this construction, we prove the following lemma.

Lemma 8 *Let $u \in \mathbb{R}^V$. Then $\mathcal{W}_u \neq \emptyset$ if and only if for every node i with $u_i = 0$, either all its neighbors satisfy $u_j = 0$, or it has neighbors both with $u_j < 0$ and $u_j > 0$.*

Proof By the remark above, it suffices to specify negative numbers M_{ij} for the edges ij so that (4) is satisfied for each i with $u_i = 0$. The edges between two nodes with $u_i = 0$ play no role, and so the conditions (4) can be considered separately. For a fixed i, the single linear equation for the M_{ij} can be satisfied by negative numbers if and only if the condition in the lemma holds. □

We need the following fact about the neighbors of the other nodes.

Lemma 9 *Let $u \in \mathbb{R}^V$, $M \in \mathcal{W}_u$, and suppose that M has a negative eigenvalue $\lambda < 0$, with eigenvector $\pi > 0$. Then every node i with $u_i > 0$ has a neighbor j for which $u_j/\pi_j < u_i/\pi_i$.*

Proof Suppose not. Then $u_j \geq \pi_j u_i/\pi_i$ for every $j \in N(i)$, and so

$$0 = \sum_j M_{ij}u_j \leq M_{ii}u_i + \sum_{j\in N(i)} M_{ij}\frac{\pi_j}{\pi_i}u_i = \frac{u_i}{\pi_i}\Big(\sum_j M_{ij}\pi_j\Big) = \lambda u_i < 0,$$

a contradiction. □

Algorithm 4 (Double cover)

Input: a vector $u \in \mathbb{R}^V$, two edges ab and cd with $u_a u_b \leq 0$, $u_c u_d \leq 0$, $b \neq d$, $u_a \neq 0$, $u_c \neq 0$, and a matrix $M \in \mathcal{W}_u^1$.

Output: a matrix $M' \in \mathcal{W}_u^1$ of corank at least 2.

Define the symmetric matrix $N^{ab} \in \mathbb{R}^{V\times V}$ by

$$(N^{ab})_{ij} = \begin{cases} u_a u_b, & \text{if } \{i,j\} = \{a,b\},\\ -u_b^2, & \text{if } i = j = a,\\ -u_a^2, & \text{if } i = j = b,\\ 0, & \text{otherwise,}\end{cases}$$

and define N^{cd} analogously. Then $N^{ab}u = N^{cd}u = 0$, and so $M' = M + tN^{ab} + tN^{cd} \in \mathcal{W}_u$ for every $t > 0$. Moreover, $N^{ab} + N^{cd}$ has two negative eigenvalues, as one may (case-)check. So $M + tN^{ab} + tN^{cd} \in \mathcal{W}_u^2$ for some t, by Lemma 5. So with the Interpolation Algorithm 1 we find M' as required.

3.2 Embedding in the Line

Now we come to the main algorithm for dimension 1.

Algorithm 5
Input: A connected graph $G = (V, E)$.
Output: Either an embedding $u : V \to \mathbb{R}$ of G (then G is a path), or a well-signed G-matrix with one negative eigenvalue and corank at least 2.

Preparation We find a matrix $M \in \mathcal{W}^1$. This is easy by creating any well-signed G-matrix and subtracting its second smallest eigenvalue from the diagonal. We may assume that $\text{corank}(M) = 1$, else we are done.

Let $u \neq 0$ be a vector in the nullspace of M, and let π be an eigenvector belonging to its negative eigenvalue. Then the matrix $M' = \text{diag}(\pi)M\text{diag}(\pi)$ is in $\mathcal{W}^1(G)$ and the vector $w = (u_i/\pi_i : i \in V)$ is in its nullspace. By Lemma 9, this means that if we replace M by M' and u by w, then we get a vector $u \in \mathbb{R}^n$ and a matrix $M \in \mathcal{W}_u^1$ such that every node i with $u_i > 0$ has a neighbor j with $u_j < u_i$, and every node i with $u_i < 0$ has a neighbor j with $u_j > u_i$.

If $u_i = u_j = 0$ for some distinct i, j, we can apply Algorithm 2. So we can assume that $u_i = 0$ for at most one i.

Let us define a *cell* as an open interval between two consecutive points u_i. If every cell is covered by only one edge, then G is a path and u defines an embedding of G in the line, and we are done. Indeed, suppose first $u_i = u_j$ with $j \neq i$. By assumption $u_i \neq 0$. If $u_i > 0$, then both i and j have a neighbour i' and j' respectively, with $u_{i'} < u_i$ and $u_{j'} < u_j$, hence some cell is covered twice by edges. Similarly if $u_i < 0$. So the u_i are all distinct. Assuming that each cell is covered at most once by an edge, u must be an embedding of G into $\mathbb{R}$, and so G is a path.

So we can assume that there exists a cell (a, b) covered by at least two edges. We choose (a, b) nearest to the origin. Replacing u by $-u$ if necessary, we may assume that $b > 0$.

Main step Below, we are going to maintain the following conditions. We have a vector $u \in \mathbb{R}^V$ and a matrix $M \in \mathcal{W}_u^1$; every node i with $u_i > 0$ has a neighbor j with $u_j < u_i$; there is a cell (a, b) with $b > 0$ that is doubly covered, and that is nearest the origin among such cells.

We have to distinguish some cases.

Case 1. If $a < 0$, then we use the Double Cover Algorithm 4 to obtain a matrix with the desired properties.

Case 2. If $a \geq 0$, then let u_p be the smallest nonnegative entry of u.

Case 2.1. Assume that $u_p = 0$. Let $(0, c)$ be the cell incident with 0 and with $c > 0$, and let M' be obtained from M by replacing the (p, p) diagonal entry by 0. Then $M' \in \mathcal{W}_u$. It follows by Lemma 2(a) that M' is not positive semidefinite. If M' has more than one negative eigenvalue, then we can run the Interpolation Algorithm 1. So we may assume that $M' \in \mathcal{W}_u^1$.

For $0 < t < c$, consider the G-matrices $A(t)$ defined for edges ij by

$$A(t)_{ij} = A(t)_{ji} = \begin{cases} M_{ij}, & \text{if } i, j \neq p, \\ \dfrac{u_j}{u_j - t} M_{pj}, & \text{if } i = p, \end{cases}$$

and on the diagonal by

$$A(t)_{ii} = -\frac{1}{u_i - t} \sum_{j \in N(i)} A(t)_{ij}(u_j - t).$$

Clearly $A(t) \in \mathcal{W}_{u-t}$. Lemma 2(a) implies that $A(t)$ has at least one negative eigenvalue. Furthermore, if $t \to 0$, then $A(t)_{ij} \to M_{ij}$; this is trivial except for $i = j = p$, when, using that $\sum_{j \in N(p)} M_{pj} u_j = -M_{pp} u_p = 0$, we have

$$A(t)_{pp} = \frac{1}{t} \sum_{j \in N(p)} M_{pj} u_j = 0.$$

Thus defining $A(0) = M'$ the family $A(t)$ remains continuous.

If the matrix $A(c/2)$ has one negative eigenvalue, then replace M by $A(c/2)$ and u by $u - c/2$, and return to the Main Step. Note that the number of nodes with $u_i \geq 0$ has decreased, while those with $u_i > 0$ did not change.

If $A(c/2)$ has at least two negative eigenvalues, then the Interpolation Algorithm 1 can be applied to the families $(A(t))$ and $(u - t)$ to get a number $0 \leq s \leq c/2$ with $A(s) \in \mathcal{W}_{u-t}^1$ and $\operatorname{corank}(A(s)) > 1$.

Case 2.2. Assume that $u_p > 0$. Let σ and τ denote the cells to the left and to the right of u_p (so $0 \in \sigma$). There is no other node q with $u_q = u_p$ (since from both nodes, an edge would start to the left, whereas 0 is covered only once). From u_p, there is an edge starting to the left, and also one to the right (since by connectivity, there is an edge covering τ, and this must start at p, since σ is covered only once). Therefore, $\mathcal{W}_{u-u_p} \neq \emptyset$ by Lemma 8. Following the proof of this Lemma, we can construct a matrix $B \in \mathcal{W}_{u-u_p}$ with $B_{pp} = 0$. Since $u - u_p$ has a zero entry, Lemma 2(a) implies that B has at least one negative eigenvalue.

For $t \in [0, u_p)$, consider the G-matrices $B(t)$ defined for edges ij by

$$B(t)_{ij} = B(t)_{ji} = \begin{cases} B_{ij}, & \text{if } i, j \neq p, \\ \dfrac{u_j - u_p}{u_j - t} B_{pj}, & \text{if } i = p, \end{cases}$$

and on the diagonal by

$$B(t)_{ii} = -\frac{1}{u_i - t} \sum_{j \in N(i)} B(t)_{ij}(u_j - t).$$

Clearly, $B(t) \in \mathcal{W}_{u-t}$. Furthermore, $\lim_{t \to u_p} B(t) = B$.

If B has one negative eigenvalue, then replace M by B and u by $u - u_p$, and go to the Main Step. Note that the number of nodes with $u_i > 0$ has decreased, while those with $u_i \geq 0$ did not change.

If B has more than one negative eigenvalue and $B(0)$ has only one, then the Interpolation Algorithm 1 gives a value $0 \leq s < u_p$ such that $B(s) \in \mathcal{W}^1_{u-s}$ and $\text{corank}(B(s)) > 1$.

Finally, if $B(0)$ has more than one negative eigenvalue, then we call the Interpolation Algorithm 1 for the family of matrices $(1-t)M + tB(0)$, keeping u fixed.

4 2-Dimensional Nullspace Representations

4.1 G-Matrices and Circulations

Our goal in this section is to provide a characterization of G-matrices and their nullspace representations in dimension 2.

A *circulation* on an undirected simple graph G is a real $V \times V$ matrix f such that it is supported on adjacent pairs, is skew-symmetric and satisfies the flow conditions:

$$f_{ij} = 0 \ (ij \notin E), \quad f_{ij} = -f_{ji} \ (i,j \in V), \quad \sum_j f_{ij} = 0 \ (i \in V).$$

If we fix an orientation of the graph, then it suffices to specify the values of f on the oriented edges; the values on the reversed edges follow by skew symmetry. A *positive circulation* on an oriented graph (V, A) is a circulation on the underlying undirected graph that takes positive values on the arcs in A.

For any representation $u : V \to \mathbb{R}^2$, we define its *area-matrix* as the (skew-symmetric) matrix $T = T(u)$ by $T_{ij} := \det(u_i, u_j)$. This number is the signed area of the parallelogram spanned by u_i and u_j. If R denotes counterclockwise rotation by 90°, then $T_{ij} = u_i^{\mathsf{T}} R u_j$.

Given a graph G and a representation $u : V \to \mathbb{R}^2$ by nonzero vectors, we define a directed graph (V, A_u) and an undirected graph (V, E_u) by

$$\begin{aligned} A_u &:= \{(i,j) \in V \times V \mid ij \in E, T(u)_{ij} > 0\} \\ E_u &:= \{ij \in E \mid T(u)_{ij} = 0\}. \end{aligned}$$

So E is partitioned into A_u and E_u, where (V, A_u) is an oriented graph in which each edge is oriented counterclockwise as seen from the origin. The graph (V, E_u) consists of all edges that are contained in a line through the origin.

Given a representation $u : V \to \mathbb{R}^2$, a circulation f on (V, A_u), and a function $g : E_u \to \mathbb{R}$, we define a G-matrix $M(u,f,g)$ by

$$M(u,f,g)_{ij} = \begin{cases} -f_{ij}/T(u)_{ij}, & \text{if } ij \in A_u, \\ g(ij), & \text{if } ij \in E_u. \end{cases}$$

We define the diagonal entries by (1), and let the other entries be 0. The first main ingredient of our proof and algorithm is the following representation of G-matrices with a given nullspace.

Lemma 10 *Let $G = (V, E)$ be a graph and let $u : V \to \mathbb{R}^2$ be a labeling of V by nonzero vectors. Then*

$$\mathcal{M}_u = \big\{M(u,f,g) : f \textit{ is a circulation on } (V, A_u) \text{ and } g : E_u \to \mathbb{R}\big\}.$$

Proof First, we prove that $M(u,f,g) \in \mathcal{M}_u$ for every circulation on (V, A_u) and every $g : E_u \to \mathbb{R}$. Using that $M(u,f,g) = M(u,f,0) + M(u,0,g)$, it suffices to prove that $M(u,f,g) \in \mathcal{M}_u$ if either $f = 0$ or $g = 0$. If $M = M(u,f,0)$, then using that f is a circulation, we have

$$\Big(\sum_j M_{ij}u_j\Big)^{\mathsf{T}} Ru_i = \sum_j f_{ij} = 0.$$

This means that $\sum_j M_{ij}u_j^{\mathsf{T}}$ is orthogonal to Ru_i, and so parallel to u_i. As remarked above, this means that $M(u,f,0) \in \mathcal{M}_u$. If $M = M(u,0,g)$, then for every $i \in V$,

$$\sum_{j \in N(i)} M_{ij}u_j = \sum_{j:\, ij \in E_u} g(ij)u_j$$

This vector is clearly parallel to u_i, proving that $M(u,0,g) \in \mathcal{M}_u$.

Second, given a matrix $M \in \mathcal{M}_u$, define $f_{ij} = -T_{ij}M_{ij}$ for $ij \in A_u$ and $g_{ij} = M_{ij}$ for $ij \in E_u$. Then f is a circulation. Indeed, for $i \in V$,

$$\sum_{ij \in A_u} f_{ij} = -\sum_{ij \in A_u} M_{ij}u_j^{\mathsf{T}} Ru_i = -\sum_{j \in V} M_{ij}u_j^{\mathsf{T}} Ru_i = \Big(-\sum_{j \in V} M_{ij}u_j\Big)^{\mathsf{T}} Ru_i = 0.$$

Furthermore, $M(u,f,g) = M$ by simple computation. □

Note that the G-matrix $M(u,f,g)$ is well-signed if and only if f is a positive circulation on (V, A_u) and $g < 0$. Thus,

Corollary 11 *Let $G = (V, E)$ be a graph, let $u: V \to \mathbb{R}^2$ be a representation of V by nonzero vectors. Then*

$$\mathcal{W}_u = \big\{M(u,f,g): \text{ } f \text{ is a positive circulation on } (V, A_u), \\ g: E_u \to \mathbb{R},\ g < 0\big\}.$$

In particular, it follows that $\mathcal{W}_u \neq \emptyset$ if and only if A_u carries a positive circulation. This happens if and only if each arc in A_u is contained in a directed cycle in A_u; that is, if and only if each component of the directed graph (V, A_u) is strongly connected.

The signature of eigenvalues of $M(u,f,g)$ is a more difficult question, but we can say something about $M(u,0,g)$ if $g < 0$. Let H be a connected component of the graph (V, E_u), and let M_H be the submatrix of $M(u,0,g)$ formed by the rows and columns whose index belongs to $V(H)$. Then M_H is a well-signed H-matrix. The vectors u_i representing nodes $i \in V(H)$ are contained in a single line through the origin. Lemma 2 implies that M_H has at least one negative eigenvalue unless $u(V(H))$ is contained in a semiline starting at the origin. Let us call such a component *degenerate*. Then we can state:

Lemma 12 *Let $u: V \to \mathbb{R}^2$ be a representation of V with nonzero vectors, and let $g: E_u \to \mathbb{R}$ be a function with negative values. Then the number of negative eigenvalues of $M(u,0,g)$ is at least the number of nondegenerate components of (V, E_u).*

4.2 Shifting the Origin

Consider the cell complex made by the (two-way infinite) lines through distinct points u_i and u_j with $ij \in E$. The 1- and 2-dimensional cells are called 1-*cells* and 2-*cells*, respectively. Two cells c and d are *incident* if $d \subseteq \overline{c} \setminus c$ or $c \subseteq \overline{d} \setminus d$.

Two points p and q belong to the same cell if and only if $A_{u-p} = A_{u-q}$ and $E_{u-p} = E_{u-q}$. Hence, for any cell c, we can write A_c and E_c for A_{u-p} and E_{u-p}, where p is an arbitrary element of c. For any cell c, set $\mathcal{W}_c := \bigcup_{p \in c} \mathcal{W}_{u-p}$. It follows by Lemma 10 that if $\mathcal{W}_c \neq \emptyset$, then $\mathcal{W}_{u-p} \neq \emptyset$ for every $p \in c$. It also follows that $\mathcal{W}_c$ is connected for each cell c, as it is the range of the continuous function $M(u-p,f,g)$ on the connected topological space of triples (p,f,g) where $p \in c, f$ is a positive circulation on A_c, and g is a negative function on E_c.

The following lemma is an essential tool in the proof.

Lemma 13 *Let c be a cell with $\mathcal{W}_c \neq \emptyset$ and let $q \in \overline{c}$. Then $M(u-q,0,g) \in \overline{\mathcal{W}_c}$ for some negative function g on E_{u-q}.*

Proof Choose any $p \in c$. Note that $q \in \overline{c}$ implies that $E_{u-p} \subseteq E_{u-q}$. Let $M \in \mathcal{W}_{u-p}$. Define $g(ij) = M_{ij}$ for $ij \in E_{u-q}$ and let $N = M(u-q, 0, g)$. We prove that N belongs to $\overline{\mathcal{W}_c}$.

By Lemma 10 we can write $M = M(u-p, f, g')$ with some positive circulation f on A_{u-p} and negative function g' on E_{u-p}. Define $g(ij) = M_{ij}$ for $ij \in E_{u-q}$ and let $N = M(u-q, 0, g)$. For $\alpha \in (0, 1]$, define $p_\alpha = (1-\alpha)q + \alpha p$, and consider the G-matrices $M_\alpha = M(u - p_\alpha, \alpha f, g')$. Clearly $M_\alpha \in \mathcal{W}_c$. We show $\lim_{\alpha \to 0} M_\alpha = N$.

Let $f_0 = \max_{ij \in E} |f_{ij}|$, $\ell = \min_{u_i \neq u_j} |u_i - u_j|$, $\beta = \max_{i,j} |u_i|/|u_j|$, and let δ denote the distance of q from the closest edge in A_{u-q}. Let $0 < \alpha \leq \delta/|q-p|$. It suffices to prove that

$$\|M_\alpha - N\|_\infty \leq \frac{4\alpha\beta f_0}{\delta\ell}, \tag{6}$$

which implies that $M_\alpha \to N$ as $\alpha \to 0$.

- If $ij \in E_{u-p}$, then $(M_\alpha)_{ij} = N_{ij} = g'(ij)$, independently of α.
- If $ij \in E_{u-q} \setminus E_{u-p}$, then for each $\alpha \in (0, 1]$ we have $ij \notin E_{u-\alpha p}$. The points u_i, u_j, and q are collinear, hence $T(u-p_\alpha)_{ij} = \alpha T(u-p)_{ij}$ for each $\alpha \in (0, 1]$. Thus

$$(M_\alpha)_{ij} = \frac{-\alpha f_{ij}}{T(u-p_\alpha)_{ij}} = \frac{-f_{ij}}{T(u-p)_{ij}} = M_{ij} = g_{ij} = N_{ij}. \tag{7}$$

- If $ij \in E \setminus E_{u-q}$, then $N_{ij} = 0$ and

$$|T(u-p_\alpha)_{ij}| \geq |T(u-q)_{ij}| - \frac{1}{2}|q-p_\alpha|\,|u_i - u_j| \geq \frac{1}{2}(\delta - \alpha|q-p|)|u_i - u_j| \geq \frac{1}{4}\delta\ell.$$

  So

$$|N_{ij} - (M_\alpha)_{ij}| = |(M_\alpha)_{ij}| \leq \alpha\frac{4f_0}{\delta\ell}.$$

- If $i, j \in V$ with $ij \notin E$ and $i \neq j$, then $(M_\alpha)_{ij} = 0 = N_{ij}$.
- For the diagonal, (1) gives that

$$|N_{ii} - (M_\alpha)_{ii}| \leq \sum_{j \in N(i)} |N_{ij} - (M_\alpha)ij| \frac{|u_j^\mathsf{T} u_i|}{u_i^\mathsf{T} u_i} \leq \alpha\beta\frac{4f_0}{\delta\ell}.$$

This proves (6). □

Corollary 14 *Let c be a cell with $\mathcal{W}_c \neq \emptyset$ and $q \in \overline{c}$. Then for every matrix $M \in \mathcal{W}_{u-q}$ there is a matrix $M' \in \mathcal{W}_{u-q} \cap \overline{\mathcal{W}_c}$ that differs from M only on entries corresponding to edges in E_{u-q} and on the diagonal entries.*

Proof By Lemma 10 we can write $M = M(u-q,f,g)$ with some positive circulation f on A_{u-q} and negative function g on E_{u-q}. By Lemma 13, there is a negative function g' on E_{u-q} such that $M(u-q,0,g') \in \overline{\mathcal{W}}_c$. There are points $p_k \in c$ and matrices $M_k \in \mathcal{W}_{u-p_k}$ such that $M_k \to M(u-q,0,g')$ as $k \to \infty$. Then $M_k + M(u-p_k,f,0)$ belongs to $\mathcal{W}_{u-p_k}$ and $M_k + M(u-p_k,f,0) \to M(u-q,0,g') + M(u-q,f,0) = M(u-q,f,g')$ as $k \to \infty$, showing that $M' = M(u-q,f,g')$ belongs to $\overline{\mathcal{W}}_c$. Furthermore, $M - M' = M(u-q,0,g-g')$ is nonzero on entries in E_{u-q} and on the diagonal entries only. □

Corollary 15 *If c and d are incident cells, then $\mathcal{W}_c \cup \mathcal{W}_d$ is connected.*

Proof We may assume that $d \subseteq \overline{c} \setminus c$, and that both $\mathcal{W}_c$ and $\mathcal{W}_d$ are nonempty (otherwise the assertion follows from the connectivity of $\mathcal{W}_c$ and $\mathcal{W}_d$).

Choose $q \in d$. Since $\mathcal{W}_d \neq \emptyset$, Corollary 14 implies that $\mathcal{W}_d$ and $\overline{\mathcal{W}}_c$ intersect, and by the connectivity of $\mathcal{W}_c$ and $\mathcal{W}_d$, this implies that $\mathcal{W}_c \cup \mathcal{W}_d$ is connected. □

Call a segment σ in the plane *separating*, if σ connects points u_a and u_b for some $a, b \in V$, with the property that $V \setminus \{a, b\}$ can be partitioned into two nonempty sets X and Y such that no edge of G connects X and Y and such that the sets $\{u_i \mid i \in X\}$ and $\{u_i \mid i \in Y\}$ are on distinct sides of the line through σ. Note that this implies that σ is a 1-cell.

Lemma 16 *Let G be a connected graph, and let σ be a separating segment connecting u_i and u_j, with incident 2-cells R and Q. If $\mathcal{W}_\sigma \cup \mathcal{W}_R \neq \emptyset$, then A_Q contains a directed circuit traversing ij.*

Proof We may assume that σ connects u_1 and u_2, and that edge 12 of G is oriented from 1 to 2 in A_Q. Let ℓ be the line through σ, and let H and H' be the open halfplanes with boundary ℓ containing Q and R, respectively.

Choose $p \in \sigma \cup R$ with $\mathcal{W}_{u-p} \neq \emptyset$. Note that A_Q and A_{u-p} differ only for edge 12. Any edge $ij \neq 12$ has the same orientation in A_Q as in A_{u-p}.

Since H contains points u_i, since G is connected, and since ℓ crosses no u_iu_j with $ij \in E$, G has an edge $1k$ or $2k$ with $u_k \in H$. By symmetry, we can assume that $2k$ is an edge. Then in A_{u-p}, edge $2k$ is oriented from 2 to k. As $\mathcal{W}_{u-p} \neq \emptyset$, A_{u-p} has a positive circulation. So A_{u-p} contains a directed circuit D containing $2k$. The edge preceding $2k$, say $j2$, must have $u_j \in H'$, as p belongs to $\sigma \cup R$. Therefore, since $\{1, 2\}$ separates nodes k and j, D traverses node 1. So the directed path in D from 2 to 1 together with the edge 12 forms the required directed circuit C in A_{u-q}. □

Corollary 17 *Let G be a connected graph, let σ be a separating segment, and let R be a 2-cell incident with σ. Then $\mathcal{W}_\sigma \neq \emptyset$ if and only if $\mathcal{W}_R \neq \emptyset$.*

Proof Let σ connect u_1 and u_2. If $\mathcal{W}_\sigma \neq \emptyset$, then A_σ has a positive circulation f'. By Lemma 16, A_R contains a directed circuit C traversing 12. Let f be the incidence vector of C. Then $f' + f$ is a positive circulation on A_R. So $\mathcal{W}_R \neq \emptyset$.

Conversely, if $\mathcal{W}_R \neq \emptyset$, then A_R has a positive circulation f. By Lemma 16, A_R contains a directed cycle through the arc 21, which gives a directed path P from 1 to 2 not using 12. It follows that by rerouting f_{12} over P, we obtain a positive circulation on A_σ, showing that $\mathcal{W}_\sigma \neq \emptyset$. □

4.3 Outerplanar Nullspace Embeddings

Let $G = (V, E)$ be a graph. A mapping $u : V \to \mathbb{R}^2$ is called *outerplanar* if its extension to the edges gives an embedding of G in the plane, and each u_i is incident with the unbounded face of this embedding.

Theorem 18 *Let G be a 2-connected graph with $\kappa(G) = 2$. Then the normalized nullspace representation defined by any well-signed G-matrix with one negative eigenvalue and with corank 2 is an outerplanar embedding of G.*

Proof Let u be such a normalized nullspace representation (this exists by Corollary 7). Let K be the convex hull of $u(V)$. Since all u_i have unit length, each u_i is a vertex of K. We define a *diagonal edge* as the line segment connecting points $u_i \neq u_j$, where $ij \in E$. We don't know at this point that the points u_i are different and that diagonal edges do not cross; so the same diagonal edge may represent several edges of G, and may consist of several 1-cells.

Let P denote the set of points $p \in \mathbb{R}^2 \setminus u(V)$ with $\mathcal{W}^1_{u-p} \neq \emptyset$. Clearly, the origin belongs to P. Lemma 2(b) implies that

Claim 1 *P is contained in the interior of K.*
(It will follow below that P is equal to the interior of K.)

Consider again the cell complex into which the diagonal edges cut K. By the connectivity of the sets $\mathcal{W}_c$ and by Lemma 3, P is a union of cells.

Claim 2 *$\overline{P}$ cannot contain a point $u_i = u_j$ for two distinct nodes i and j.*

Indeed, since $u_i = u_j$ is a vertex of the convex hull of $u(V)$, we can choose $p \in P$ close enough to u_i so that it is not in the convex hull of $u(V) \setminus \{u_i\}$. This, however, contradicts Lemma 4.

Claim 3 *No point $p \in \overline{P} \setminus u(V)$ is contained in two different diagonal edges.*

Indeed, consider any cell $c \subseteq P$ with $p \in \overline{c}$. Since $\mathcal{W}_c \neq \emptyset$, Lemma 13 implies that there is a negative function g on E_{u-p} such that $M(u - p, 0, g) \in \overline{\mathcal{W}_c}$. As all matrices in $\mathcal{W}_c$ have exactly one negative eigenvalue, $M(u - p, 0, g)$ has at most one negative eigenvalue. Lemma 12 implies that (V, E_{u-p}) has at most one nondegenerate component. But every diagonal containing p is contained in a nondegenerate component of (V, E_{u-p}), and these components are different for different diagonals, so p can be contained in at most one diagonal. This proves Claim 3.

It is easy to complete the proof now. Clearly, P is bounded by one or more polygons. Let p be a vertex of $\overline{P}$, and assume that $p \notin u(V)$. Then p belongs to two diagonals (defining the edges of P incident with p), contradicting Claim 3. Thus all vertices of P are contained in $u(V)$. This implies that $\overline{P}$ is a convex polygon spanned by an appropriate subset of $u(V)$.

To show that $\overline{P} = K$, assume that the boundary of P has an edge σ contained in the interior of K and let $R \subseteq P$ be a 2-cell incident with σ, and let Q be the 2-cell incident with σ on the other side. Clearly, $\mathcal{W}_R \neq \emptyset$, and by Corollary 17, $\mathcal{W}_\sigma \neq \emptyset$ and by the same Corollary, $\mathcal{W}_Q \neq \emptyset$. The sets $\mathcal{W}_\sigma \cup \mathcal{W}_R$ and $\mathcal{W}_\sigma \cup \mathcal{W}_Q$

are connected by Corollary 15, and hence so is $\mathcal{W}_\sigma \cup \mathcal{W}_R \cup \mathcal{W}_Q$. We also know that $\mathcal{W}^1 \cap \mathcal{W}_R \neq \emptyset$. Since $\mathcal{W}^1$ is open and closed in $\mathcal{W}$ (Lemma 3, note that in this case $\mathcal{W}^1 = \mathcal{W}^1 \cap \mathcal{W}^=$ as $\kappa(G) = 2$), we conclude that $\mathcal{W}^1 \cap \mathcal{W}_Q \neq \emptyset$, i.e., $Q \subseteq P$. But this contradicts the definition of σ.

Thus P is equal to the interior of K. Claim 2 implies that the points u_i are all different, and Claim 3 implies that the diagonals do not cross. ■

4.4 Algorithm

The considerations in this section give rise to a polynomial algorithm achieving the following.

Algorithm 6

Input: A 2-connected graph $G = (V, E)$.

Output: Either an outerplanar embedding $u : V \to \mathbb{R}^2$ of G, or a well-signed G-matrix with one negative eigenvalue and corank at least 3.

The algorithm progresses along the same lines as the algorithm in Sect. 3.2. We describe the main steps, omitting some details. It will be useful to remember that by Lemma 2(a), no well-signed G-matrix with two zero eigenvalues is positive semidefinite.

Step 1. We call Algorithm 5, which returns a well-signed G-matrix M with one negative and at least two zero eigenvalues (since the graph is not a path). If it has three zero eigenvalues, we are done, so suppose that this is not the case. We compute its nullspace representation u. We compute a positive circulation f on (G, u) and a negative function g on E_u such that $M = M(u,f,g)$, following the simple formulas in the proof of Lemma 10.

If $M(u,f,0)$ has two negative eigenvalues, then the Interpolation Algorithms, applied with the matrix family $M(s) = (1-t)M + tM(u,f,0)$, returns a number $0 \leq s < 1$ for which $M(s)$ a well-signed G-matrix with one negative eigenvalue and corank at least 3. So suppose that $M(u,f,0)$ has one negative eigenvalue.

Step 2. If there is an i with $u_i = 0$, then Algorithm 3 gives a matrix $M'' \in \mathcal{W}_u$ with one negative and at least three zero eigenvalues. So we may assume that $u_i \neq 0$ for every i. We scale M so that $|u_i| = 1$. (All we are going to use of this condition is that every u_i is a vertex of the convex hull K of the vectors u_i.) Lemma 2 implies that $0 \in \text{int}(K)$; let c be the cell containing 0 (this may be a point, and edge, or a polygon).

If u is an outerplanar embedding, we are done. Otherwise, we have either two nodes $i,j \in V$ with $u_i = u_j$, or two (diagonal) edges that intersect. Let $z \in K$ be a point that is either the intersection point of two diagonal edges, or $z = u_i = u_j$ for two nodes i and j. Choose z so that the number of diagonal edges separating z from 0 is minimal.

Step 3. If $z = 0$ (equivalently, c is 0-dimensional), then the origin is the intersection point of two diagonal edges, and hence $M(u, 0, -1)$ has at least two negative eigenvalues. So we can apply Algorithm 1 with the matrix family $tM + (1-t)M(u,0,-1)$ (keeping u fixed).

Step 4. Suppose that we find two matrices $M \in \mathcal{W}^1_{u-p}$ and $M' \in \mathcal{W}^2_{u-q}$ where $p,q \in c$. Since p and q belong to the same cell, the matrix $M(u-q,f,g)$ is well defined and $M(u-q,f,g) \in \mathcal{W}_{u-q}$. If $M(u-q,f,g)$ has one negative eigenvalue, then we invoke Algorithm 1 with the family $(1-t)M' + tM(u-q,f,g)$ (keeping $u-q$ fixed). If $M(u-q,f,g)$ has at least two negative eigenvalues, then similarly invoke Algorithm 1 with the family $M(u-tp-(1-t)q,f,g)$ and $(u-tp-(1-t)q)$ for $0 \le t \le 1$.

Step 5. Suppose that no diagonal edge separates z from the origin, and z is the intersection point of at least two diagonal edges. Choose a number α such that

$$0 < \alpha < \min\left\{1, \frac{\delta}{|z|}, \frac{\delta\ell n}{4\beta f_0}\right\},$$

where the numbers β, f_0, δ, ℓ are defined as in the proof of Lemma 13 and are easily computed. As in the proof of Lemma 13, we construct a negative function g on E_{u-z} and a matrix $M_\alpha \in \mathcal{W}_{u-(1-\alpha)z}$ such that the matrix $N = M(u-(1-\alpha)z, 0, g)$ satisfies $\|M_\alpha - N\|_\infty < 4\alpha\beta f_0/(\delta\ell)$. The matrix N has at least two negative eigenvalues. Then elementary linear algebra gives that the matrix M_α has at least two negative eigenvalues. We conclude by Step 4.

Step 6. Suppose that every vertex of c is in $u(V)$, and c has a vertex $z = u_i = u_j$. Let q be a point in the interior of c but not in $\mathrm{conv}(u(V) \setminus \{z\}$. Then by Lemma 4, the matrix $M(u-q,f,g)$ has either corank at least 3 or two negative eigenvalues. In the first case, we are done; in the second, we invoke Step 4. So we may assume that z is not a vertex of c.

Step 7. If $c = [u_i, u_j]$ is a diagonal (intersecting no other diagonal), then let $\varepsilon > 0$ be small enough so that εz belongs to a region R bounded by c. By the construction in the proof of Lemma 17, we find a directed cycle C in G that passes every edge in the positive direction when viewed from R. Let h denote the unit flow around C, and let $M' = M(u-\varepsilon z, f + \varepsilon^2 h, 0) \in \mathcal{W}_{\varepsilon z}$.

If M' has one negative eigenvalue, then we can replace M by M' and u by $u-\varepsilon z$, to get an instance where the segment $[\varepsilon z, z]$ intersects fewer diagonal edges than $[0,z]$. If $M' \in \mathcal{W}^2_{u-\varepsilon z}$, then we apply the interpolation argument to the family $M(t) = M(u-tz, f+t^2 h, 0)$, using that $M(0)$ has one negative eigenvalue (as it is the limit of $M(u,f,\beta g)$ as $\beta \to 0$) and $M(1) = M'$. (The coefficient of h is t^2 to make sure that $M(t)$ depends continuously on t at $t = 0$.)

Step 8. So we may assume that c is a 2-dimensional polygon, 0 is an internal point of it, every vertex of c is the position of exactly one node, and so every edge of c is a full diagonal edge. Let q be the intersection point of $[0,z]$ with the boundary of c. Let $ij \in A_u$ be the edge for which $q \in [u_i, u_j]$, and let Q be the region on the other side of e, let C be a cycle through e in A_Q whose edges are counterclockwise when viewed from Q (constructed as in Lemma 17). Let h denote the unit flow around C, and let $M' = M(u-q, f + f_{ij}h, -1)$. Then $M' \in \mathcal{W}_{u-q}$. If it has one negative eigenvalue, then we can replace M by M' and u by $u - q$. If $M' \in \mathcal{W}^2_{u-q}$, then we apply the Interpolation Algorithm 1 to the matrix family $M(t) = M(tq, f + tf_{ij}h, 0)$.

Acknowledgements We thank Bart Sevenster for helpful discussion on $\kappa(G)$, and two referees for useful suggestions improving the presentation.

References

1. Y. Colin de Verdière, Sur un nouvel invariant des graphes et un critère de planarité. J. Combin. Theory Ser. B **50**, 11–21 (1990) [English transl.: On a new graph invariant and a criterion for planarity, in: *Graph Structure Theory*, ed. by N. Robertson, P. Seymour (American Mathematical Society, Providence, 1993), pp. 137–147]
2. R. Connelly, Rigidity and energy. Invent. Math. **66**, 11–33 (1982)
3. M. Laurent, A. Varvitsiotis, Positive semidefinite matrix completion, universal rigidity and the Strong Arnold Property. Linear Algebra Appl. **452**, 292–317 (2014)
4. L. Lovász, Steinitz representations and the Colin de Verdière number. J. Combin. Theory B **82**, 223–236 (2001)
5. L. Lovász, A. Schrijver, A Borsuk theorem for antipodal links and a spectral characterization of linklessly embeddable graphs. Proc. Am. Math. Soc. **126**, 1275–1285 (1998)
6. L. Lovász, A. Schrijver, On the null space of a Colin de Verdière matrix. Annales de l'Institut Fourier, Université de Grenoble **49**, 1017–1026 (1999)
7. N. Robertson, P. Seymour, R. Thomas, Sachs' linkless embedding conjecture. J. Combin. Theory Ser. B **64**, 185–227 (1995)
8. A. Schrijver, B. Sevenster, The Strong Arnold Property for 4-connected flat graphs. Linear Algebra Appl. **522**, 153–160 (2017)
9. H. van der Holst, A short proof of the planarity characterization of Colin de Verdière. J. Combin. Theory Ser. B **65**, 269–272 (1995)
10. H. van der Holst, Topological and spectral graph characterizations. Ph.D. thesis, University of Amsterdam, Amsterdam (1996)
11. H. van der Holst, L. Lovász, A. Schrijver, The Colin de Verdière graph parameter, in *Graph Theory and Combinatorial Biology*. Bolyai Society Mathematical Studies, vol. 7 (János Bolyai Mathematical Society, Budapest, 1999), pp. 29–85

Homology of Spaces of Directed Paths in Euclidean Pattern Spaces

Roy Meshulam and Martin Raussen

In memory of Jirka Matoušek.

Abstract Let $\mathcal{F}$ be a family of subsets of $\{1, \ldots, n\}$ and let

$$Y_{\mathcal{F}} = \bigcup_{F \in \mathcal{F}} \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_i \in \mathbb{Z} \text{ for all } i \in F\}.$$

Let $X_{\mathcal{F}} = \mathbb{R}^n \setminus Y_{\mathcal{F}}$. For a vector of positive integers $\mathbf{k} = (k_1, \ldots, k_n)$ let $\vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k}+\mathbf{1}}$ denote the space of monotone paths from $\mathbf{0} = (0, \ldots, 0)$ to $\mathbf{k} + \mathbf{1} = (k_1 + 1, \ldots, k_n + 1)$ whose interior is contained in $X_{\mathcal{F}}$. The path spaces $\vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k}+\mathbf{1}}$ appear as natural examples in the study of Dijkstra's PV-model for parallel computations in concurrency theory.

We study the topology of $\vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k}+\mathbf{1}}$ by relating it to a subspace arrangement in a product of simplices. This, in particular, leads to a computation of the homology of $\vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k}+\mathbf{1}}$ in terms of certain order complexes associated with the hypergraph $\mathcal{F}$.

The first author is supported by ISF and GIF grants.

Both authors acknowledge support from the ESF research networking programme ACAT.

R. Meshulam (✉)
Department of Mathematics, Technion, 32000 Haifa, Israel
e-mail: meshulam@math.technion.ac.il

M. Raussen
Department of Mathematical Sciences, Aalborg University, Fredrik Bajersvej 7G, 9220 Aalborg Øst, Denmark
e-mail: raussen@math.aau.dk

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_24

1 Introduction

Concurrency theory in computer systems deals with properties of systems in which several computations are executing simultaneously and potentially interacting with each other. Among the many models suggested for the study of concurrency are the Higher Dimensional Automata (HDA) introduced by Pratt [10]. Those arise as cubical complexes in which individual cubes (of varying dimension) with directed paths on each of them, are glued together consistently. Compared to other concurrency models, HDA have the highest expressive power based on their ability to represent causal dependence [15]. On the other hand, only little is known in general about the topology of the space of directed paths of a general HDA [12].

A specific simple case of linear HDA's consists of the PV-model suggested by Dijkstra [4] back in the 1960s. In this model there are m resources (e.g. shared memory sites) $a_1, \dots, a_m$ with positive integer capacities $\kappa(a_1), \dots, \kappa(a_m)$ where $\kappa(a_i)$ indicates the maximal number of processes that a_i can serve at any given time, and n linear processes $T_1, \dots, T_n$ (without branchings or loops) that require access to these resources. Given a resource a and a process T, denote by Pa and Va the locking and respectively unlocking of a by T. A process T_i is specified by a sequence of locking and unlocking operations on the various resources in a certain order. Modeling each process T_i as an ordered sequence of integer points on the interval $(0, k_i]$, one can view a legal execution of $\mathbf{T} = (T_1, \dots, T_n)$ as a coordinate-wise non-decreasing continuous path from $\mathbf{0} = (0, \dots, 0)$ to $\mathbf{k} + \mathbf{1} = (k_1 + 1, \dots, k_n + 1)$ that avoids a forbidden region determined by the processes and by the capacities of the resources. If two such paths are homotopic via a homotopy respecting the monotonicity condition then corresponding concurrent computations along the two paths have always the same result [5, 6].

Let $X_{\mathbf{T},\kappa}$ denote the complement of the forbidden region in $\prod_{i=1}^{n}[0, k_i + 1]$. The trace space $\overrightarrow{P}(X_{\mathbf{T},\kappa})_{\mathbf{0}}^{\mathbf{k+1}}$ associated with the pair $(\mathbf{T}, \kappa)$ consists of all paths as above endowed with the compact-open topology. For example, for the two processes sharing two resources depicted in Fig. 1, the forbidden region is the "Swiss Flag" and the trace space is homotopy equivalent to the two point space S^0. For an analysis of the PV spaces $X_{\mathbf{T},\kappa}$ and their associated trace spaces $\overrightarrow{P}(X_{\mathbf{T},\kappa})_{\mathbf{0}}^{\mathbf{k+1}}$, we refer to [5, 6, 11, 18].

In this paper we consider a special class of PV models in which the access and release of every resource happen *without time delay*. In this case, the forbidden region is a union of sets of the form $B \cap (K_1 \times \cdots \times K_n)$, where B is a fixed aligned box and each K_i is either $\mathbb{Z}$ or $\mathbb{R}$. Our main result (see Theorem 1.3 below) is a formula for the Poincaré series of the trace spaces associated to such special PV models. We proceed with some formal definitions leading to the statement of Theorem 1.3.

Let X be a subspace of $\mathbb{R}^n$. A continuous path $\mathbf{p} = (p_1, \dots, p_n) : I = [0, 1] \to X \subset \mathbb{R}^n$ is called *directed* if all components $p_i : I \to \mathbb{R}$ are non-decreasing. For two points y_0 and y_1 in the closure of X, let $\vec{P}(X)_{\mathbf{y}_0}^{\mathbf{y}_1}$ be the space of all directed paths in $\bar{X}$ (endowed with the compact-open topology) starting at $\mathbf{y}_0$ and ending at $\mathbf{y}_1$ whose interior is contained in X.

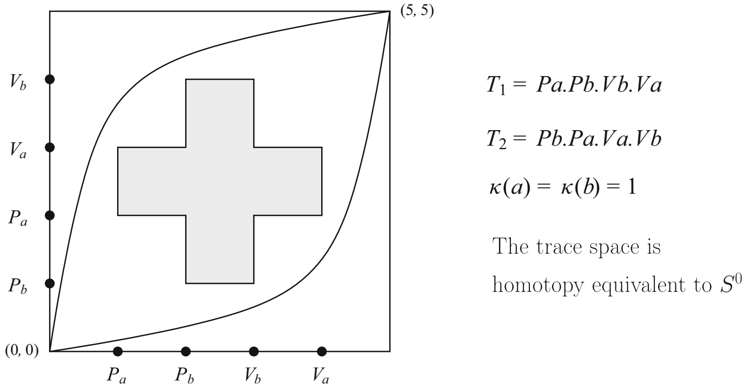


Fig. 1 The Swiss flag example – two processes sharing two resources

Let $\mathbb{N}$ denote the non-negative integers and let $\mathbb{N}_+$ denote the positive integers. Let $\mathbf{k} = (k_1, \ldots, k_n) \in \mathbb{N}_+^n$ be a fixed vector, and let $\mathbf{0} = (0, \ldots, 0)$, $\mathbf{1} = (1, \ldots, 1)$, $\mathbf{k} + \mathbf{1} = (k_1 + 1, \ldots, k_n + 1)$. In this paper we study the topology of $\vec{P}(X)_{\mathbf{0}}^{\mathbf{k+1}}$ for spaces X that are associated with the special PV programs described above. Let $\mathcal{F}$ be a family of subsets of $[n] = \{1, \ldots, n\}$ and let

$$Y_{\mathcal{F}} = \bigcup_{F \in \mathcal{F}} \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_i \in \mathbb{Z} \text{ for all } i \in F\}. \tag{1}$$

The *Euclidean Pattern Space* associated with $\mathcal{F}$ is defined by $X_{\mathcal{F}} = \mathbb{R}^n \setminus Y_{\mathcal{F}}$, with a corresponding *Path Space* $\vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$.

Example If $\mathcal{F}$ consists of the single set $[n]$ then $X_{\mathcal{F}} = \mathbb{R}^n \setminus \mathbb{Z}^n$. Raussen and Ziemiański [13] investigated the path space $\vec{P}(\mathbb{R}^n \setminus \mathbb{Z}^n)_{\mathbf{0}}^{\mathbf{k+1}}$ and determined its homology groups and its cohomology ring. Their result concerning homology is the following:

Theorem 1.1 (Raussen and Ziemiański [13]) *For $n \geq 3$*

$$\tilde{H}_\ell(\vec{P}(\mathbb{R}^n \setminus \mathbb{Z}^n)_{\mathbf{0}}^{\mathbf{k+1}}) = \begin{cases} \mathbb{Z}^{\prod_{i=1}^n \binom{k_i}{m}} & \ell = (n-2)m, \ m > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The Betti number $\prod_{i=1}^n \binom{k_i}{m}$ in (2) corresponds to the number of strictly increasing integer sequences of length m strictly between $\mathbf{0}$ and $\mathbf{k} + \mathbf{1}$.

In this paper we consider $\vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$ for general $\mathcal{F}$. Without loss of generality we may assume that $\mathcal{F}$ is *upward closed*, i.e. if $F \in \mathcal{F}$ and $F \subset F' \subset [n]$ then $F' \in \mathcal{F}$. It will also be assumed that $|F| \geq 2$ for all $F \in \mathcal{F}$ (otherwise $\vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$ is empty). We first introduce some terminology.

Definition 1.2

(i) A subset $\mathcal{G} \subset \mathcal{F}$ is a *matching* if $G \cap G' = \emptyset$ for all $G \neq G' \in \mathcal{G}$. Let $M(\mathcal{F})$ denote the family of all nonempty matchings of $\mathcal{F}$, with partial order $\preceq$ given by $\mathcal{G} \preceq \mathcal{G}'$ if for every $G \in \mathcal{G}$ there exists a $G' \in \mathcal{G}'$ such that $G \subset G'$. For $K \subset [n]$ let

$$M(\mathcal{F})_{\preceq K} = \{\mathcal{G} \in M(\mathcal{F}) : G \subset K \text{ for all } G \in \mathcal{G}\}$$

and let $M(\mathcal{F})_{\prec K} = M(\mathcal{F})_{\preceq}(K) \setminus \{\{K\}\}$. The order complex of $M(\mathcal{F})_{\prec K}$ is denoted by $\Delta(M(\mathcal{F})_{\prec K})$.

(ii) For a function $\mathbf{m} : \mathcal{F} \to \mathbb{N}$ let $T_{\mathcal{F}}(\mathbf{m})$ be the (simple undirected) graph on the vertex set $\cup_{F \in \mathcal{F}}\{F\} \times [\mathbf{m}(F)]$, where two vertices $(F, i) \neq (F', i')$ are connected by an edge if $F \cap F' \neq \emptyset$.

(iii) An *orientation* of a simple undirected graph $G = (V, E)$ will be determined by a function $\alpha : E \to V^2$ that maps an edge $\{u, v\} \in E$ to either (u, v) or (v, u). An orientation is *acyclic* if the resulting directed graph does not contain directed cycles. Let $\mathfrak{A}(G)$ denote the set of acyclic orientations of G and let $a(G) = |\mathfrak{A}(G)|$. By a result of Stanley [14], $a(G)$ can be computed by evaluating the chromatic polynomial of G at -1.

For $\mathbf{m} : \mathcal{F} \to \mathbb{N}$ let

$$\begin{aligned} b_{\mathcal{F},\mathbf{k}}(\mathbf{m}) &= \frac{a(T_{\mathcal{F}}(\mathbf{m}))}{\prod_{F \in \mathcal{F}} \mathbf{m}(F)!} \prod_{i=1}^{n} \binom{k_i}{\sum_{F \ni i} \mathbf{m}(F)}, \\ c_{\mathcal{F}}(\mathbf{m}) &= \sum_{F \in \mathcal{F}} \mathbf{m}(F)(|F| - 2) + 1. \end{aligned} \tag{3}$$

The *reduced Poincaré series* of a space Y over a field $\mathbb{K}$ is defined by

$$f_{\mathbb{K}}(Y, t) = \sum_{i \geq 0} \dim \tilde{H}_{i-1}(Y; \mathbb{K}) t^i. \tag{4}$$

Our main result is the following

Theorem 1.3

(i) If $H_*(\Delta(M(\mathcal{F})_{\prec F}); \mathbb{Z})$ is free for all $F \in \mathcal{F}$ then $H_*(\vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}; \mathbb{Z})$ is free.

(ii) For any field $\mathbb{K}$

$$f_{\mathbb{K}}\left(\vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}, t\right) = \sum_{0 \neq \mathbf{m} \in \mathbb{N}^{\mathcal{F}}} b_{\mathcal{F},\mathbf{k}}(\mathbf{m}) t^{c_{\mathcal{F}}(\mathbf{m})} \prod_{F \in \mathcal{F}} f_{\mathbb{K}}\left(\Delta(M(\mathcal{F})_{\prec F}), t^{-1}\right)^{\mathbf{m}(F)}. \tag{5}$$

The paper is organized as follows: In Sect. 2 we describe a subspace arrangement $D_{\mathcal{F}}$ that is homotopy equivalent to $\vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$. In Sect. 3 we state Theorem 3.1 that describes the homotopy type of the Alexander dual of $D_{\mathcal{F}}$ and then use it to prove Theorem 1.3. The proof of Theorem 3.1 is given in Sect. 4 which constitutes the main technical part of the paper. In Sect. 5 we discuss several applications arising from particular cases of Theorem 1.3. The easy conclusions about (higher) connectivity of path spaces in Sect. 5.4 are probably the most notable ones for applications in concurrency theory. Some open problems are mentioned in Sect. 6.

2 Directed Paths via Subspace Arrangements

Spaces of directed paths in a PV-model have been shown to be homotopy equivalent to certain finite prod-simplicial complexes that make homology computations possible – at least in principle [6, 11]. Unfortunately, these complexes grow very fast in dimension and size. Here we give an alternative description as *complement of a subspace arrangement*. Remark that directedness has the consequence that such an arrangement has to be considered as a subset of a product of simplices and *not* of Euclidean space; this is the reason why classical results are not immediately applicable.

Let $\mathcal{F}$ be an upward closed hypergraph on $[n]$ and let $\mathbf{k} = (k_1, \dots, k_n) \in \mathbb{N}_+^n$. In this section we describe a model for $\vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$ up to homotopy equivalence.

Definition 2.1

(i) For $k \geq 1$ let $\overset{\circ}{\Delta}_k$ denote the open k-simplex

$$\overset{\circ}{\Delta}_k = \{(x_1, \dots, x_k) \in \mathbb{R}^k : 0 < x_1 < \cdots < x_k < 1\}.$$

For $k = 0$ let $\overset{\circ}{\Delta}_0$ denote the one point space $\{*\}$.
For $\mathbf{k} = (k_1, \dots, k_n) \in \mathbb{N}_+^n$ let

$$N = \sum_{i=1}^{n} k_i, \ \ [\mathbf{k}] = \prod_{i=1}^{n} [k_i], \text{ and } \overset{\circ}{\Delta}_{\mathbf{k}} = \prod_{i=1}^{n} \overset{\circ}{\Delta}_{k_i} \subset \mathbb{R}^N. \tag{6}$$

(ii) For $F \subset [n]$ let $[\mathbf{k}_F] = \prod_{i \in F} [k_i]$. For $\mathbf{j} = (\mathbf{j}(i))_{i \in F} \in [\mathbf{k}_F]$ and $F' \subset F$, the *restriction* $(\mathbf{j})_{|F'} \in [\mathbf{k}_{F'}]$ of $\mathbf{j}$ to F' is given by $(\mathbf{j})_{|F'}(i) = \mathbf{j}(i)$ for all $i \in F'$. A *partial sequence* is a pair $(F, \mathbf{j})$ where $F \subset [n]$ and $\mathbf{j} = (\mathbf{j}(i))_{i \in F} \in [\mathbf{k}_F]$. Let $S_{\mathcal{F}}$ be the family of all partial sequences $(F, \mathbf{j})$ where $F \in \mathcal{F}$ and $\mathbf{j} \in [\mathbf{k}_F]$.

(iii) For a partial sequence $(F, \mathbf{j})$ let

$$G_{(F,\mathbf{j})} = \{(x_{i1}, \dots, x_{ik_i})_{i=1}^n \in \prod_{i=1}^{n} \overset{\circ}{\Delta}_{k_i} : x_{i\mathbf{j}(i)} = x_{i'\mathbf{j}(i')} \text{ for all } i, i' \in F\}.$$

Let

$$E_{\mathcal{F}} = \bigcup_{(F,\mathbf{j})\in S_{\mathcal{F}}} G_{(F,\mathbf{j})} \quad , \quad D_{\mathcal{F}} = \mathring{\Delta}_{\mathbf{k}} - E_{\mathcal{F}}.$$

(iv) The one-point compactification of $\mathring{\Delta}_{\mathbf{k}}$ is given by

$$\widehat{\mathring{\Delta}}_{\mathbf{k}} = \mathring{\Delta}_{\mathbf{k}} \cup \{\infty\} = \Delta_{\mathbf{k}}/_{\partial\Delta_{\mathbf{k}}} \cong S^N.$$

For $(F,\mathbf{j}) \in S_{\mathcal{F}}$, the compactification of $G_{(F,\mathbf{j})}$ in $\widehat{\mathring{\Delta}}_{\mathbf{k}}$ is given by $\Gamma_{(F,\mathbf{j})} = G_{(F,\mathbf{j})} \cup \{\infty\}$. The compactification of $E_{\mathcal{F}}$ in $\widehat{\mathring{\Delta}}_{\mathbf{k}}$ is

$$\widehat{E_{\mathcal{F}}} = E_{\mathcal{F}} \cup \{\infty\}.$$

Let $\vec{P}_<(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}} \subset \vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$ denote the space of *increasing* directed paths $\mathbf{p} = (p_1,\dots,p_n) : I \to X_{\mathcal{F}} \subset \mathbf{R}^n$ characterized by $t < t' \Rightarrow p_i(t) < p_i(t')$ (instead of $\leq$) for all i. Remark that every component p_i is a homeomorphism of the unit interval.

A correspondence between the space $D_{\mathcal{F}}$ from Definition 2.1(iii) and this path space $\vec{P}_<(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$ and may be established as follows: For every $k \in \mathbb{N}_+$ and $\mathbf{x} = (x_1,\dots,x_k) \in \mathring{\Delta}_k$ let $p_{\mathbf{x}} : I \to [0,k+1]$ denote the (directed) path with $p_{\mathbf{x}}(0) = 0$, $p_{\mathbf{x}}(1) = k+1$, $p_{\mathbf{x}}(x_i) = i$, $1 \leq i \leq k$, and connected by line segments inbetween. For every $\mathbf{0} < \mathbf{k} \in \mathbb{N}_+^n$ and every $\mathbf{x} = (\mathbf{x}_1,\dots\mathbf{x}_n) \in \mathring{\Delta}_{\mathbf{k}}$ (cf. 6), let $\mathbf{p}(\mathbf{x})(t) = (p_{\mathbf{x}_1}(t),\dots,p_{\mathbf{x}_n}(t))$. This recipe defines a continuous map $P : \mathring{\Delta}_{\mathbf{k}} \to \vec{P}_<(\mathbf{R}^n)_{\mathbf{0}}^{\mathbf{k+1}}$ that restricts to a map $P_{\mathcal{F}}^< : D_{\mathcal{F}} \to \vec{P}_<(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$: For $\mathbf{x} = (\mathbf{x}_1,\dots,\mathbf{x}_n) \in \mathring{\Delta}_{\mathbf{k}}$ and $\mathbf{x}_i = (x_{i1},\dots,x_{ik_i}) \in \mathring{\Delta}_{k_i}$, $(F,\mathbf{j}) \in S_{\mathcal{F}}$ and $0 < t < 1$ assume that $p_{\mathbf{x}_i}(t) = \mathbf{j}(i) \in \mathbf{Z}, i \in F$. Then $t = x_{i\mathbf{j}(i)} = x_{i'\mathbf{j}(i')}$ for $i, i' \in F$ and hence $\mathbf{x} \in E_{\mathcal{F}}$.

The composition of $P_{\mathcal{F}}^<$ with the inclusion map $i : \vec{P}_<(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}} \hookrightarrow \vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$ will be denoted by $\vec{P}_{\mathcal{F}} : D_{\mathcal{F}} \to \vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$.

Proposition 2.2 *The map $\vec{P}_{\mathcal{F}} : D_{\mathcal{F}} \to \vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$ is a homotopy equivalence.*

We prove Proposition 2.2 via the following two lemmas:

Lemma 2.3 *The map $P_{\mathcal{F}}^< : D_{\mathcal{F}} \to \vec{P}_<(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$ is a homotopy equivalence.*

Proof Define a reverse continuous map $Q : \vec{P}_<(\mathbf{R}^n)_{\mathbf{0}}^{\mathbf{k+1}} \to \mathring{\Delta}_{\mathbf{k}}$ as follows: For $\mathbf{p} = (p_1,\dots,p_n) \in \vec{P}_<(\mathbf{R}^n)_{\mathbf{0}}^{\mathbf{k+1}}$ such that $p_j(x_{ij_i}) = j_i$ let $Q(\mathbf{p}) = (x_{11},\dots,x_{1k_1};\dots;x_{n1},\dots,x_{nk_n})$. Remark that Q is well-defined and continuous since every p_i is a homeomorphism; and that Q cannot be extended to the space $\vec{P}(\mathbf{R}^n)_{\mathbf{0}}^{\mathbf{k+1}}$ of non-decreasing directed paths. Remark moreover that Q restricts to a map $Q_{\mathcal{F}} : \vec{P}_<(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}} \to D_{\mathcal{F}}$.

It is obvious from the definitions that $Q \circ P$ is the identity map on $\overset{\circ}{\Delta}_{\mathbf{k}}$ and hence that $Q_{\mathcal{F}} \circ P_{\mathcal{F}}$ is the identity map on $D_{\mathcal{F}}$. The map $P \circ Q : \vec{P}_<(\mathbf{R}^n)_{\mathbf{0}}^{\mathbf{k+1}} \to \vec{P}_<(\mathbf{R}^n)_{\mathbf{0}}^{\mathbf{k+1}}$ has the property: $((P \circ Q)(\mathbf{p}))_i(p_i^{-1}(j)) = j = p_i(p_i^{-1}(j))$ and $((P \circ Q)(\mathbf{p}))_i(t) \notin \mathbf{Z}$ for $\mathbf{p} \in \vec{P}_<(\mathbf{R}^n)_{\mathbf{0}}^{\mathbf{k+1}}$ and $t \notin Q_i(\mathbf{p}) \cup \{0, 1\}$. That same property holds for all directed paths in the linear homotopy on $\vec{P}_<(\mathbf{R}^n)_{\mathbf{0}}^{\mathbf{k+1}}$ given by $s \mapsto (1-s)\mathbf{p}+s(P \circ Q)(\mathbf{p}), 0 \leq s \leq 1$. Hence $P \circ Q$ restricts to a map $P_{\mathcal{F}} \circ Q_{\mathcal{F}} : \vec{P}_<(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}} \to \vec{P}_<(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$ that is homotopic to the identity map on $\vec{P}_<(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$. □

Lemma 2.4 *The inclusion map* $i : \vec{P}_<(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}} \hookrightarrow \vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$ *is a homotopy equivalence for every positive integer vector* $\mathbf{k}$.

Proof Let $\delta_{\mathbf{k}} \in \vec{P}_<(\mathbf{R}^n)_{\mathbf{0}}^{\mathbf{k+1}}$ denote the linear path given by $\delta_{\mathbf{k}}(t) = t(\mathbf{k} + \mathbf{1})$. Then, for every $\mathbf{p} \in \vec{P}(\mathbf{R}^n)_{\mathbf{0}}^{\mathbf{k+1}}$ and $0 < s \leq 1$, the convex combination $\mathbf{p}_s := (1 - s)\mathbf{p} + s\delta_{\mathbf{k}}$ is *strictly* increasing and hence contained in $\vec{P}_<(\mathbf{R}^n)_{\mathbf{0}}^{\mathbf{k+1}}$. For a given $\mathbf{p} \in \vec{P}_<(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$, we want to choose $s > 0$ small enough to ensure that $\mathbf{p}_s$ avoids $Y_{\mathcal{F}}$ (see (1)) and hence so that $\mathbf{p}_s$ is contained in $\vec{P}_<(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$; and this in a way that makes the parameter s depend continuously on the path $\mathbf{p}$.

Fix a norm and the associated metric d on $\mathbf{R}^n$, e.g., the box norm. For every path $\mathbf{p} \in \vec{P}_<(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$, the spaces $\mathbf{p}(I)$ and $Y_{\mathcal{F}} \cap [\mathbf{0}, \mathbf{k} + \mathbf{1}]$ are disjoint closed and hence compact subspaces of $[\mathbf{0}, \mathbf{k}+\mathbf{1}]$ with a positive distance $d(\mathbf{p}) := \max_{t \in I}(d(\mathbf{p}(t), Y_{\mathcal{F}})$ depending continuously on $\mathbf{p}$. Let $K := \max_1^n k_i$ and let $s(\mathbf{p}) = \frac{d(\mathbf{p})}{K}$. Then, for every $\mathbf{p} \in \vec{P}_<(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$ one obtains: $d(\mathbf{p}, \mathbf{p}_s) = s(\mathbf{p}) \parallel \mathbf{p} - \delta_{\mathbf{k}} \parallel < d(\mathbf{p})$ and in particular $d(\mathbf{p}_s, Y_{\mathcal{F}}) > 0$.

Let $i : \vec{P}_<(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}} \to \vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$ denote the inclusion map, and let $r : \vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}} \to \vec{P}_<(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$ denote the continuous map given by $r(\mathbf{p}) = (1 - s(\mathbf{p}))\mathbf{p} + s(\mathbf{p})\delta_{\mathbf{k}}$. The continuous map $R : \vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}} \times I \to \vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$ given by $R(\mathbf{p}, t) = (1 - ts(\mathbf{p}))\mathbf{p} + ts(\mathbf{p})\delta_{\mathbf{k}}$ is a homotopy between the identity and $i \circ r$; its restriction to $\vec{P}_<(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$ is a homotopy between the identity and $r \circ i$. □

Remark A variant of the proof above shows that spaces of increasing and of non-decreasing directed paths (as they arise in models for concurrency theory) are homotopy equivalent in a more general context.

3 The Homology of $D_{\mathcal{F}}$

In this section we state Theorem 3.1 that describes the homotopy type of the Alexander dual of $D_{\mathcal{F}}$ in the one-point compactification of $\overset{\circ}{\Delta}_{\mathbf{k}}$ – a sphere of dimension $N = \sum_{i=1}^n k_i$. This result is then used to prove Theorem 1.3. Our main observation is the following homotopy decomposition of $\widehat{E_{\mathcal{F}}}$ (see (3) and Definitions 1.2(i) and 2.1(iii)).

Theorem 3.1

$$\widehat{E_{\mathcal{F}}} \simeq \bigvee_{0\neq \mathbf{m}\in\mathbb{N}^{\mathcal{F}}} \bigvee^{b_{\mathcal{F},\mathbf{k}}(\mathbf{m})} S^{N-c_{\mathcal{F}}(\mathbf{m})} * \mathop{\ast}_{F\in\mathcal{F}} \Delta(M(\mathcal{F})_{\prec F})^{*\mathbf{m}(F)}. \tag{7}$$

The proof of Theorem 3.1 is deferred to Sect. 4.

Proof of Theorem 1.3 (i) If the integral homology $\tilde{H}_*\left(\Delta(M(\mathcal{F})_{\prec F});\mathbb{Z}\right)$ is free for all $F\in\mathcal{F}$, then (7) implies that $\tilde{H}_*(\widehat{E_{\mathcal{F}}};\mathbb{Z})$ is free. Recalling that $\widehat{\mathring{\Delta}}_{\mathbf{k}} \cong S^N$, it follows by Alexander duality that for all ℓ

$$\begin{aligned}\tilde{H}_\ell(D_{\mathcal{F}};\mathbb{Z}) &= \tilde{H}_\ell(\mathring{\Delta}_{\mathbf{k}} - E_{\mathcal{F}};\mathbb{Z})\\ &= \tilde{H}_\ell(\widehat{\mathring{\Delta}}_{\mathbf{k}} - \widehat{E_{\mathcal{F}}};\mathbb{Z}) \cong \tilde{H}_{N-\ell-1}(\widehat{E_{\mathcal{F}}};\mathbb{Z}).\end{aligned}$$

Therefore $\tilde{H}_\ell(D_{\mathcal{F}};\mathbb{Z})$ is free.
(ii) Recall that the behavior of the reduced Poincaré series $f_{\mathbb{K}}(\cdot)$ (cf. (4)); as a consequence, with respect to the wedge and join operations is given by

$$\begin{aligned} f_{\mathbb{K}}(Y_1\vee Y_2,t) &= f_{\mathbb{K}}(Y_1,t)+f_{\mathbb{K}}(Y_2,t)\ , \\ f_{\mathbb{K}}(Y_1 * Y_2,t) &= f_{\mathbb{K}}(Y_1,t)f_{\mathbb{K}}(Y_2,t). \end{aligned} \tag{8}$$

Furthermore, if Y is a subcomplex of S^N then by Alexander duality

$$f_{\mathbb{K}}(S^N - Y,t) = t^{N+1} f_{\mathbb{K}}(Y,t^{-1}). \tag{9}$$

Theorem 3.1 together with (8) imply that for any field $\mathbb{K}$

$$f_{\mathbb{K}}(\widehat{E_{\mathcal{F}}},t) = \sum_{0\neq\mathbf{m}\in\mathbb{N}^{\mathcal{F}}} b_{\mathcal{F},\mathbf{k}}(\mathbf{m}) t^{N-c_{\mathcal{F}}(\mathbf{m})+1} \prod_{F\in\mathcal{F}} f_{\mathbb{K}}(\Delta(M(\mathcal{F})_{\prec F}),t)^{\mathbf{m}(F)}. \tag{10}$$

Combining Proposition 2.2 with (9) and (10) it follows that

$$\begin{aligned} f_{\mathbb{K}}\left(\vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}},t\right) &= f_{\mathbb{K}}\left(D_{\mathcal{F}},t\right) = t^{N+1} f_{\mathbb{K}}\left(\widehat{E_{\mathcal{F}}},t^{-1}\right)\\ &= \sum_{0\neq\mathbf{m}\in\mathbb{N}^{\mathcal{F}}} b_{\mathcal{F},\mathbf{k}}(\mathbf{m}) t^{c_{\mathcal{F}}(\mathbf{m})} \prod_{F\in\mathcal{F}} f_{\mathbb{K}}\left(\Delta(M(\mathcal{F})_{\prec F}),t^{-1}\right)^{\mathbf{m}(F)}. \end{aligned}$$

□

4 Homotopy Decomposition of $\widehat{E_{\mathcal{F}}}$

In this Section we prove Theorem 3.1. Our basic approach is to apply the Wedge Lemma of Ziegler and Živaljević [17] to the cover $\{\Gamma_{(F,\mathbf{j})} : (F,\mathbf{j}) \in S_{\mathcal{F}}\}$ of $\widehat{E_{\mathcal{F}}}$. The actual proof depends on a number of preliminary results. For notations, we refer the reader to Definition 2.1.

Definition 4.1 Let $R \subset S_{\mathcal{F}}$.

(i) Let

$$G_R = \bigcap_{(F,\mathbf{j})\in R} G_{(F,\mathbf{j})} \quad , \quad \Gamma_R = \bigcap_{(F,\mathbf{j})\in R} \Gamma_{(F,\mathbf{j})} = G_R \cup \{\infty\}.$$

(ii) R is *separated* if $\mathbf{j}(i) \neq \mathbf{j}'(i)$ for any $(F,\mathbf{j}) \neq (F',\mathbf{j}') \in R$ and $i \in F \cap F'$.

For separated families $R \subset S_{\mathcal{F}}$ it is sometimes useful to represent G_R by a diagram with n rows such that the i-th row contains the coordinates $x_{i1} < \cdots < x_{ik_i}$ of $\overset{\circ}{\Delta}_{k_i}$, and such that $x_{ij}, x_{i'j'}$ are connected by a dashed line iff $x_{ij} = x_{i'j'}$ for all $\mathbf{x} \in G_R$, i.e. iff there exists an $(F,\mathbf{j}) \in R$ such that $i, i' \in F$ and $\mathbf{j}(i) = j$ and $\mathbf{j}(i') = j'$.

Example 4.2 Let $k_1 = k_2 = k_3 = 2$ and let $R = \{(F_i,\mathbf{j}_i)\}_{i=1}^{3}$ where $F_1 = \{1,2\}$, $F_2 = \{2,3\}$, $F_3 = \{1,3\}$ and $(\mathbf{j}_1(1),\mathbf{j}_1(2)) = (1,1)$, $(\mathbf{j}_2(2),\mathbf{j}_2(3)) = (2,1)$, $(\mathbf{j}_3(1),\mathbf{j}_3(3)) = (2,2)$. The diagram of G_R is depicted in Fig. 2.

Definition 4.3 For $R \subset S_{\mathcal{F}}$ let K_R be the directed graph on the vertex set R with edges $(F,\mathbf{j}) \to (F',\mathbf{j}')$, where $(F,\mathbf{j})$ and $(F',\mathbf{j}')$ are distinct elements of R that satisfy $F \cap F' \neq \emptyset$ and $\mathbf{j}(i) < \mathbf{j}'(i)$ for all $i \in F \cap F'$. The family $R \subset S_{\mathcal{F}}$ is *acyclic* if R is separated and if K_R does not contain directed cycles. Let $A_{\mathcal{F}}$ denote the set of all acyclic subfamilies of $S_{\mathcal{F}}$.
The next two Propositions describe some properties of Γ_R for separated families R.

Proposition 4.4 *Let $R \subset S_{\mathcal{F}}$ be a separated family. Then:*

(i) If $R \notin A_{\mathcal{F}}$ then $\Gamma_R = \{\infty\}$.
(ii) If $R \in A_{\mathcal{F}}$ then there is a homeomorphism

$$\Gamma_R \cong S^{N-\sum_{(F,\mathbf{j})\in R}(|F|-1)}. \tag{11}$$

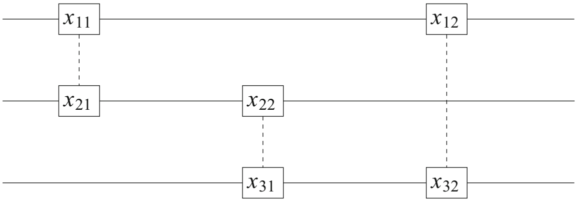


Fig. 2 Diagram of G_R (cf. Example 4.2)

Proof (i) Let

$$(F_1, \mathbf{j}_1) \to \cdots \to (F_r, \mathbf{j}_r) \to (F_1, \mathbf{j}_1)$$

be a directed cycle in K_R. Then there exist

$$i_1 \in F_1 \cap F_2, i_2 \in F_2 \cap F_3, \ldots, i_r \in F_r \cap F_1$$

such that

$$\mathbf{j}_1(i_1) < \mathbf{j}_2(i_1)\ ,\ \mathbf{j}_2(i_2) < \mathbf{j}_3(i_2)\ ,\ \cdots\ ,\ \mathbf{j}_r(i_r) < \mathbf{j}_1(i_r). \tag{12}$$

We will show that $G_R = \emptyset$ and hence $\Gamma_R = \{\infty\}$. Indeed, suppose that $((x_{i,1}, \ldots, x_{i,k_i}))_{i=1}^n$ is contained in G_R. We conclude from (12) – since $i_j, i_{j+1} \in F_{j+1}, j < r$, and $i_1, i_r \in F_1$:

$$x_{i_1,\mathbf{j}_1(i_1)} < x_{i_1,\mathbf{j}_2(i_1)} = x_{i_2,\mathbf{j}_2(i_2)} < x_{i_2,\mathbf{j}_3(i_2)} = x_{i_3,\mathbf{j}_3(i_3)} <$$
$$\cdots < x_{i_{r-1},\mathbf{j}_r(i_{r-1})} = x_{i_r,\mathbf{j}_r(i_r)} < x_{i_r,\mathbf{j}_1(i_r)} = x_{i_1,\mathbf{j}_1(i_1)},$$

a contradiction.

Example 4.5 Let $k_1 = k_2 = k_3 = 2$ and let $R = \{(F_i, \mathbf{j}_i)\}_{i=1}^3$ where $F_1 = \{1, 2\}$, $F_2 = \{2, 3\}$, $F_3 = \{1, 3\}$ and $(\mathbf{j}_1(1), \mathbf{j}_1(2)) = (2, 1)$, $(\mathbf{j}_2(2), \mathbf{j}_2(3)) = (2, 1)$, $(\mathbf{j}_3(1), \mathbf{j}_3(3)) = (1, 2)$. Then $(F_1, \mathbf{j}_1) \to (F_2, \mathbf{j}_2) \to (F_3, \mathbf{j}_3) \to (F_1, \mathbf{j}_1)$ is a cycle in K_R and thus $G_R = \emptyset$, as is also clear from the diagram of G_R in Fig. 3.

(ii) For $(F, \mathbf{j}) \in S_{\mathcal{F}}$ define

$$V_{(F,\mathbf{j})} = \{(x_{i1}, \ldots, x_{ik_i})_{i=1}^n \in \prod_{i=1}^n \mathbb{R}^{k_i} : x_{i\mathbf{j}(i)} = x_{i'\mathbf{j}(i')} \text{ for all } i, i' \in F\}.$$

For $R \subset S_{\mathcal{F}}$, let $V_R = \bigcap_{(F,\mathbf{j})\in R} V_{(F,\mathbf{j})} \subset \mathbb{R}^N$. If the family R is separated, then V_R is a linear subspace of codimension $\sum_{(F,\mathbf{j})\in R}(|F| - 1)$. If R is acyclic, one can easily find an element $\mathbf{x} \in V_R \cap \mathring{\Delta}_{\mathbf{k}}$.

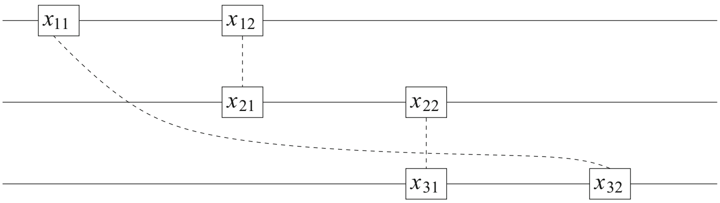


Fig. 3 If K_R has directed cycles then $G_R = \emptyset$ (cf. Example 4.5)

For such a chosen solution $\mathbf{x}$ and for $\mathbf{v} \in S(V_R) = \{\mathbf{u} \in V_R : \|\mathbf{u}\| = 1\} \subset D(V_R) = \{\mathbf{u} \in V_R : \|\mathbf{u}\| \leq 1\}$ let $\alpha(\mathbf{v}) = \min\{t > 0 | \, \mathbf{x} + t\mathbf{v} \in \partial\Delta_{\mathbf{k}}\}$; hence, for $\mathbf{w} \in \partial\Delta_{\mathbf{k}}$, one has that $\alpha(\frac{\mathbf{w}-\mathbf{x}}{\|\mathbf{w}-\mathbf{x}\|}) = \|\mathbf{w}-\mathbf{x}\|$. This recipe defines a continuous map α from $S(V_R)$ to the positive reals since $\alpha(\mathbf{v})$ is locally obtained as the minimum among the solutions to a number of linear equations.

We define a (scaling) map

$$\Phi_R : \Gamma_R = V_R \cap \Delta_{\mathbf{k}}/_{V_R \cap \partial\Delta_{\mathbf{k}}} \to D(V_R)/_{S(V_R)} \cong S^{N-\sum_{(F,\mathbf{j})\in R}(|F|-1)}$$

by

$$\Phi_R(\mathbf{w}) = \begin{cases} \frac{1}{\alpha(\frac{\mathbf{w}-\mathbf{x}}{\|\mathbf{w}-\mathbf{x}\|})}(\mathbf{w}-\mathbf{x}) & \mathbf{w} \neq \mathbf{x}, \\ \mathbf{0} & \mathbf{w} = \mathbf{x}. \end{cases}$$

The map Φ_R is indeed a *homeomorphism* with inverse $\Psi_R : D(V_R)/_{S(V_R)} \to \Gamma_R$ given by

$$\Psi_R(\mathbf{v}) = \begin{cases} \mathbf{x} + \alpha(\frac{\mathbf{v}}{\|\mathbf{v}\|})\mathbf{v} & \mathbf{v} \neq \mathbf{0}, \\ \mathbf{x} & \mathbf{v} = \mathbf{0}. \end{cases}$$

□

Proposition 4.6 *Let $R, R' \in A_{\mathcal{F}}$. Then the following two conditions are equivalent:*

(a) $\Gamma_R \subset \Gamma_{R'}$.
(b) For any $(F', \mathbf{j}') \in R'$ there exists an $(F, \mathbf{j}) \in R$ such that $F' \subset F$ and $\mathbf{j}' = (\mathbf{j})_{|F'}$.

Proof Clearly (b) implies (a). To show the other direction, assume that $\Gamma_R \subset \Gamma_{R'}$ and let $(F', \mathbf{j}') \in R'$. Let $R_1 = R \cup \{(F', \mathbf{j}')\}$, then

$$\Gamma_R = \Gamma_R \cap \Gamma_{R'} \subset \Gamma_{R_1} \subset \Gamma_R,$$

hence $\Gamma_{R_1} = \Gamma_R \neq \{\infty\}$. It follows that if R_1 is separated then it must be acyclic. But this would imply, using Proposition 4.4(ii), that

$$\dim \Gamma_{R_1} = \dim \Gamma_R - (|F'| - 1) < \dim \Gamma_R,$$

in contradiction with $\Gamma_{R_1} = \Gamma_R$. Hence R_1 is not separated and therefore

$$S = \{(F, \mathbf{j}) \in R : F \cap F' \neq \emptyset \ \ \& \ \ (\mathbf{j})_{|F\cap F'} = (\mathbf{j}')_{|F\cap F'}\} \neq \emptyset.$$

We claim that $|S| = 1$. Otherwise there exist $(F_1, \mathbf{j}_1) \neq (F_2, \mathbf{j}_2) \in R$ and $i_1 \in F_1 \cap F'$, $i_2 \in F_2 \cap F'$ such that $\mathbf{j}_1(i_1) = \mathbf{j}'(i_1)$ and $\mathbf{j}_2(i_2) = \mathbf{j}'(i_2)$. It follows that if

$\mathbf{x} = (x_{i1}, \ldots, x_{ik_i})_{i=1}^n \in G_{R_1}$ then for all $i_1' \in F_1$, $i_2' \in F_2$

$$\begin{aligned} x_{i_1'\mathbf{j}_1(i_1')} &= x_{i_1\mathbf{j}_1(i_1)} = x_{i_1\mathbf{j}'(i_1)} \\ &= x_{i_2\mathbf{j}'(i_2)} = x_{i_2\mathbf{j}_2(i_2)} = x_{i_2'\mathbf{j}_2(i_2')}. \end{aligned} \tag{13}$$

Since R is separated, (13) implies that $F_1 \cap F_2 = \emptyset$. Let $F_3 = F_1 \cup F_2$ and let $\mathbf{j}_3 \in [\mathbf{k}_{F_3}]$ be given by

$$\mathbf{j}_3(i) = \begin{cases} \mathbf{j}_1(i) & i \in F_1, \\ \mathbf{j}_2(i) & i \in F_2. \end{cases}$$

Writing

$$R_2 = R \setminus \{(F_1, \mathbf{j}_1), (F_2, \mathbf{j}_2)\} \cup \{(F_3, \mathbf{j}_3)\},$$

it follows from (13) that $\Gamma_R = \Gamma_{R_1} \subset \Gamma_{R_2} \subset \Gamma_R$. Therefore $\Gamma_{R_2} = \Gamma_R \neq \{\infty\}$. As R_2 is separated, it follows that $R_2 \in A_{\mathcal{F}}$ and hence by Proposition 4.4(ii):

$$\dim \Gamma_{R_2} = \dim \Gamma_R + (|F_1| - 1) + (|F_2| - 1) - (|F_1 \cup F_2| - 1) = \dim \Gamma_R - 1,$$

in contradiction with $\Gamma_{R_2} = \Gamma_R$. Therefore $|S| = 1$.

Write $S = \{(F_1, \mathbf{j}_1)\}$ and let $i_1 \in F_1 \cap F'$. Then $\mathbf{j}_1(i_1) = \mathbf{j}'(i_1)$. It follows that if $\mathbf{x} = (x_{i1}, \ldots, x_{ik_i})_{i=1}^n \in G_{R_1}$ then for all $i_1' \in F_1$, $i' \in F'$

$$x_{i_1'\mathbf{j}_1(i_1')} = x_{i_1\mathbf{j}_1(i_1)} = x_{i_1\mathbf{j}'(i_1)} = x_{i'\mathbf{j}'(i')}. \tag{14}$$

Let $F_4 = F_1 \cup F'$ and let $\mathbf{j}_4 \in [\mathbf{k}_{F_4}]$ be given by

$$\mathbf{j}_4(i) = \begin{cases} \mathbf{j}_1(i) & i \in F_1, \\ \mathbf{j}'(i) & i \in F'. \end{cases}$$

Note that $\mathbf{j}_4$ is well defined by (14). Writing

$$R_3 = R \setminus \{(F_1, \mathbf{j}_1)\} \cup \{(F_4, \mathbf{j}_4)\},$$

it follows from (14) that

$$\Gamma_R = \Gamma_{R_1} \subset \Gamma_{R_3} \subset \Gamma_R,$$

hence $\Gamma_{R_3} = \Gamma_R \neq \{\infty\}$. Furthermore, $|S| = 1$ implies that R_3 is separated. Therefore by Proposition 4.4(ii):

$$\dim \Gamma_{R_3} = \dim \Gamma_R + (|F_1| - 1) - (|F_4| - 1).$$

It follows that $|F_1| = |F_4| = |F_1 \cup F'|$. Hence $F' \subset F_1$ and $\mathbf{j}' = (\mathbf{j}_1)_{|F'}$. □

Let Q be the intersection poset of the cover $\{\Gamma_{(F,\mathbf{j})} : (F,\mathbf{j}) \in S_{\mathcal{F}}\}$ of $\widehat{E_{\mathcal{F}}}$ ordered by reverse inclusion: An element q of Q corresponds to an intersection U_q of sets in the cover, i.e. $U_q = \Gamma_R$ for $R \subset S_{\mathcal{F}}$, and $q' \leq q$ in Q iff $U_q \subset U_{q'}$. Q has a maximal element $\widehat{1}$ that corresponds to $U_{\widehat{1}} = \{\infty\}$. Fix a $\widehat{1} \neq q \in Q$ and let $R \subset S_{\mathcal{F}}$ be a family of minimal cardinality such that $U_q = \Gamma_R$. The assumption that $\mathcal{F}$ is upward closed implies that R is a separated family. Indeed, suppose $u' = (F',\mathbf{j}') \neq u'' = (F'',\mathbf{j}'') \in R$ and there exists some $i_0 \in F' \cap F''$ such that $\mathbf{j}'(i_0) = \mathbf{j}''(i_0)$. Let $\infty \neq \mathbf{x} = (x_{i1},\ldots,x_{ik_i})_{i=1}^n \in \Gamma_R$. Then for any $i \in F' \cap F''$

$$x_{i\mathbf{j}'(i)} = x_{i_0\mathbf{j}'(i_0)} = x_{i_0\mathbf{j}''(i_0)} = x_{i\mathbf{j}''(i)},$$

hence $\mathbf{j}'(i) = \mathbf{j}''(i)$. Let $u = (F,\mathbf{j}) \in S_{\mathcal{F}}$ where $F = F' \cup F''$ and

$$\mathbf{j}(i) = \begin{cases} \mathbf{j}'(i) & i \in F', \\ \mathbf{j}''(i) & i \in F''. \end{cases}$$

Then $\Gamma_R = \Gamma_{R-\{u',u''\}\cup\{u\}}$, contradicting the minimality of R. Thus R is separated. By Proposition 4.4(i), the assumption $q \neq \widehat{1}$ implies that $R \in A_{\mathcal{F}}$; cf. Definition 4.3. We next study the topology of the order complex $\Delta(Q_{<q})$.

Proposition 4.7 *Fix $\widehat{1} \neq q \in Q$ and write $U_q = \Gamma_R$ where $R = \{(F_\ell,\mathbf{j}_\ell)\}_{\ell=1}^r \in A_{\mathcal{F}}$. Then there is a homeomorphism*

$$\Delta(Q_{<q}) \cong \Delta(M(\mathcal{F})_{\prec F_1}) * \cdots * \Delta(M(\mathcal{F})_{\prec F_r}) * S^{r-2}. \tag{15}$$

Proof Let $M(\mathcal{F})^*_{\preceq F_\ell}$ denote the poset obtained by appending to $M(\mathcal{F})_{\preceq F_\ell}$ a minimal element 0_ℓ. Denote

$$C_q = M(\mathcal{F})^*_{\preceq F_1} \times \cdots \times M(\mathcal{F})^*_{\preceq F_r} \setminus \{(0_1,\ldots,0_r)\}.$$

Define a mapping $\gamma : C_q \to A_{\mathcal{F}}$ as follows. Let $\alpha = (\alpha_1,\ldots,\alpha_r) \in C_q$ and let

$$L(\alpha) = \{1 \leq \ell \leq r : \alpha_\ell \neq 0_\ell\}.$$

Note that $L(\alpha) \neq \emptyset$. For $\ell \in L(\alpha)$ write $\alpha_\ell = \mathcal{G}_\ell \in M(\mathcal{F})$ and let

$$\gamma(\alpha) = \bigcup_{\ell \in L(\alpha)} \{(F,(\mathbf{j}_\ell)_{|F}) : F \in \mathcal{G}_\ell\}.$$

Define an order preserving map $\theta : C_q \to Q_{\leq q}$ as follows: For $\alpha \in C_q$ let $\theta(\alpha)$ be the element of Q that satisfies $U_{\theta(\alpha)} = \Gamma_{\gamma(\alpha)}$.

Lemma 4.8 *θ is a poset isomorphism.*

Proof To show surjectivity, let $q' \leq q$ with $U_{q'} = \Gamma_{R'}$ for some $R' \in A_{\mathcal{F}}$. Then $\Gamma_R = U_q \subset U_{q'} = \Gamma_{R'}$ and hence, by Proposition 4.6, there exists an $\alpha \in C_q$ such that $\gamma(\alpha) = R'$. Therefore $U_{\theta(\alpha)} = \Gamma_{\gamma(\alpha)} = \Gamma_{R'} = U_{q'}$ and so $\theta(\alpha) = q'$. To show injectivity, assume that $\theta(\alpha) = \theta(\alpha')$ for some $\alpha, \alpha' \in C_q$. Then $\Gamma_{\gamma(\alpha)} = U_{\theta(\alpha)} = U_{\theta(\alpha')} = \Gamma_{\gamma(\alpha')}$ and therefore $\gamma(\alpha) = \gamma(\alpha')$ by Proposition 4.6. As γ is clearly injective, it follows that $\alpha = \alpha'$. □

Recall the following result of Walker (Theorem 5.1(d) in [16]).

Theorem 4.9 (Walker [16]) *For* $1 \leq i \leq r$ *let* T_i *be a finite poset with minimal element* 0_i *and maximal element* 1_i*. Let* $T = T_1 \times \cdots \times T_r$ *and* $\mathbf{0} = (0_1, \ldots, 0_r)$, $\mathbf{1} = (1_1, \ldots, 1_r) \in T$*. Let* $\widehat{T_i} = T_i - \{(0_i, 1_i)\}$ *and* $\widehat{T} = T - \{\mathbf{0}, \mathbf{1}\}$*. Then there is a homeomorphism*

$$\Delta(\widehat{T}) \cong \Delta(\widehat{T_1}) * \cdots * \Delta(\widehat{T_r}) * S^{r-2}.$$

For $1 \leq i \leq r$ let $T_i = M(\mathcal{F})^*_{\preceq F_i}$. Then $\widehat{T_i} = M(\mathcal{F})_{\prec F_i}$ and $\widehat{T} = C_q - \{(\{F_1\}, \ldots, \{F_r\})\}$. Lemma 4.8 thus implies that $\widehat{T} \cong Q_{<q}$. Therefore by Theorem 4.9:

$$\begin{aligned}\Delta(Q_{<q}) &\cong \Delta(\widehat{T}) \cong \Delta(\widehat{T_1}) * \cdots * \Delta(\widehat{T_r}) * S^{r-2} \\ &= \Delta(M(\mathcal{F})_{\prec F_1}) * \cdots * \Delta(M(\mathcal{F})_{\prec F_r}) * S^{r-2}.\end{aligned}$$

□

Definition 4.10 For a function $0 \neq \mathbf{m} \in \mathbb{N}^{\mathcal{F}}$, let $A_{\mathcal{F}}(\mathbf{m})$ denote the set of all $R \in A_{\mathcal{F}}$ such that $|\{\mathbf{j} \in [\mathbf{k}_F]| \ (F, \mathbf{j}) \in R\}| = \mathbf{m}(F)$ for all $F \in \mathcal{F}$.

The final ingredient needed for the proof of Theorem 3.1 is the following computation (see (3) and Definition 1.2(i)).

Proposition 4.11 *Let* $0 \neq \mathbf{m} \in \mathbb{N}^{\mathcal{F}}$*. Then:*

$$|A_{\mathcal{F}}(\mathbf{m})| = b_{\mathcal{F},\mathbf{k}}(\mathbf{m}) = \frac{a(T_{\mathcal{F}}(\mathbf{m}))}{\prod_{F \in \mathcal{F}} \mathbf{m}(F)!} \prod_{j=1}^{n} \binom{k_j}{\sum_{F \ni j} \mathbf{m}(F)}. \tag{16}$$

Proof Let $\tilde{\mathfrak{A}}(T_{\mathcal{F}}(\mathbf{m}))$ denote the set of all acyclic orientations of $T_{\mathcal{F}}(\mathbf{m})$ such that $(F, i) \to (F, i')$ for all $F \in \mathcal{F}$ and $1 \leq i < i' \leq \mathbf{m}(F)$. Then

$$|\tilde{\mathfrak{A}}(T_{\mathcal{F}}(\mathbf{m}))| = \frac{a(T_{\mathcal{F}}(\mathbf{m}))}{\prod_{F \in \mathcal{F}} \mathbf{m}(F)!}.$$

Define a mapping

$$\tau : A_{\mathcal{F}}(\mathbf{m}) \to \tilde{\mathfrak{A}}(T_{\mathcal{F}}(\mathbf{m})) \times \prod_{i=1}^{n} \binom{[k_i]}{\sum_{F \ni i} \mathbf{m}(F)}$$

as follows. Let $R \in A_{\mathcal{F}}(\mathbf{m})$. For $1 \leq i \leq n$ let

$$B_i = \{\mathbf{j}(i) : (F, \mathbf{j}) \in R \text{ and } i \in F\} \in \binom{[k_i]}{\sum_{F \ni i} \mathbf{m}(F)}.$$

Write

$$R = \bigcup_{\{F \in \mathcal{F} : \mathbf{m}(F) > 0\}} \{(F, \mathbf{j}_{F,\ell}) : 1 \leq \ell \leq \mathbf{m}(F)\}$$

where $\mathbf{j}_{F,\ell} \in \mathbf{k}_F$ for all $1 \leq \ell \leq \mathbf{m}(F)$ and

$$\mathbf{j}_{F,1}(i) < \cdots < \mathbf{j}_{F,\mathbf{m}(F)}(i)$$

for all $i \in F$. Define an orientation $\alpha \in \tilde{\mathfrak{A}}(T_{\mathcal{F}}(\mathbf{m}))$ as follows. Let $e = \{(F, s), (F', s')\}$ be an edge of $T_{\mathcal{F}}(\mathbf{m})$. Define $\alpha(e) = ((F, s), (F', s'))$ if either $F = F'$ and $s < s'$, or if $F \neq F'$ and $\mathbf{j}_{F,s}(i) < \mathbf{j}_{F',s'}(i)$ for some (and therefore all) $i \in F \cap F'$. Now let

$$\tau(R) = (\alpha, B_1, \ldots, B_n).$$

It is straightforward to check that τ is bijective. This proves Proposition 4.11. □

Example 4.12 To illustrate the bijection τ from the proof of Claim 4.11 consider the family $\mathcal{F} = \{F \subset [4] : |F| \geq 2\}$ and let $n = 4$, $(k_1, k_2, k_3, k_4) = (4, 5, 4, 2)$. Let $F_1 = \{1, 2\}$, $F_2 = \{2, 3\}$ and $F_3 = \{1, 3, 4\}$ and for $F \in \mathcal{F}$ let

$$\mathbf{m}(F) = \begin{cases} 2 \; F = F_1 \; \textit{or} \; F = F_2, \\ 1 \; F = F_3, \\ 0 \; \textit{otherwise.} \end{cases}$$

Let $R \in A_{\mathcal{F}}(\mathbf{m})$ satisfy $\tau(R) = (\alpha, B_1, B_2, B_3, B_4)$ where

$$(B_1, B_2, B_3, B_4) = (\{2, 3, 4\}, \{1, 2, 4, 5\}, \{1, 3, 4\}, \{2\})$$

and the orientation α on the (complete) graph $T_{\mathcal{F}}(\mathbf{m})$ is given by the total order

$$(F_2, 1) \rightarrow (F_1, 1) \rightarrow (F_1, 2) \rightarrow (F_3, 1) \rightarrow (F_2, 2).$$

The reconstruction of R from $\tau(R)$ is depicted in Fig. 4.

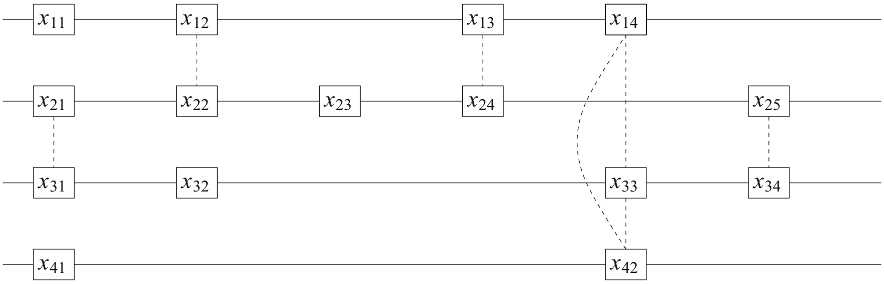


Fig. 4 Reconstruction of R from $\tau(R)$ (cf. Example 4.12)

Proof of Theorem 3.1 Consider the cover $\{\Gamma_{(F,\mathbf{j})} : (F,\mathbf{j}) \in S_{\mathcal{F}}\}$ of $\widehat{E_{\mathcal{F}}}$, and its associated intersection poset Q as above. By Proposition 4.4(ii), if $\widehat{1} \neq q \in Q$ then U_q is a sphere pointed at ∞. Furthermore, if $q < p \in Q$ then the injection $U_p \to U_q$ is a pointed embedding of a sphere (or the point ∞ if $p = \widehat{1}$) into a higher dimensional sphere and is thus homotopic to the constant map $U_p \to \infty \in U_q$. Therefore by the Wedge Lemma (Lemma 1.8 in [17]) there is a homotopy equivalence

$$\widehat{E_{\mathcal{F}}} \simeq \bigvee_{q\in Q} \Delta(Q_{<q}) * U_q. \tag{17}$$

We next determine the contribution of each $q \in Q$ to (17). If $q = \widehat{1}$ then $\Delta(Q_{<q}) * U_q$ is contractible to the point ∞ and hence does not contribute to (17). Suppose $q < \widehat{1}$ and let $U_q = \Gamma_R$ where $R = \{(F_\ell, \mathbf{j}_\ell)\}_{\ell=1}^r \in A_{\mathcal{F}}$. Combining Proposition 4.4(ii) and (15) it follows that

$$\begin{aligned}\Delta(Q_{<q}) * U_q &\cong \Delta(M(\mathcal{F})_{\prec F_1}) * \cdots * \Delta(M(\mathcal{F})_{\prec F_r}) * S^{r-2} * S^{N-\sum_{(F,\mathbf{j})\in R}(|F|-1)}\\ &\cong S^{N-\sum_{i=1}^{r}(|F_i|-2)-1} * \mathop{\ast}\limits_{i=1}^{r} \Delta(M(\mathcal{F})_{\prec F_i}).\end{aligned}$$

Therefore, if $R \in A_{\mathcal{F}}(\mathbf{m})$ and $U_q = \Gamma_R$ then (cf. (3))

$$\Delta(Q_{<q}) * U_q \cong S^{N-c_{\mathcal{F}}(\mathbf{m})} * \mathop{\ast}\limits_{F\in\mathcal{F}} \Delta(M(\mathcal{F})_{\prec F})^{*\mathbf{m}(F)}. \tag{18}$$

Theorem 3.1 now follows from (17), (18) and Proposition 4.11. □

5 Applications

In this section we use Theorem 1.3 to study several specific Euclidean pattern spaces.

5.1 The Homology of $\vec{P}(\mathbb{R}^n \setminus \mathbb{Z}^n)_{\mathbf{0}}^{\mathbf{k+1}}$

As noted earlier, $\mathbb{R}^n \setminus \mathbb{Z}^n = X_{\mathcal{F}}$ where $\mathcal{F}$ consists of the single set $[n]$. Since $\Delta(M(\mathcal{F})_{\prec[n]})$ is the empty complex $\{\emptyset\}$ it follows that $f_{\mathbb{K}}(\Delta(M(\mathcal{F})_{\prec[n]}), t) = 1$. If $\mathbf{m}([n]) = m > 0$ then $b_{\mathcal{F},\mathbf{k}}(\mathbf{m}) = \prod_{i=1}^{n} \binom{k_i}{m}$ and $c_{\mathcal{F}}(\mathbf{m}) = m(n-2)+1$. Theorem 1.3 implies that

$$f_{\mathbb{K}}(\vec{P}(\mathbb{R}^n \setminus \mathbb{Z}^n)_{\mathbf{0}}^{\mathbf{k+1}}, t) = \sum_{m\geq 1} \prod_{i=1}^{n} \binom{k_i}{m} t^{m(n-2)+1}.$$

Since $\tilde{H}_*(\Delta(M(\mathcal{F})_{\prec[n]})) = \tilde{H}_{-1}(\{\emptyset\}) = \mathbb{Z}$ is free, it follows that $\tilde{H}_\ell(\vec{P}(\mathbb{R}^n \setminus \mathbb{Z}^n)_{\mathbf{0}}^{\mathbf{k+1}})$ is free of rank $\prod_{i=1}^{n} \binom{k_i}{m}$ if $\ell = (n-2)m > 0$, and is zero otherwise. This recovers the above mentioned Theorem 1.1 of Raussen and Ziemiański [13].

5.2 Binary Path Spaces

The *binary path space* associated with an upward closed $\mathcal{F} \subset 2^{[n]}$ is $\vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{2}}$ where $\mathbf{2} = (2, \ldots, 2)$. Note that $\vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{2}}$ is homotopy equivalent to the diagonal subspace arrangement

$$\mathbb{R}^n - \bigcup_{F=\{i_1,\ldots,i_\ell\}\in\mathcal{F}} \{x = (x_1, \ldots, x_n) \in \mathbb{R}^n : x_{i_1} = \cdots = x_{i_\ell}\}.$$

The general formula (5) for the Poincaré series of $\vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$ simplifies in this case as follows. Let $\mathbf{k} = \mathbf{1}$ and let $0 \neq \mathbf{m} \in \mathbb{N}^{\mathcal{F}}$. Then $b_{\mathcal{F},\mathbf{1}}(\mathbf{m}) = 1$ if both $\mathbf{m}(F) \leq 1$ for all $F \in \mathcal{F}$, and $\{F : \mathbf{m}(F) = 1\} \in M(\mathcal{F})$. Otherwise $b_{\mathcal{F},\mathbf{1}}(\mathbf{m}) = 0$. Hence, by (5)

$$f_{\mathbb{K}}\left(\vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{2}}, t\right) = \sum_{\mathcal{G}\in M(\mathcal{F})} t^{\sum_{F\in\mathcal{G}}(|F|-2)+1} \prod_{F\in\mathcal{G}} f_{\mathbb{K}}\left(\Delta(M(\mathcal{F})_{\prec F}), t^{-1}\right). \tag{19}$$

Equation (19) can also be obtained from the general Goresky–MacPherson formula for the homology of subspace arrangements [7].

5.3 The $(s, \mathbf{k})$-Equal Path Space

Let $1 \le s \le n$ and $\mathbf{k} = (k_1, \dots, k_n) \in \mathbb{N}_+^n$. The $(s, \mathbf{k})$-*equal path space* is defined as $\vec{P}(X_{\mathcal{F}_{n,s}})_{\mathbf{0}}^{\mathbf{k+1}}$ where $\mathcal{F}_{n,s} = \{F \subset [n] : |F| \ge s\}$. This path space occurs when every process T_i calls upon a single resource *a of capacity* $s-1$ a number k_i of times.

We use Formula (5) to obtain some information on the homology of this space. For $m \ge s$, let $\overline{\Pi}_{m,s}$ denote the poset of nontrivial partitions of $[m]$ such that every non-singleton block has cardinality at least s. The homology of the order complex $\Delta(\overline{\Pi}_{m,s})$ had been determined by Björner and Welker [3] and was further studied in [2, 9]. We will need the following result:

Theorem 5.1 (Theorem 4.5 in [3], Corollary 6.2 in [2]) $\Delta(\overline{\Pi}_{m,s})$ *has the homotopy type of a wedge of spheres. The d-th Betti number of* $\Delta(\overline{\Pi}_{m,s})$ *is nonzero iff* $d = m - 3 - \ell(s-2)$ *for some* $1 \le \ell \le \lfloor \frac{n}{s} \rfloor$*, and*

$$\tilde{\beta}_{m-3-\ell(s-2)}(\Delta(\overline{\Pi}_{m,s})) = \sum_{\substack{j_1+\cdots+j_\ell=m \\ j_i \ge s}} \binom{m-1}{j_1-1, j_2, \dots, j_\ell} \prod_{i=0}^{\ell-1} \binom{j_i-1}{s-1}. \tag{20}$$

Note that $\Delta(M((\mathcal{F}_{n,s})_{\prec F})) \cong \Delta(\overline{\Pi}_{|F|,s})$ for any $F \in \mathcal{F}_{n,s}$. Theorem 1.3(i) implies that $H_*(\vec{P}(X_{\mathcal{F}_{n,s}})_{\mathbf{0}}^{\mathbf{k+1}})$ is free. Moreover, by Theorem 1.3(ii)

$$\begin{aligned} & f_{\mathbb{K}}\left(\vec{P}(X_{\mathcal{F}_{n,s}})_{\mathbf{0}}^{\mathbf{k+1}}, t\right) \\ &= \sum_{0 \ne \mathbf{m} \in \mathbb{N}^{\mathcal{F}_{n,s}}} b_{\mathcal{F},\mathbf{k}}(\mathbf{m}) t^{c_{\mathcal{F}_{n,s}}(\mathbf{m})} \prod_{F \in \mathcal{F}_{n,s}} f_{\mathbb{K}}\left(\Delta(M((\mathcal{F}_{n,s})_{\prec F})), t^{-1}\right)^{\mathbf{m}(F)} \\ &= \sum_{0 \ne \mathbf{m} \in \mathbb{N}^{\mathcal{F}_{n,s}}} b_{\mathcal{F},\mathbf{k}}(\mathbf{m}) t^{\sum_{F \in \mathcal{F}_{n,s}} \mathbf{m}(F)(|F|-2)+1} \mu(\mathbf{m}, t) \end{aligned} \tag{21}$$

where

$$\mu(\mathbf{m}, t) = \prod_{F \in \mathcal{F}_{n,s}} \left(\sum_{\ell=1}^{\left\lfloor \frac{|F|}{s} \right\rfloor} \tilde{\beta}_{|F|-3-\ell(s-2)}(\Delta(\overline{\Pi}_{|F|,s})) t^{-|F|+2+\ell(s-2)} \right)^{\mathbf{m}(F)}.$$

It follows that t^α appears in $f_{\mathbb{K}}\left(\vec{P}(X_{\mathcal{F}_{n,s}})_{\mathbf{0}}^{\mathbf{k+1}}, t\right)$ with nonzero coefficient only if $\alpha \equiv 1(\mathrm{mod}(s-2))$.

Corollary 5.2 $\tilde{H}_\ell(\vec{P}(X_{\mathcal{F}_{n,s}})_{\mathbf{0}}^{\mathbf{k+1}}; \mathbb{Z}) = 0$ *unless* $\ell = m(s-2)$ *for some* $m > 0$.

5.4 The Connectivity of Path Spaces

The following result determines the homological connectivity of $\overrightarrow{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$.

Proposition 5.3 *Let* $s(\mathcal{F}) = \min_{F\in\mathcal{F}} |F|$. *Then*

$$\min\{i : \tilde{H}_i(\overrightarrow{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}; \mathbb{Z}) \neq 0\} = s(\mathcal{F}) - 2.$$

Proof Choose an $F \in \mathcal{F}$ such that $|F| = s(\mathcal{F})$. Then $\Delta(M(\mathcal{F})_{\prec F})$ is the empty complex $\{\emptyset\}$ and therefore $f_{\mathbb{K}}(\Delta(M(\mathcal{F})_{\prec F}), t^{-1}) = 1$. Letting $\mathbf{m}(F') = 1$ if $F = F'$ and zero otherwise, it follows from Theorem 1.3(ii) that $t^{c_{\mathcal{F}}(\mathbf{m})} = t^{|F|-1} = t^{s(\mathcal{F})-1}$ appears in $f_{\mathbb{K}}(\overrightarrow{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}, t)$ with a positive coefficient, and therefore $\tilde{H}_{s(\mathcal{F})-2}(\overrightarrow{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}; \mathbb{K}) \neq 0$.
For the other direction, first note that for any $F \in \mathcal{F}$

$$\dim \Delta(M(\mathcal{F})_{\prec F}) \leq |F| - s(\mathcal{F}) - 1.$$

Therefore for any $F_1, \ldots, F_r \in \mathcal{F}$

$$\begin{aligned}\dim \mathop{\ast}_{i=1}^{r} \Delta(M(\mathcal{F})_{\prec F_i}) &\leq \sum_{i=1}^{r}(|F_i| - s(\mathcal{F}) - 1) + r - 1\\ &= \sum_{i=1}^{r} |F_i| - rs(\mathcal{F}) - 1\\ &< \sum_{i=1}^{r}(|F_i| - 2) - s(\mathcal{F}) + 2.\end{aligned}$$

Thus

$$\tilde{H}^j(S^{N-\sum_{i=1}^{r}(|F_i|-2)-1} * \mathop{\ast}_{i=1}^{r} \Delta(M(\mathcal{F})_{\prec F_i}) = 0$$

for all

$$\begin{aligned}j &\geq (N - \sum_{i=1}^{r}(|F_i| - 2) - 1) + (\sum_{i=1}^{r}(|F_i| - 2) - s(\mathcal{F}) + 2) + 1\\ &= N - s(\mathcal{F}) + 2.\end{aligned}$$

As $\widehat{E_{\mathcal{F}}}$ is a wedge of spaces of the form

$$S^{N-\sum_{i=1}^{r}(|F_i|-2)-1} * \mathop{\ast}_{i=1}^{r} \Delta(M(\mathcal{F})_{\prec F_i})$$

where $F_1, \ldots, F_r \in \mathcal{F}$, it follows that $\tilde{H}^j(\widehat{E_{\mathcal{F}}};\mathbb{Z}) = 0$ for all $j \geq N - s(\mathcal{F}) + 2$. Finally, Alexander duality $\tilde{H}_i(\overrightarrow{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}};\mathbb{Z}) \cong \tilde{H}^{N-i-1}(\widehat{E_{\mathcal{F}}})$ implies that $\tilde{H}_i(\overrightarrow{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}};\mathbb{Z}) = 0$ for all $i \leq s(\mathcal{F}) - 3$. □

In fact, we establish the following stronger result:

Proposition 5.4 *Let* $\mathbf{p}$ *denote any directed path in* $\vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$. *Then* $\pi_i(\vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}};\mathbf{p}) = 0$ *for all* $i \leq s(\mathcal{F}) - 3$.

Proof According to Proposition 2.2, we may replace $\vec{P}(X_{\mathcal{F}})_{\mathbf{0}}^{\mathbf{k+1}}$ with the homotopy equivalent space $D_{\mathcal{F}} \subset \mathring{\Delta}_{\mathbf{k}}$. Proposition 5.3 tells us that $D_{\mathcal{F}}$ is connected; hence we can choose any base point $\mathbf{p} \in D_{\mathcal{F}}$ in the following. Connectedness can also be concluded from the subsequent argument in the case $i = 0$.

Let $F : S^i \to D_{\mathcal{F}}$ denote any continuous map. Its image $F(S^i)$ is compact and has thus positive distance from the compact set $\overline{E_{\mathcal{F}}} \subset \Delta_{\mathbf{k}}$. F admits a smooth approximation $\tilde{F} : S^i \to \mathring{\Delta}_{\mathbf{k}}$ homotopic to F and so close to F that the image of the homotopy does not intersect $E_{\mathcal{F}}$. Extend $\tilde{F}$ to a smooth map $G : D^{i+1} \to \mathring{\Delta}_{\mathbf{k}}$ by defining $G(0) = \mathbf{p}$ and by convex combination with $\tilde{F}$ on the boundary S^i. The image $G(D^{i+1})$ may intersect $E_{\mathcal{F}}$.

By multiple application of the transversality theorem (see e.g. [8, Theorem III.2.1], [1, Ch. I.2]), one can find a smooth approximation H to G that is transversal to all strata in $E_{\mathcal{F}}$. Moreover, since the compact sets $G(D^{i+1})$ and $\partial\Delta_{\mathbf{k}}$ have a positive distance, we may assume that $H(D^{i+1})$ is contained in $\mathring{\Delta}_{\mathbf{k}}$, as well. Each of the subspaces $G_{F,\mathbf{j}}$ in the definition of $E_{\mathcal{F}}$ has codimension $|F| - 1$ in $\mathbf{R}^N$, and intersections have higher codimensions. In particular, if $i + 1 < |F| - 1$, then $H(D^{i+1}) \cap G_{F,\mathbf{j}} = \emptyset$ by transversality. If $i + 1 < s(\mathcal{F}) - 1$, then $H(D^{i+1}) \cap E_{\mathcal{F}} = \emptyset$ and H establishes that $\tilde{F}$ and hence F are nul-homotopic in $D_{\mathcal{F}}$. □

6 Concluding Remarks

We conclude with a few remarks about possible extensions of the results of this paper that we hope to deal with in future work. One obvious challenge concerns finding maps from spheres, and more generally products of spheres, into path space such that the images of the fundamental classes may serve as generators for homology in the appropriate dimensions, aiming at a generalization of [13, Corollary 3.10] in the paper of Raussen and Ziemiański. This is work in progress.

On the other hand, the situation we analysed is perhaps characterized by more regularity than what is needed for the method to work. The paper of Raussen and Ziemiański [13] calculates the homology of the path space $\vec{P}(X)_{\mathbf{0}}^{\mathbf{k+1}}$ with $X = \mathbb{R}^n \setminus Y$ with Y a *subset* of $\mathbb{Z}^n$. It seems likely that it is possible to extend our results to the following more general situation (with $\mathcal{F}$ an upward closed hypergraph on $[n]$ as previously):

For $F \in \mathcal{F}$ and $\alpha : F \to \mathbb{Z}$ a function, let $Y_\alpha := \{(x_1, \dots, x_n) | \ x_i = \alpha(i), i \in F\}$. For any non-empty subset $\beta(F) \subset \mathbb{Z}^F$ let $Y_{\beta(F)} := \bigcup_{\alpha \in \beta(F)} Y_\alpha$. In the present paper, we only considered $\beta(F) = \mathbb{Z}^F$.

Now we assume that for every $F \in \mathcal{F}$ such a subset $\beta(F)$ has been chosen. Coherence suggests either to make a choice only for minimal elements of the family or to ask that $\beta(F_2)$ consists of *all extensions* of functions in $\beta(F_1)$ to F_2 in case $F_1 \subset F_2$. The set to be excluded is then the union of hyperplanes $Y = \bigcup_{F \in \mathcal{F}} Y_{\beta(F)}$. It seems likely that one can determine the homology of $\vec{P}(X)_{\mathbf{0}}^{\mathbf{k+1}}$ with $X = \mathbb{R}^n \setminus Y$, as well.

It is less obvious how to analyse topological properties of path spaces associated to general PV spaces (cf. Sect. 1) via arrangements – those would no longer be given by restrictions of *linear* subspaces. Instead, one has to remove *thickened* subspace arrangements within products of simplices leading to pattern spaces that are more difficult to analyse. For such thickened arrangements, our method – that makes essential use of the Wedge Lemma – is in general no longer applicable.

Since Ziemiański has shown [18] that every finite simplicial complex can arise as a connected component of the path space for some PV-space, one cannot expect a simple algorithmic determination of the homology of such a path space in general.

References

1. V.I. Arnold, S.M. Gusein-Zade, A.N. Varchenko, *Singularities of Differentiable Maps*, vol. 1 (Birkhäuser, Basel, 1985)
2. A. Björner, M.L. Wachs, Shellable nonpure complexes and posets. I. Trans. Am. Math. Soc. **348**, 1299–1327 (1996)
3. A. Björner, V. Welker, The homology of "*k*-equal" manifolds and related partition lattices. Adv. Math. **110**, 277–313 (1995)
4. E.W. Dijkstra, Co-operating sequential processes, in *Programming Languages*, ed. by F. Genuys (Academic, New York, 1968), pp. 43–110
5. L. Fajstrup, É. Goubault, M. Raussen, Algebraic topology and concurrency. Theor. Comput. Sci. **357**, 241–278 (2006). Revised version of Aalborg University (1999, preprint)
6. L. Fajstrup, É. Goubault, E. Haucourt, S. Mimram, M. Raussen, *Directed Algebraic Topology and Concurrency* (Springer, Berlin, 2016)
7. M. Goresky, R. MacPherson, *Stratified Morse Theory* (Springer, Berlin, 1988)
8. M. Hirsch, *Differential Topology* (Springer, New York, 1976)
9. I. Peeva, V. Reiner, V. Welker, Cohomology of real diagonal subspace arrangements via resolutions. Compositio Math. **117**, 99–115 (1999)
10. V. Pratt, Modelling concurrency with geometry, in *Proceedings of the 18th ACM Symposium on Principles of Programming Languages* (1991), pp. 311–322
11. M. Raussen, Simplicial models for trace spaces. Algebr. Geom. Topol. **10**, 1683–1714 (2010)
12. M. Raussen, Simplicial models for trace spaces II: general higher-dimensional automata. Algebr. Geom. Topol. **12**, 1745–1765 (2012)
13. M. Raussen, K. Ziemiański, Homology of spaces of directed paths on Euclidean cubical complexes. J. Homotopy Relat. Struct. **9**, 67–84 (2014)
14. R.P. Stanley, Acyclic orientations of graphs. Discrete Math. **5**, 171–178 (1973)
15. R. van Glabbeek, On the Expressiveness of higher dimensional automata. Theor. Comput. Sci. **368**, 168–194 (2006)

16. J.W. Walker, Canonical homeomorphisms of posets. Eur. J. Combin. **9**, 97–107 (1988)
17. G. Ziegler, R. Živaljević, Homotopy types of subspace arrangements via diagrams of spaces. Math. Ann. **295**, 527–548 (1993)
18. K. Ziemiański, On execution spaces of PV-programs. Theoret. Comput. Sci. **619**, 87–98 (2016)

Sperner's Colorings and Optimal Partitioning of the Simplex

Maryam Mirzakhani and Jan Vondrák

In memory and dedication to Jirka Matoušek

Abstract We discuss coloring and partitioning questions related to Sperner's Lemma, originally motivated by an application in hardness of approximation. Informally, we call a partitioning of the $(k-1)$-dimensional simplex into k parts, or a labeling of a lattice inside the simplex by k colors, "Sperner-admissible" if color i avoids the face opposite to vertex i. The questions we study are of the following flavor: What is the Sperner-admissible labeling/partitioning that makes the total area of the boundary between different colors/parts as small as possible?

First, for a natural arrangement of "cells" in the simplex, we prove an optimal lower bound on the number of cells that must be non-monochromatic in any Sperner-admissible labeling. This lower bound is matched by a simple labeling where each vertex receives the minimum admissible color.

Second, we show for this arrangement that in contrast to Sperner's Lemma, there is a Sperner-admissible labeling such that every cell contains at most 4 colors.

Finally, we prove a geometric variant of the first result: For any Sperner-admissible partition of the regular simplex, the total surface area of the boundary shared by at least two different parts is minimized by the Voronoi partition $(A_1^*, \dots, A_k^*)$ where A_i^* contains all the points whose closest vertex is $\mathbf{e}_i$. We also discuss possible extensions of this result to general polytopes and some open questions.

The first two results of this paper have also appeared in [10].

M. Mirzakhani • J. Vondrák (✉)
Department of Mathematics, Stanford University, 450 Serra Mall, 94305 Stanford, CA, USA
e-mail: mmirzakh@stanford.edu; jvondrak@stanford.edu

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_25

1 Introduction

Sperner's Lemma is a gem in combinatorics which was originally discovered by Emmanuel Sperner [12] as a tool to derive a simple proof of Brouwer's Fixed Point Theorem. Since then, Sperner's Lemma has seen numerous applications, notably in the proof of existence of mixed Nash equilibria [11], in fair division [13], and recently it played an important role in the study of computational complexity of finding a Nash equilibrium [2, 3]. At a high level, Sperner's Lemma states that for any coloring of a simplicial subdivision of a simplex satisfying certain boundary conditions, there must be a "rainbow cell" that receives all possible colors. We review Sperner's Lemma in Sect. 3.

The starting point of this work was a question that arises in the study of approximation algorithms for a certain hypergraph labeling problem [6, 9, 1, 5, 4]. The question posed by [4], while in some ways reminiscent of Sperner's Lemma, is different in the following sense: Instead of asking whether there exists a rainbow cell for any admissible coloring, the question is what is the minimum possible number of cells that must be *non-monochromatic*. (Also, the question arises for a particular regular lattice inside the simplex rather than an arbitrary subdivision.) In this paper, we resolve this question and investigate some related problems.

Before we state our results, let us note the following connection. As the granularity of the subdivision tends to zero, Sperner's Lemma becomes a statement about certain geometric partitions of the simplex: for any Sperner-admissible partition, where part *i* avoids the face opposite to vertex *i*, there must be a point where all parts meet. This result is known as the Knaster–Kuratowski–Mazurkiewicz Lemma [7]. In contrast, the questions we are studying are concerned with the measure of the boundary where at least two different parts meet: This can be viewed as a multi-colored isoperimetric inequality, where we try to partition the simplex in a certain way, so that the surface area of the union of all pairwise boundaries (what we call a *separating set*) is minimized. The way we measure the separating set also affects the problem; the discrete version of the question that is of primary interest to us is mandated by the application in [4]. In the geometric setting, a natural notion of surface area is the Minkowski content of the separating set (which coincides with other notions of volume for well-behaved sets). We give an optimal answer to this question for a regular simplex and discuss other related questions.

To state our results formally, we need some notation that we introduce in Sect. 2. We postpone our contributions to Sects. 4, 5, and 6, after a discussion of Sperner's Lemma in Sect. 3.

2 Preliminaries

We denote vectors in boldface, such as $\mathbf{v} \in \mathbb{R}^k$. The coordinates of $\mathbf{v}$ are written in italics, such as $\mathbf{v} = (v_1, \ldots, v_k)$. By $\mathbf{e}_i$, we denote the canonical basis vectors $(0, \ldots, 1, \ldots, 0)$. By $\mathrm{conv}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$, we denote the convex hull of the respective vectors.

2.1 *Simplicial Subdivisions of the Simplex*

Consider the $(k-1)$-dimensional simplex defined by

$$\Delta_k = \mathrm{conv}(\mathbf{e}_1, \ldots, \mathbf{e}_k) = \left\{ \mathbf{x} = (x_1, x_2, \ldots, x_k) \in \mathbb{R}^k : \mathbf{x} \geq 0, \sum_{i=1}^{k} x_i = 1 \right\}.$$

Simplicial subdivision A simplicial subdivision of Δ_k is a collection of simplices ("cells") Σ such that

- The union of the cells in Σ is the simplex Δ_k.
- For any two cells $\sigma_1, \sigma_2 \in \Sigma$, their intersection is either empty or a full face of a certain dimension shared by σ_1, σ_2.

The Simplex-Lattice Hypergraph Next, we describe a specific configuration of cells in a simplex; this configuration is actually not a full subdivision since its cells do not cover the full volume of the simplex. It can be completed to a subdivision if desired.[1]

Let $q \geq 1$ be an integer and define

$$\Delta_{k,q} = \left\{ \mathbf{x} = (x_1, x_2, \ldots, x_k) \in \mathbb{R}^k : \mathbf{x} \geq 0, \sum_{i=1}^{k} x_i = q \right\}.$$

We consider a vertex set of all the points in $\Delta_{k,q}$ with integer coordinates:

$$V_{k,q} = \left\{ \mathbf{a} = (a_1, a_2, \ldots, a_k) \in \mathbb{Z}^k : \mathbf{a} \geq 0, \sum_{i=1}^{k} a_i = q \right\}.$$

[1] This specific configuration arises in [4] as an integrality gap example for a certain hypergraph labeling problem; see also [10] for more details.

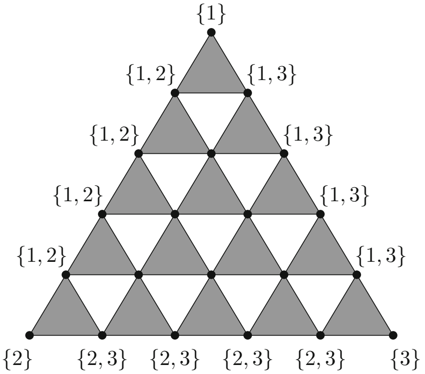


Fig. 1 The Simplex Lattice Hypergraph for $k = 3$ and $q = 5$, with hyperedges shaded in *gray*. The *gray triangles* together with the *white triangles* form a simplicial subdivision. The lists of admissible colors are given on the boundary; for internal vertices the lists are all $\{1, 2, 3\}$

The *Simplex-Lattice Hypergraph* is a k-uniform hypergraph $H_{k,q} = (V_{k,q}, E_{k,q})$ whose hyperedges (which we also call *cells* due to their geometric interpretation) are indexed by $\mathbf{b} \in \mathbb{Z}_+^k$ such that $\sum_{i=1}^k b_i = q - 1$ (Fig. 1): we have

$$E_{k,q} = \left\{ e(\mathbf{b}) : \mathbf{b} \in \mathbb{Z}^k, \mathbf{b} \geq 0, \sum_{i=1}^k b_i = q - 1 \right\}$$

where $e(\mathbf{b}) = \{\mathbf{b} + \mathbf{e}_1, \mathbf{b} + \mathbf{e}_2, \ldots, \mathbf{b} + \mathbf{e}_k\} = \{(b_1 + 1, b_2, \ldots, b_k), (b_1, b_2 + 1, \ldots, b_k), \ldots, (b_1, b_2, \ldots, b_k + 1)\}$. For each vertex $\mathbf{a} \in V_{k,q}$, we have a list of admissible colors $L(\mathbf{a})$, which is

$$L(\mathbf{a}) = \{i \in [k] : a_i > 0\}.$$

3 Sperner's Lemma

First, let us recall the statement of Sperner's Lemma [12]. We consider labelings $\ell : V_{k,q} \to [k]$. We call a labeling ℓ Sperner-admissible if $\ell(\mathbf{a}) \in L(\mathbf{a})$ for each $\mathbf{a} \in V$; i.e. , if $\ell(\mathbf{a}) = j$ then $a_j > 0$.

Lemma 1 (Sperner's Lemma) *For every Sperner-admissible labeling of the vertices of a simplicial subdivision of Δ_k, there is a cell whose vertices receive all k colors.*

We remark that this does not say anything about the Simplex-Lattice Hypergraph: Even if the subdivision uses the point set $V_{k,q}$, the rainbow cell given by Sperner's Lemma might not be a member of $E_{k,q}$ since $E_{k,q}$ consists only of scaled copies of $\Delta_{k,q}$ without rotation; it is not a full subdivision of the simplex. (See Fig. 2.)

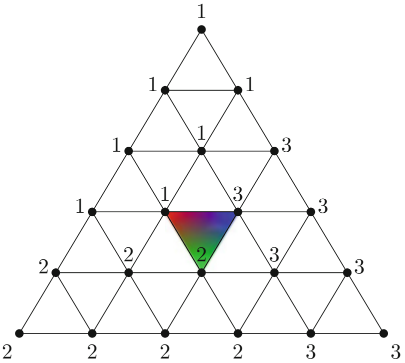


Fig. 2 A Sperner-admissible labeling for $k = 3$ and $q = 5$. At least one cell in the triangulation (not necessarily in $E_{k,q}$) must be k-colored (rainbow)

4 The Simplex-Lattice Coloring Lemma

Instead of rainbow cells, the statement proposed (and proved for $k = 3$) in [4] involves non-monochromatic cells.

Proposition 1 (Simplex-Lattice Coloring Lemma) *For any Sperner-admissible labeling $\ell : V_{k,q} \to [k]$, there are at least $\binom{q+k-3}{k-2}$ hyperedges $e \in E_{k,q}$ that are non-monochromatic under ℓ.*

The first-choice labeling In particular, Proposition 1 is that a Sperner-admissible labeling minimizing the number of non-monochromatic cells is a "first-choice one" which labels each vertex $\mathbf{a}$ by the smallest coordinate i such that $a_i > 0$. Under this labeling, all the hyperedges $e(\mathbf{b})$ such that $b_1 > 0$ are labeled monochromatically by 1. The only hyperedges that receive more than 1 color are those where $b_1 = 0$, and the number of such hyperedges is exactly $\binom{q+k-3}{k-2}$ (see [4]). Here we give a proof of Proposition 1 (Fig. 3).

Proof Consider the set of hyperedges $E_{k,q}$: observe that it can be written naturally as

$$E_{k,q} = \{e(\mathbf{b}) : \mathbf{b} \in V_{k,q-1}\}.$$

I.e., the hyperedges can be identified one-to-one with the vertices in $V_{k,q-1}$. Recall that $e(\mathbf{b}) = \{\mathbf{b} + \mathbf{e}_1, \mathbf{b} + \mathbf{e}_2, \ldots, \mathbf{b} + \mathbf{e}_k\}$. Two hyperedges $e(\mathbf{b}), e(\mathbf{b}')$ share a vertex if and only if $\mathbf{b}' + \mathbf{e}_j = \mathbf{b} + \mathbf{e}_i$ for some pair $i, j \in [k]$; or in other words if $\mathbf{b}, \mathbf{b}'$ are nearest neighbors in $V_{k,q-1}$ (differ by ±1 in exactly two coordinates).

Consider a labeling $\ell : V_{k,q} \to [k]$. For each $i \in [k]$, let C_i denote the set of points in $V_{k,q-1}$ representing the monochromatic hyperedges in color i,

$$C_i = \{\mathbf{b} \in V_{k,q-1} : \forall \mathbf{v} \in e(\mathbf{b}); \ell(\mathbf{v}) = i\}.$$

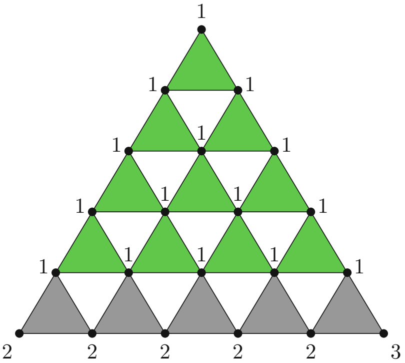


Fig. 3 The first-choice labeling

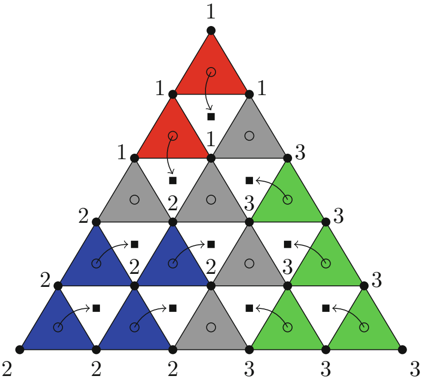


Fig. 4 The mappings $\phi_i : C_i \to V_{k,q-2}$. The hyperedges are represented by the *empty circles*; C_i is the subset of them monochromatic in color i. The *black squares* represent $V_{k,q-2}$; note that each point in $V_{k,q-2}$ is the image of at most one monochromatic hyperedge

Define an injective mapping $\phi_i : C_i \to V_{k,q-2}$ as follows:

$$\phi_i(\mathbf{b}) = \mathbf{b} - \mathbf{e}_i.$$

The image is indeed in $V_{k,q-2}$: if $\mathbf{b} \in C_i$, we have $b_i > 0$, or else $e(\mathbf{b})$ would contain a vertex $\mathbf{a}$ such that $a_i = 0$ and hence $e(\mathbf{b})$ could not be monochromatic in color i. Therefore, $\mathbf{b} - \mathbf{e}_i \in \mathbb{Z}_+^k$ and $(\mathbf{b} - \mathbf{e}_i) \cdot \mathbf{1} = q - 2$ which means $\mathbf{b} - \mathbf{e}_i \in V_{k,q-2}$ (Fig. 4). (Here, $\mathbf{1}$ denotes the all-1's vector.)

Further, we claim that $\phi_i[C_i] \cap \phi_j[C_j] = \emptyset$ for every $i \neq j$. If not, there would be $\mathbf{b} \in C_i$ and $\mathbf{b}' \in C_j$ such that $\mathbf{b} - \mathbf{e}_i = \mathbf{b}' - \mathbf{e}_j$. Then, the point $\mathbf{a} = \mathbf{b} + \mathbf{e}_j = \mathbf{b}' + \mathbf{e}_i$ would be an element of both the hyperedge $e(\mathbf{b})$ and the hyperedge $e(\mathbf{b}')$. This contradicts the assumption that $e(\mathbf{b})$ is monochromatic in color i and $e(\mathbf{b}')$ is monochromatic in color j. So the sets $\phi_i[C_i]$ are pairwise disjoint subsets of $V_{k,q-2}$.

By the definition of ϕ_i, we clearly have $|\phi_i[C_i]| = |C_i|$. We conclude that the total number of monochromatic hyperedges is

$$\sum_{i=1}^{k} |C_i| = \sum_{i=1}^{k} |\phi_i[C_i]| \le |V_{k,q-2}|.$$

The total number of hyperedges is $|E_{k,q}| = |V_{k,q-1}|$. Considering that $|V_{k,q}| = \binom{q+k-1}{k-1}$ (the number of partitions of q into a sum of k nonnegative integers), we obtain that the number of non-monochromatic hyperedges is

$$|E_{k,q}| - \sum_{i=1}^{k} |C_i| \ge |V_{k,q-1}| - |V_{k,q-2}| = \tbinom{q+k-2}{k-1} - \tbinom{q+k-3}{k-1} = \tbinom{q+k-3}{k-2}.$$

□

5 A Labeling of $H_{k,q}$ with at Most 4 Colors on Each Hyperedge

We recall that Sperner's lemma states that any Sperner-admissible labeling of a subdivision of the simplex must contain a simplex with all k colors. The hypergraph $H_{k,q}$ defined in Sect. 2.1 is not a subdivision since it covers only a subset of the large simplex. It is easy to see that the conclusion of Sperner's lemma does not hold for $H_{k,q}$ – for example for $k = 3$, we can label a 2-dimensional triangulation so that exactly one triangle has 3 different colors, and this triangle is not in $E_{3,q}$. (See Fig. 2.) Hence, each triangle in $E_{3,q}$ has at most 2 colors. By an extension of this argument, we can label $H_{k,q}$ so that each hyperedge in $E_{k,q}$ contains at most $k-1$ colors. The question we ask in this section is, what is the minimum c such that there is a Sperner-admissible labeling with at most c different colors on each hyperedge in $E_{k,q}$? We prove the following result.

Proposition 2 *For any $k \ge 4$ and $q \ge k^2$, there is a Sperner-admissible labeling of $H_{k,q} = (V_{k,q}, E_{k,q})$ such that every hyperedge in $E_{k,q}$ contains at most 4 different colors.*

We note that this statement is not true for $q = 1$ and $k > 4$ (since $E_{k,1}$ consists of a single simplex which has k different colors). We have not identified the optimal lower bound on q that allows our statement to hold. Also, the statement could possibly hold with 2 or 3 colors instead of 4; the number 4 is just an artifact of our proof and we have no reason to believe that it is tight.

The intuition behind our construction is as follows: We want to label the vertices so that the number of different colors on each hyperedge is small. A natural choice is to label each vertex $\mathbf{v}$ by its maximum-value coordinate. However, this does not work since a hyperedge in the center of the simplex may receive all k colors. The problem is that this labeling is possibly very sensitive to small changes in $\mathbf{v}$.

A more "robust" labeling is one where we select a subset of "top coordinates" and choose one among them according to another rule. This rule should be such that incrementing the coordinates one at a time does not change the label too many times. One such rule that works well is described below.

Proof We define a labeling $\ell : V_{k,q} \to [k]$ as follows:

- Given $\mathbf{a} \in V_{k,q}$, let $\pi : [k] \to [k]$ be a permutation such that $a_{\pi(1)} \geq a_{\pi(2)} \geq \ldots \geq a_{\pi(k)}$ (and if $a_{\pi(i)} = a_{\pi(i+1)}$, we order π so that $\pi(i) < \pi(i+1)$).
- Define $t(\mathbf{a})$ to be the maximum $t \in [k]$ such that $\forall 1 \leq j \leq t$, $a_{\pi(j)} \geq k - j + 1$. We define the "Top coordinates" of $\mathbf{a}$ to be $Top(\mathbf{a}) = (\pi(1), \ldots, \pi(t(\mathbf{a})))$ (an ordered set).
- We define the label of $\mathbf{a}$ to be $\ell(\mathbf{a}) = \pi(t(\mathbf{a}))$, the index of the "last Top coordinate".

First, we verify that this is a well-defined Sperner-admissible labeling. Since $\sum_{i=1}^{k} a_i = q \geq k^2$, we have $a_{\pi(1)} = \max a_i \geq k$ and hence $1 \leq t(\mathbf{a}) \leq k$. For each $\mathbf{a} \in V_{k,q}$, we have: $a_{\ell(\mathbf{a})} = a_{\pi(t(\mathbf{a}))} \geq k - t(\mathbf{a}) + 1 > 0$, since $t(\mathbf{a}) \leq k$. Therefore, ℓ is Sperner-admissible.

Now, consider a hyperedge $e(\mathbf{b}) = (\mathbf{b} + \mathbf{e}_1, \mathbf{b} + \mathbf{e}_2, \ldots, \mathbf{b} + \mathbf{e}_k)$ where $\mathbf{b} \geq 0$, $\sum_{i=1}^{k} b_i = q - 1$. We claim that $\ell(\mathbf{b} + \mathbf{e}_i)$ attains at most 4 different values for $i = 1, \ldots, k$. Without loss of generality, assume that $b_1 \geq b_2 \geq \ldots \geq b_k$. Define ℓ^* to be the label assigned to $\mathbf{b}$ by our construction (note that $\mathbf{b}$ is not a vertex in $V_{k,q}$ but we can still apply our definition): ℓ^* is the maximum value in $[k]$ such that for all $1 \leq j \leq \ell^*$, $b_j \geq k - j + 1$. Hence, we have $Top(\mathbf{b}) = \{1, 2, \ldots, \ell^*\}$.

Let $i \in [k]$, $\mathbf{a} = \mathbf{b} + \mathbf{e}_i$, and let π be the permutation such that $a_{\pi(1)} \geq \ldots \geq a_{\pi(k)}$ as above. (Recall that for $\mathbf{b}$, we assumed that the respective permutation is the identity.) We consider the following cases:

- If $1 \leq i < \ell^*$, then we claim that $\ell(\mathbf{a}) = \ell(\mathbf{b} + \mathbf{e}_i) = \ell(\mathbf{b}) = \ell^*$. In the rule for selecting $t(\mathbf{a})$, one of the first $\ell^* - 1$ coordinates has been incremented compared to $\mathbf{b}$, which possibly pushes i forward in the ordering of the Top coordinates. However, the other coordinates remain unchanged, the condition $a_{\pi(j)} \geq k - j + 1$ is still satisfied for $1 \leq j \leq \ell^*$, and $Top(\mathbf{a}) = Top(\mathbf{b})$. In particular ℓ^* is still the last coordinate included in $Top(\mathbf{a})$ and hence $\ell(\mathbf{a}) = \ell^*$.
- If $i = \ell^*$, then $\ell(\mathbf{a}) = \ell(\mathbf{b} + \mathbf{e}_{\ell^*})$ is still one of the coordinates in $Top(\mathbf{b})$, possibly different from ℓ^* (due to a change in order, although we still have $Top(\mathbf{a}) = Top(\mathbf{b})$) – let us call this label ℓ_2^*.
- If $\ell^* < i \leq k$, then it is possible that in $\mathbf{a} = \mathbf{b} + \mathbf{e}_i$, we obtain additional Top coordinates ($Top(\mathbf{a}) \supset Top(\mathbf{b})$). It could be $a_i = b_i + 1$ itself which is now included among the Top coordinates, and possibly additional coordinates that already satisfied the condition $b_j \geq k - j + 1$ but were not selected due to the condition being false for b_{ℓ^*+1}. If this does not happen and we have $Top(\mathbf{a}) = Top(\mathbf{b})$, the label of $\mathbf{a}$ is still $\ell(\mathbf{a}) = \ell^*$ (because the ordering of the Top coordinates remains the same).

Assume now that *Top*(**a**) has additional coordinates beyond *Top*(**b**). By the definition of ℓ^*, we have $b_{\ell^*} \geq k - \ell^* + 1$, and for each $j > \ell^*$, we have $b_j < k - \ell^*$; otherwise j would have been still chosen in *Top*(**b**). For $Top(\mathbf{a}) = Top(\mathbf{b} + \mathbf{e}_i)$ to grow beyond *Top*(**b**), a_i must become the $(\ell^* + 1)$-largest coordinate and satisfy $a_i \geq k - \ell^*$. The only way this can happen is that $b_i = k - \ell^* - 1$ and hence $a_i = b_i + 1 = k - \ell^*$. In this case, a_i is the maximum coordinate among $\{a_j : j > \ell^*\}$, and still smaller than a_{ℓ^*}. Therefore, i will be included in *Top*(**a**). Now, *Top*(**a**) may grow further. However, note that the construction of *Top*(**a**) will proceed in the same way for every $\mathbf{a} = \mathbf{b} + \mathbf{e}_i$ such that $b_i = k - \ell^* - 1$. This is because all the coordinates equal to $k - \ell^* - 1$ will be certainly included in *Top*(**a**), and coordinates smaller than $k - \ell^* - 1$ remain the same in each of these cases (equal to the coordinates of **b**). Therefore, the set *Top*(**a**) will be the same in all these cases; let us call this set Top_+.

The label assigned to $\mathbf{a} = \mathbf{b} + \mathbf{e}_i$ is the index of the last coordinate included in $Top_+ = Top(\mathbf{a})$. Since Top_+ is the same whenever $Top(\mathbf{a}) \neq Top(\mathbf{b})$, the label of **a** will be the coordinate j^* minimizing b_j (and maximizing j to break ties) among all $j \in Top_+$, unless $j^* = i$ in which case the last included coordinate might be another one. This gives potentially two additional colors, let us call them ℓ_3^*, ℓ_4^*, that are assigned to $\mathbf{a} = \mathbf{b} + \mathbf{e}_i$ for all $i > \ell^*$ where $b_i = k - \ell^* - 1$. For other choices of $i > \ell^*$, we have $Top(\mathbf{b} + \mathbf{e}_i) = Top(\mathbf{b})$ and the label assigned to $\mathbf{b} + \mathbf{e}_i$ is $\ell(\mathbf{b} + \mathbf{e}_i) = \ell^*$.

To summarize, all the colors that appear in the labeling of $e(\mathbf{b})$ are included in $\{\ell^*, \ell_2^*, \ell_3^*, \ell_4^*\}$. □

6 Boundary-Minimizing Partitioning of the Simplex

Let us turn now to a geometric variant of Proposition 1. We recall that Sperner's Lemma has a geometric variant known as the Knaster–Kuratowski–Mazurkiewicz Lemma [7]:

Consider a covering of the simplex Δ_k *by closed sets* $A_1, \ldots, A_k$ *such that each point* $\mathbf{x} \in \Delta_k$ *is contained in some set* A_i *such that* $x_i > 0$. *Then* $\bigcap_{i=1}^k A_i \neq \emptyset$.

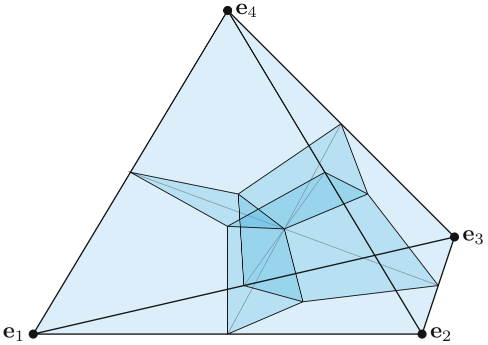
Here we consider a similar setup, but instead of the intersection of all sets, we are interested in the measure of the boundaries between pairs of adjacent sets. To avoid technicalities, let us assume that the A_i's are closed, disjoint except on the boundary, and each A_i is disjoint from the face $\{\mathbf{x} \in \Delta_k : x_i = 0\}$.

Definition 1 A Sperner-admissible partition of Δ_k is a k-tuple of closed sets $(A_1, \ldots, A_k)$ such that

- $\bigcup_{i=1}^k A_i = \Delta_k$,
- $A_1, \ldots, A_k$ are disjoint except on their boundary,
- $x_i > 0$ for every $\mathbf{x} \in A_i$.

We call the union of pairwise boundaries $\bigcup_{i \neq j}(A_i \cap A_j)$ the *separating set*.

Fig. 5 The Voronoi partition of a simplex



The question we ask here is, in analogy with Proposition 1, what is the Sperner-admissible partition with the separating set of minimum measure? A candidate partition is depicted in Fig. 5, where A_i is the set of all points in Δ_k for whom $\mathbf{e}_i$ is the closest vertex. We call this the *Voronoi partition*.

We prove that for the regular simplex Δ_k this is indeed the optimal partition (along with other, similar configurations). In the following, we denote by μ_k the usual Lebesgue measure on $\mathbb{R}^k$, and by μ_ℓ ($\ell < k$) the ℓ-dimensional Minkowski content.

Definition 2 For $A \subset \mathbb{R}^k$, the ℓ-dimensional Minkowski content is (if the limit exists)

$$\mu_\ell(A) = \lim_{\epsilon \to 0^+} \frac{\mu_k(A_\epsilon)}{\alpha_{k-\ell}\epsilon^{k-\ell}}$$

where $A_\epsilon = \{\mathbf{y} \in \mathbb{R}^k : \exists \mathbf{x} \in A, \|\mathbf{x} - \mathbf{y}\| \le \epsilon\}$ is the ϵ-neighborhood of A and $\alpha_{k-\ell}$ is the volume of a unit ball in $\mathbb{R}^{k-\ell}$. We also define $\mu_\ell^+(A)$ to be the upper limit and $\mu_\ell^-(A)$ the lower limit of the expression above.

We remark that for ℓ-rectifiable sets (polyhedral faces, smooth surfaces, etc.) the notion of Minkowski content coincides with that of Hausdorff measure (under suitable normalization).

Theorem 1 *For every Sperner-admissible partition* $(A_1, \dots, A_k)$ *of* Δ_k*,*

$$\mu_{k-2}^-\left(\bigcup_{i \ne j}(A_i \cap A_j)\right) \ge \frac{k-1}{\sqrt{2}}\mu_{k-1}(\Delta_k)$$

and the Voronoi partition achieves this with equality.

First, let us analyze the Voronoi partition and more generally the following kind of partition.

Lemma 2 *For any* $\mathbf{z}$ *in the interior of* Δ_k*, the partition* $(A_1^z, \ldots, A_k^z)$ *where*

$$A_i^z = \{\mathbf{x} \in \Delta_k : x_i - z_i = \max_{1 \le j \le k} (x_j - z_j)\}$$

satisfies

$$\mu_{k-2}\left(\bigcup_{i \neq j} (A_i^z \cap A_j^z)\right) = \frac{k-1}{\sqrt{2}} \mu_{k-1}(\Delta_k) = \frac{1}{(k-2)!}\sqrt{\frac{k}{2}}.$$

We call this kind of partition "Voronoi-type".[2] We note that for $\mathbf{z} = (\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k})$ we obtain the Voronoi partition in Fig. 5. Other choices of $\mathbf{z}$ correspond to similar configurations where all the colors meet at the point $\mathbf{z}$. Note that $\mathbf{z}$ is the "rainbow point" guaranteed by the Knaster-Kuratowski-Mazurkiewicz Lemma.

Proof First let us compute some basic quantities that we will need. The sides of our simplex Δ_k have length $\sqrt{2}$. Denote by h_k the height of Δ_k, that is the distance of any vertex from the opposite facet. We have

$$h_k = \left\| (1, 0, \ldots, 0) - \left(0, \frac{1}{k-1}, \ldots, \frac{1}{k-1}\right) \right\| = \sqrt{1 + (k-1) \cdot \frac{1}{(k-1)^2}} = \sqrt{\frac{k}{k-1}}.$$

The volume of the simplex can be computed inductively as follows; we have $\mu_1(\Delta_2) = \sqrt{2}$, and $\mu_k(\Delta_{k+1}) = \frac{1}{k} h_{k+1} \cdot \mu_{k-1}(\Delta_k)$. This implies

$$\mu_{k-1}(\Delta_k) = \frac{\sqrt{k}}{(k-1)!}.$$

Now let us compute the measure of the separating set for the partition $(A_1^z, \ldots, A_k^z)$ defined above, by induction. The separating set can be described explicitly as

$$\bigcup_{i \neq j} (A_i^z \cap A_j^z) = \{\mathbf{x} \in \Delta_k : \exists i \neq j, x_i - z_i = x_j - z_j = \max_{1 \le \ell \le k} x_\ell - z_\ell\}.$$

For $k = 2$, $A_1^z \cap A_2^z$ is just a single point, and $\mu_0(A_1^z \cap A_2^z) = 1$. For $k \ge 3$, denote by S the separating set for $(A_1^z, \ldots, A_k^z)$ and define $S_i = S \cap \mathrm{conv}(\{\mathbf{e}_j : j \neq i\})$, the separating set restricted to the facet opposite vertex $\mathbf{e}_i$. Since S_i is a Voronoi-type separating set for Δ_{k-1}, by induction we assume that $\mu_{k-3}(S_i) = \frac{1}{(k-3)!}\sqrt{\frac{k-1}{2}}$.

[2] We note that these partitions are also known as "power diagrams".

The separating set S can be written as $S = \bigcup_{i=1}^{k} \mathrm{conv}(S_i \cup \{\mathbf{z}\})$, see Fig. 5. Denote by h'_i the distance of $\mathbf{z}$ from the facet containing S_i. By the pyramid formula in dimension $k-2$,

$$\mu_{k-2}(\mathrm{conv}(S_i \cup \{\mathbf{z}\})) = \frac{1}{k-2} h'_i \mu_{k-3}(S_i) = \frac{h'_i}{(k-2)!}\sqrt{\frac{k-1}{2}}.$$

By a simple calculation, $\sum_{i=1}^{k} h'_i = h_k = \sqrt{\frac{k}{k-1}}$. The sets $\mathrm{conv}(S_i \cup \{\mathbf{z}\})$ are disjoint except for lower-dimensional intersections. Hence,

$$\mu_{k-2}(S) = \sum_{i=1}^{k} \mu_{k-2}(\mathrm{conv}(S_i \cup \{\mathbf{z}\})) = \sum_{i=1}^{k} \frac{h'_i}{(k-2)!}\sqrt{\frac{k-1}{2}} = \frac{1}{(k-2)!}\sqrt{\frac{k}{2}}.$$

□

Thus the proof of Theorem 1 will be complete if we prove the following bound.

Lemma 3 *For every Sperner-admissible partition* $(A_1, \ldots, A_k)$ *of* Δ_k,

$$\mu^{-}_{k-2}\left(\bigcup_{i \neq j} (A_i \cap A_j)\right) \geq \frac{k-1}{\sqrt{2}} \mu_{k-1}(\Delta_k) = \frac{1}{(k-2)!}\sqrt{\frac{k}{2}}.$$

Proof We pursue an approach similar to the proof of Proposition 1, with some additional technicalities. The high-level approach is to shrink the sets A_i somewhat, by excluding a small neighborhood of the separating set. This creates a buffer zone between the shrunk sets A'_i (yellow in Fig. 6) whose measure corresponds to the measure of the separating set. Since we have this extra space, we are able to push the sets A'_i closer together and obtain sets A''_i that fit inside a slightly smaller simplex. The difference between the volume of this simplex and the original one gives a bound on the measure of the separating set.

First, let $\epsilon_0 = \inf_{i \in [k], \mathbf{x} \in A_i} x_i$. Recall that $x_i > 0$ for each $\mathbf{x} \in A_i$, and moreover each A_i is closed. Hence $\epsilon_0 > 0$.

Define $S = \bigcup_{i \neq j} (A_i \cap A_j)$, the separating set whose measure we are trying to lower-bound. Fix $\epsilon \in (0, \frac{1}{2}\epsilon_0)$ (eventually we will let $\epsilon \to 0$) and define S_ϵ as the ϵ-neighborhood of S,

$$S_\epsilon = \{\mathbf{x} \in \Delta_k : \exists \mathbf{y} \in S, \|\mathbf{x} - \mathbf{y}\| \leq \epsilon\}.$$

We define subsets $A'_i \subset A_i$ as follows:

$$A'_i = A_i \setminus S_\epsilon.$$

Thus we have $\bigcup_{i=1}^{k} A'_i = \Delta_k \setminus S_\epsilon$. Also, the sets A'_i are clearly disjoint (see Fig. 6).

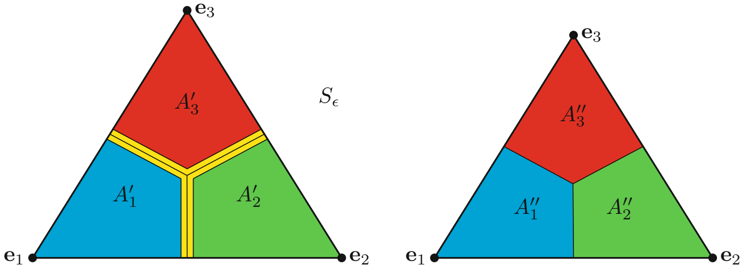


Fig. 6 The construction of A_i' and A_i''

Next, we set $\epsilon' = \epsilon\sqrt{2}$ and define

$$A_i'' = A_i' - \epsilon' \mathbf{e}_i = \{\mathbf{x} - \epsilon' \mathbf{e}_i : \mathbf{x} \in A_i'\}.$$

Thus A_i'' is a shifted copy of A_i', where we push A_i' slightly away from vertex $\mathbf{e}_i$. The sets A_i'' live in the hyperplane $\sum_{i=1}^k x_i = 1 - \epsilon'$ rather than $\sum_{i=1}^k x_i = 1$. We claim that the sets A_i'' are still disjoint: Suppose that $A_i'' \cap A_j'' = (A_i' - \epsilon' \mathbf{e}_i) \cap (A_j' - \epsilon' \mathbf{e}_j) \neq \emptyset$. This would mean that there are points $\mathbf{x} \in A_i', \mathbf{y} \in A_j'$ such that $\mathbf{x} - \epsilon' \mathbf{e}_i = \mathbf{y} - \epsilon' \mathbf{e}_j$. In other words, $\|\mathbf{x} - \mathbf{y}\| = \epsilon' \|\mathbf{e}_i - \mathbf{e}_j\| = \epsilon' \sqrt{2} = 2\epsilon$. Take the midpoint $\frac{1}{2}(\mathbf{x} + \mathbf{y})$: this point is in the simplex Δ_k (by convexity), and hence it is in some set A_ℓ, where either $\ell \neq i$ or $\ell \neq j$ (possibly both). Assume without loss of generality that $\ell \neq i$. Then by the closedness of A_i and A_ℓ, between $\mathbf{x}$ and $\frac{1}{2}(\mathbf{x} + \mathbf{y})$ there exists a point $\mathbf{x}' \in A_i \cap A_\ell$. We get a contradiction, because $\|\mathbf{x}' - \mathbf{x}\| \leq \epsilon$ and so $\mathbf{x}$ would not be included in A_i'.

We also observe that $A_i'' \subseteq (1 - \epsilon') \cdot \Delta_k = \{\mathbf{x} \geq 0 : \sum_{i=1}^k x_i = 1 - \epsilon'\}$. This is because for every $\mathbf{x} \in A_i''$, we have $\mathbf{y} \in A_i''$ such that $\mathbf{x} = \mathbf{y} - \epsilon' \mathbf{e}_i$. By assumption, $y_i \geq \epsilon_0 > \epsilon'$, and $\mathbf{y} \in \Delta_k$. Therefore $x_i = y_i - \epsilon' > 0$ and $\sum_{i=1}^k x_i = 1 - \epsilon'$. We conclude that $A_1'', \ldots, A_k''$ are disjoint subsets of $(1 - \epsilon') \cdot \Delta_k$, obtained by an isometry from $A_1', \ldots, A_k'$ and therefore

$$\sum_{i=1}^k \mu_{k-1}(A_i') = \sum_{i=1}^k \mu_{k-1}(A_i'') \leq (1 - \epsilon')^{k-1} \mu_{k-1}(\Delta_k).$$

Recall that $A_1', \ldots, A_k'$ are also disjoint and $\bigcup_{i=1}^k A_i' = \Delta_k \setminus S_\epsilon$. Therefore,

$$\mu_{k-1}(S_\epsilon) = \mu_{k-1}(\Delta_k) - \sum_{i=1}^k \mu_{k-1}(A_i') \geq \left(1 - (1 - \epsilon')^{k-1}\right) \mu_{k-1}(\Delta_k).$$

By the definition of Minkowski content, we have

$$\mu^-_{k-2}(S) = \liminf_{\epsilon\to 0+} \frac{\mu_{k-1}(S_\epsilon)}{2\epsilon} \geq \liminf_{\epsilon\to 0+} \frac{1-(1-\epsilon')^{k-1}}{2\epsilon}\mu_{k-1}(\Delta_k)$$
$$= \lim_{\epsilon\to 0+} \frac{1-(1-\epsilon\sqrt{2})^{k-1}}{2\epsilon}\mu_{k-1}(\Delta_k) = \frac{k-1}{\sqrt{2}}\mu_{k-1}(\Delta_k).$$

□

Alternative proof of optimality Here we give an alternative proof that the Voronoi partition has a separating set of minimum Minkowski content, avoiding an explicit computation of its volume.

Proof of Lemma 2 Let us consider the Voronoi partition $(A^*_1, \dots, A^*_k)$ (the proof for a general $\mathbf{z}$ is similar). We argue that the proof of Lemma 3 is tight for this partition. As in the proof of Lemma 3, we define $S^* = \bigcup_{i\neq j}(A^*_i \cap A^*_j)$, $S^*_\epsilon = \{\mathbf{x} \in \Delta_k : \exists \mathbf{y} \in S, \|\mathbf{x}-\mathbf{y}\| \leq \epsilon\}$, $A'_i = A^*_i \setminus S^*_\epsilon$ and $A''_i = A'_i - \epsilon'\mathbf{e}_i$, $\epsilon' = \epsilon\sqrt{2}$. In the case of the Voronoi partition, these sets are explicitly described as follows:

- $A^*_i = \{\mathbf{x} \in \Delta_k : x_i = \max_{\ell\in[k]} x_\ell\}$,
- $S^* = \{\mathbf{x} \in \Delta_k : \exists i \neq j, x_i = x_j = \max_{\ell\in[k]} x_\ell\}$,
- $A'_i = \{\mathbf{x} \in \Delta_k : x_i > \epsilon' + \max_{\ell\neq i} x_\ell\}$,
- $A''_i = \{\mathbf{x} - \epsilon'\mathbf{e}_i : \mathbf{x} \in \Delta_k, x_i > \epsilon' + \max_{\ell\neq i} x_\ell\}$.

The description of A'_i is valid because for $\mathbf{x} \in A^*_i$, it is possible to find a point in S^* within distance ϵ of $\mathbf{x}$ if and only if the maximum coordinate x_i is within $\epsilon' = \epsilon\sqrt{2}$ of the second largest coordinate – then we can replace the two largest coordinates by their average and obtain a point in S^*. The description of A''_i follows by definition.

Consider now the scaled-down simplex $(1-\epsilon')\cdot\Delta_k$. By the proof of Lemma 3, the sets A''_i are disjoint subsets of $(1-\epsilon')\cdot\Delta_k$. We show that in this case, we actually have $\sum_{i=1}^k \mu_{k-1}(A''_i) = \mu_{k-1}((1-\epsilon')\Delta_k)$. This is because for any point $\mathbf{x}' \in (1-\epsilon')\cdot\Delta_k$, if the maximum coordinate x'_i of $\mathbf{x}'$ is unique then $\mathbf{x} = \mathbf{x}' + \epsilon'\mathbf{e}_i$ is a point in Δ_k such that $x_i > \epsilon' + \max_{\ell\neq i} x_\ell$. Therefore, $\mathbf{x} \in A'_i$ which implies that $\mathbf{x}' \in A''_i$. The points $\mathbf{x}' \in (1-\epsilon')\cdot\Delta_k$ whose maximum coordinate is not unique form a set of $(k-1)$-dimensional measure zero. Therefore, $(1-\epsilon')\Delta_k$ is covered by $\bigcup_{i=1}^k A''_i$ up to a set of measure zero, and $\sum_{i=1}^k \mu_{k-1}(A'_i) = \sum_{i=1}^k \mu_{k-1}(A''_i) = \mu_{k-1}((1-\epsilon')\Delta_k) = (1-\epsilon\sqrt{2})^{k-1}\mu_{k-1}(\Delta_k)$. We also have $S^*_\epsilon = \Delta_k \setminus \bigcup_{i=1}^k A'_i$. This shows that all the inequalities in the proof of Lemma 3 are tight and the Minkowski content of the separating set S^* is exactly

$$\mu_{k-2}(S^*) = \lim_{\epsilon\to 0+} \frac{\mu_{k-1}(S^*_\epsilon)}{2\epsilon} = \lim_{\epsilon\to 0+} \frac{1-(1-\epsilon\sqrt{2})^{k-1}}{2\epsilon}\mu_{k-1}(\Delta_k) = \frac{k-1}{\sqrt{2}}\mu_{k-1}(\Delta_k).$$

□

7 Discussion and Open Questions

Sperner's Lemma extends to general polytopes in the following sense [8]:
For any coloring of a triangulation of a d-dimensional polytope with n vertices by n colors, such that each point on a face $F = conv(\{\mathbf{v}_i : i \in A\})$ must be colored with a color in A, there are at least $n - d$ full-dimensional simplices with $d + 1$ distinct colors.

It is natural ask whether our results also extend to general polytopes.

Possible extensions to polytopes Consider the example of P being a square (Fig. 7). The Voronoi partition (A_1, A_2, A_3, A_4) is not optimal with respect to the total length of the separating set. The separating set of the Voronoi partition has total length 2, whereas total length arbitrarily close to $\sqrt{2}$ is achieved by the partition (B_1, B_2, B_3, B_4).

In general, we do not know what the partition minimizing $\mu(\bigcup_{i \neq j}(A_i \cap A_j))$ looks like, even in the case of a non-regular simplex. We believe that the separating set should still be polyhedral (piecewise linear) for an optimal Sperner-admissible partition of any polytope.

We remark that depending on the coloring conditions on the surface of the polyhedron, the optimal separating set may be non-linear: For a tetrahedron, the optimal partition that separates the pair of faces $\text{conv}(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) \cup \text{conv}(\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4)$ from $\text{conv}(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_4) \cup \text{conv}(\mathbf{e}_1, \mathbf{e}_3, \mathbf{e}_4)$, is the minimal surface whose boundary is the non-planar 4-gon $\mathbf{e}_1$-$\mathbf{e}_2$-$\mathbf{e}_4$-$\mathbf{e}_3$. This is a saddle-shaped quadratic surface (see Fig. 8).

Other open questions We have proved several results about colorings of the simplex. Our first result (Proposition 1) can be viewed as being at the opposite end of the spectrum from Sperner's Lemma: Instead of the existence of a rainbow cell, we proved a lower bound on the number of non-monochromatic cells. Due to the motivating application of [4], we considered a special hypergraph embedded in the simplex rather than a full subdivision. A natural question is whether an analogous statement holds for simplicial subdivisions.

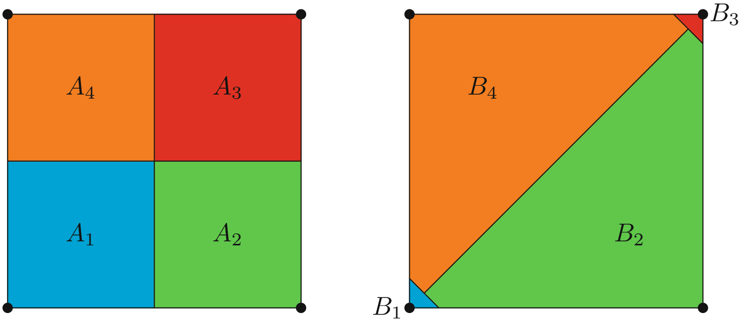


Fig. 7 Two partitions of a square

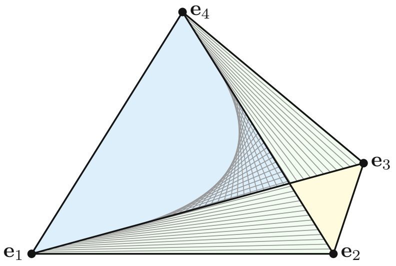


Fig. 8 An optimal partition between two pairs of faces of the tetrahedron

More generally, we might "interpolate" between Sperner's Lemma and our result, and ask: How many cells must contain at least j colors? It is clear that these questions depend on the structure of the subdivision, and some assumption of regularity would be needed to obtain a general result. Similarly, we may ask, for Sperner-admissible geometric partitions of the simplex, what is the minimum possible volume of the set where at least j colors meet? Furthermore, as we discussed above, are there generalizations of these statements to other polytopes?

Another question is, what is the Sperner-admissible labeling of the Simplex-Lattice Hypergraph $H_{k,q}$ (defined in Sect. 2) minimizing the maximum number of colors on a hyperedge? We have proved that 4 colors suffice but it is possible that 2 colors are enough (see Proposition 2). Is there a Sperner-admissible labeling of the hypergraph $H_{k,q}$, for sufficiently large q, such that each hyperedge uses at most 2 colors?

Finally, we remark that Proposition 2 does not have a continuous counterpart for geometric partitions: As we discussed earlier, for any Sperner-admissible partition of a simplex there is a point where all the parts meet, by the Knaster–Kuratowski–Mazurkiewicz Lemma [7].

Acknowledgements The second author is indebted in many ways to Jirka Matoušek, who introduced him to Sperner's Lemma in an undergradute course at Charles University a long time ago.

References

1. C. Chekuri, A. Ene, Submodular cost allocation problem and applications, in *Proceedings of the ICALP*, 2011, pp. 354–366
2. X. Chen, X. Deng, S.-H. Teng, Settling the complexity of computing two-player Nash equilibria. J. ACM **56**, 3 (2009)
3. C. Daskalakis, P.W. Goldberg, C.H. Papadimitriou, The complexity of computing a Nash equilibrium. SIAM J. Comput. **39**(1), 195–259 (2009)
4. A. Ene, J. Vondrák, Hardness of submodular cost allocation: lattice matching and a simplex coloring conjecture, in *Proceedings of the APPROX*, 2014, pp. 144–159
5. A. Ene, J. Vondrák, Y. Wu, Local distribution and the symmetry gap: approximability of multiway partitioning problems, in *Proceedings of the ACM-SIAM SODA*, 2013, pp. 306–325

6. J. Kleinberg, É. Tardos, Approximation algorithms for classification problems with pairwise relationships: metric labeling and Markov random fields. J. ACM **49**(5), 616–639 (2002)
7. B. Knaster, C. Kuratowski, S. Mazurkiewicz, Ein Beweis des Fixpunktsatzes für *n*-dimensionale Simplexe. Fundamenta Mathematicae **14**(1), 132–137 (1929)
8. J. de Loera, E. Peterson, F.E. Su, A polytopal generalization of Sperner's lemma. J. Comb. Theory A **100**, 1–26 (2002)
9. R. Manokaran, J. Naor, P. Raghavendra, R. Schwartz, SDP gaps and UGC hardness for multiway cut, 0-extension, and metric labeling, in *Proceedings of the ACM STOC*, 2008, pp. 11–20
10. M. Mirzakhani, J. Vondrák, Sperner's colorings, hypergraph labeling problems and fair division, in *Proceedings of the ACM-SIAM SODA*, 2015, pp. 873–886
11. J. Nash, Noncooperative games. Ann. Math. **54**, 289–295 (1951)
12. E. Sperner, Neuer Beweis für die Invarianz der Dimensionszahl und des Gebietes. Math. Sem. Univ. Hamburg **6**, 265–272 (1928)
13. F.E. Su, Rental harmony: Sperner's lemma in fair division. Am. Math. Mon. **106**, 930–942 (1999)

Teaching and Compressing for Low VC-Dimension

Shay Moran, Amir Shpilka, Avi Wigderson, and Amir Yehudayoff

Abstract In this work we study the quantitative relation between VC-dimension and two other basic parameters related to learning and teaching. Namely, the quality of sample compression schemes and of teaching sets for classes of low VC-dimension. Let C be a binary concept class of size m and VC-dimension d. Prior to this work, the best known upper bounds for both parameters were $\log(m)$, while the best lower bounds are linear in d. We present significantly better upper bounds on both as follows. Set $k = O(d2^d \log\log|C|)$.

We show that there always exists a concept c in C with a teaching set (i.e. a list of c-labeled examples uniquely identifying c in C) of size k. This problem was studied by Kuhlmann (On teaching and learning intersection-closed concept classes. In: EuroCOLT, pp 168–182, 1999). Our construction implies that the recursive teaching (RT) dimension of C is at most k as well. The RT-dimension was suggested by Zilles et al. (J Mach Learn Res 12:349–384, 2011) and Doliwa et al. (Recursive teaching

A preliminary version of this work, combined with the paper "Sample compression schemes for VC classes" by the first and the last authors, was published in the proceeding of FOCS'15.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement number 257575, and from the Israel Science Foundation (grant number 339/10).

Horev fellow – supported by the Taub foundation. Research is also supported by ISF and BSF.

S. Moran (✉)
Department of Computer Science, Technion-IIT, Haifa, Israel

Max Planck Institute for Informatics, Saarbrücken, Germany
e-mail: shaymrn@cs.technion.ac.il

A. Shpilka
Department of Computer Science, Tel Aviv University, Tel Aviv-Yafo, Israel
e-mail: shpilka@post.tau.ac.il

A. Wigderson
School of Mathematics, Institute for Advanced Study, Princeton, NJ, USA
e-mail: avi@ias.edu

A. Yehudayoff
Department of Mathematics, Technion-IIT, Haifa, Israel
e-mail: amir.yehudayoff@gmail.com

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_26

dimension, learning complexity, and maximum classes. In: ALT, pp 209–223, 2010). The same notion (under the name partial-ID width) was independently studied by Wigderson and Yehudayoff (Population recovery and partial identification. In: FOCS, pp 390–399, 2012). An upper bound on this parameter that depends only on d is known just for the very simple case $d = 1$, and is open even for $d = 2$. We also make small progress towards this seemingly modest goal.

We further construct sample compression schemes of size k for C, with additional information of $k \log(k)$ bits. Roughly speaking, given any list of C-labelled examples of arbitrary length, we can retain only k labeled examples in a way that allows to recover the labels of all others examples in the list, using additional $k \log(k)$ information bits. This problem was first suggested by Littlestone and Warmuth (Relating data compression and learnability. Unpublished, 1986).

1 Introduction

The study of mathematical foundations of learning and teaching has been very fruitful, revealing fundamental connections to various other areas of mathematics, such as geometry, topology, and combinatorics. Many key ideas and notions emerged from this study: Vapnik and Chervonenkis's VC-dimension [44], Valiant's seminal definition of PAC learning [43], Littlestone and Warmuth's sample compression schemes (Littlestone and Warmuth, Relating data compression and learnability. Unpublished, 1986), Goldman and Kearns's teaching dimension [19], recursive teaching dimension (RT-dimension, for short)[12, 39, 47] and more.

While it is known that some of these measures are tightly linked, the exact relationship between them is still not well understood. In particular, it is a long standing question whether the VC-dimension can be used to give a universal bound on the size of sample compression schemes, or on the RT-dimension.

In this work, we make progress on these two questions. First, we prove that the RT-dimension of a boolean concept class C having VC-dimension d is upper bounded by[1] $O(d2^d \log\log|C|)$. Secondly, we give a sample compression scheme of size $O(d2^d \log\log|C|)$ that uses additional information. Both results were subsequently improved to bounds that are independent of the size of the concept class C [9, 34].

Our proofs are based on a similar technique of recursively applying Haussler's Packing Lemma on the dual class. This similarity provides another example of the informal connection between sample compression schemes and RT-dimension. This connection also appears in other works that study their relationship with the VC-dimension [9, 12, 34].

[1] In this text $O(f)$ means at most $\alpha f + \beta$ for $\alpha, \beta > 0$ constants.

1.1 VC-Dimension

VC-dimension and size A concept class over the universe X is a set $C \subseteq \{0,1\}^X$. When X is finite, we denote $|X|$ by $n(C)$. The VC-dimension of C, denoted $\mathrm{VC}(C)$, is the maximum size of a shattered subset of X, where a set $Y \subseteq X$ is shattered if for every $Z \subseteq Y$ there is $c \in C$ so that $c(x) = 1$ for all $x \in Z$ and $c(x) = 0$ for all $x \in Y - Z$.

The most basic result concerning VC-dimension is the Sauer–Shelah–Perles Lemma, that upper bounds $|C|$ in terms of $n(C)$ and $\mathrm{VC}(C)$. It has been independently proved several times, e.g. in [41].

Theorem 1.1 (Sauer–Shelah–Perles) *Let C be a boolean concept class with VC-dimension d. Then,*

$$|C| \leq \sum_{k=0}^{d} \binom{n(C)}{k}.$$

In particular, if $d \geq 2$ then $|C| \leq n(C)^d$

VC-dimension and PAC learning The VC-dimension is one of the most basic complexity measures for concept classes. It is perhaps mostly known in the context of the PAC learning model. PAC learning was introduced in Valiant's seminal work [43] as a theoretical model for learning from random examples drawn from an unknown distribution (see the book [28] for more details).

A fundamental and well-known result of Blumer, Ehrenfeucht, Haussler, and Warmuth [8], which is based on an earlier work of Vapnik and Chervonenkis [44], states that PAC learning sample complexity is equivalent to VC-dimension. The proof of this theorem uses Theorem 1.1 and an argument commonly known as double sampling (see section "Appendix: Double Sampling" in the appendix for a short and self contained description of this well known argument).

Theorem 1.2 ([8, 44]) *Let X be a set and $C \subseteq \{0,1\}^X$ be a concept class of VC-dimension d. Let μ be a distribution over X. Let $\epsilon, \delta > 0$ and m an integer satisfying $2(2m+1)^d(1-\epsilon/4)^m < \delta$. Let $c \in C$ and $Y = (x_1, \ldots, x_m)$ be a multiset of m independent samples from μ. Then, the probability that there is $c' \in C$ so that $c|_Y = c'|_Y$ but $\mu(\{x : c(x) \neq c'(x)\}) > \epsilon$ is at most δ.*

VC-dimension and the metric structure Another fundamental result in this area is Haussler's [23] description of the metric structure of concept classes with low VC-dimension (see also the work of Dudley [14]). Roughly, it says that a concept class C of VC-dimension d, when thought of as an L_1 metric space, behaves like a d dimensional space in the sense that the size of an ϵ-separated set in C is at most $(1/\epsilon)^d$. More formally, every probability distribution μ on X induces the (pseudo) metric

$$\mathsf{dist}_\mu(c, c') = \mu(\{x : c(x) \neq c'(x)\})$$

on C. A set $S \subseteq C$ is called ϵ-separated with respect to μ if for every two concepts $c \neq c'$ in S we have $\mathsf{dist}_\mu(c, c') > \epsilon$. A set $A = A_\mu(C, \epsilon) \subseteq C$ is called an *ϵ-approximating set*[2] for C with respect to μ if it is a maximal ϵ-separated set with respect to μ. The maximality of A implies that for every $c \in C$ there is some rounding $r = r(c, \mu, C, \epsilon)$ in A so that r is a good approximation to c, that is, $\mathsf{dist}_\mu(c, r) \leq \epsilon$. We call r a rounding of c in A.

An approximating set can be thought of as a metric approximation of the possibly complicated concept class C, and for many practical purposes it is a good enough substitute for C. Haussler proved that there are always small approximating sets.

Theorem 1.3 (Haussler) *Let $C \subseteq \{0,1\}^X$ be a concept class with VC-dimension d. Let μ be a distribution on X. Let $\epsilon \in (0, 1]$. If S is ϵ-separated with respect to μ then*

$$|S| \leq e(d+1)\left(\frac{2e}{\epsilon}\right)^d \leq \left(\frac{4e^2}{\epsilon}\right)^d.$$

A proof of a weaker statement For $m = 2\log(|S|)/\epsilon$, let $x_1, \ldots, x_m$ be independent samples from μ. For every $c \neq c'$ in S,

$$\Pr_{\mu^m}\left(\forall i \in [m] \;\; c(x_i) = c'(x_i)\right) < (1-\epsilon)^m \leq e^{-m\epsilon} \leq 1/|S|^2.$$

The union bound implies that there is a choice of $Y \subseteq X$ of size $|Y| \leq m$ so that $|S|_Y| = |S|$. Theorem 1.1 implies $|S| \leq (|Y|+1)^d$. Thus, $|S| < (30d \log(2d/\epsilon)/\epsilon)^d$. □

1.2 Teaching

Imagine a teacher that helps a student to learn a concept c by picking insightful examples. The concept c is known only to the teacher, but c belongs to a class of concepts C known to both the teacher and the student. The teacher carefully chooses a set of examples that is tailored for c, and then provides these examples to the student. Now, the student should be able to recover c from these examples.

A central issue that is addressed in the design of mathematical teaching models is "collusions." Roughly speaking, a collusion occurs when the teacher and the student agree in advance on some unnatural encoding of information about c using the bit description of the chosen examples, instead of using attributes that separate c from other concepts. Many mathematical models for teaching were suggested: Shinohara and Miyano [42], Jackson and Tomkins [27], Goldman, Rivest and Schapire [21],

[2] In metric spaces such a set is called an ϵ-net, however in learning theory and combinatorial geometry the term ϵ-net has a different meaning, so we use ϵ-approximating instead.

Goldman and Kearns [19], Goldman and Mathias [20] Angluin and Krikis [2], Balbach [5], and Kobayashi and Shinohara [29]. We now discuss some of these models in more detail.

Teaching sets The first mathematical models for teaching [3, 19, 42] handle collusions in a fairly restrictive way, by requiring that the teacher provides a set of examples Y that uniquely identifies c. Formally, this is captured by the notion of a teaching set, which was independently introduced by Goldman and Kearns [19], Shinohara and Miyano [42] and Anthony et al. [3]. A set $Y \subseteq X$ is a teaching set for c in C if for all $c' \neq c$ in C, we have $c'|_Y \neq c|_Y$. The teaching complexity in these models is captured by the hardest concept to teach, i.e., $\max_{c \in C} \min\{|Y| : Y \text{ is a teaching set for } c \text{ in } C\}$.

Teaching sets also appear in other areas of learning theory: Hanneke [22] used it in his study of the label complexity in active learning, and the authors of [46] used variants of it to design efficient algorithms for learning distributions using imperfect data.

Defining the teaching complexity using the hardest concept is often too restrictive. Consider for example the concept class consisting of all singletons and the empty set over a domain X of size n. Its teaching complexity in these models is n, since the only teaching set for the empty set is X. This is a fairly simple concept class that has the maximum possible complexity.

Recursive teaching dimension Goldman and Mathias [20] and Angluin and Krikis [2] therefore suggested less restrictive teaching models, and more efficient teaching schemes were indeed discovered in these models. One approach, studied by Zilles et al. [47], Doliwa et al. [12], and Samei et al. [39], uses a natural hierarchy on the concept class C which is defined as follows. The first layer in the hierarchy consists of all concepts whose teaching set has minimal size. Then, these concepts are removed and the second layer consists of all concepts whose teaching set with respect to the remaining concepts has minimal size. Then, these concepts are removed and so on, until all concepts are removed. The maximum size of a set that is chosen in this process is called the *recursive teaching (RT)* dimension. One way of thinking about this model is that the teaching process satisfies an Occam's razor-type rule of preferring simpler concepts. For example, the concept class consisting of singletons and the empty set, which was considered earlier, has recursive teaching dimension 1: The first layer in the hierarchy consists of all singletons, which have teaching sets of size 1. Once all singletons are removed, we are left with a concept class of size 1, the concept class $\{\emptyset\}$, and in it the empty set has a teaching set of size 0.

A similar notion to RT-dimension was independently suggested in [46] under the terminology of partial IDs. There the focus was on getting a simultaneous upper bound on the size of the sets, as well as the number of layers in the recursion, and it was shown that for any concept class C both can be made at most $\log |C|$. Motivation for this study comes from the population recovery learning problem defined in [15].

Previous results Doliwa et al. [12] and Zilles et al. [47] asked whether small VC-dimension implies small recursive teaching dimension. An equivalent question was asked 10 years earlier by Kuhlmann [30]. Since the VC-dimension does not increase when concepts are removed from the class, this question is equivalent to asking whether every class with small VC-dimension has some concept in it with a small teaching set. Given the semantics of the recursive teaching dimension and the VC-dimension, an interpretation of this question is whether exact teaching is not much harder than approximate learning (i.e., PAC learning).

For infinite classes the answer to this question is negative. There is an infinite concept class with VC-dimension 1 so that every concept in it does not have a finite teaching set. An example for such a class is $C \subseteq \{0,1\}^{\mathbb{Q}}$ defined as $C = \{c_q : q \in \mathbb{Q}\}$ where c_q is the indicator function of all rational numbers that are smaller than q. The VC-dimension of C is 1, but every teaching set for some $c_q \in C$ must contain a sequence of rationals that converges to q.

For finite classes this question is open. However, in some special cases it is known that the answer is affirmative. In [30] it is shown that if C has VC-dimension 1, then its recursive teaching dimension is also 1. It is known that if C is a maximum[3] class then its recursive teaching dimension is equal to its VC-dimension [12, 38]. Other families of concept classes for which the recursive teaching dimension is at most the VC-dimension are discussed in [12]. In the other direction, [30] provided examples of concept classes with VC-dimension d and recursive teaching dimension at least $\frac{3}{2}d$.

The only bound on the recursive teaching dimension for general classes was observed by both [12, 46]. It states that the recursive teaching dimension of C is at most $\log|C|$. This bound follows from a simple halving argument which shows that for all C there exists some $c \in C$ with a teaching set of size $\log|C|$.

Our contribution Our first main result is the following general bound, which exponentially improves over the $\log|C|$ bound when the VC-dimension is small (the proof is given in Sect. 3).

Theorem 1.4 (RT-dimension) *Let C be a concept class of VC-dimension d. Then there exists $c \in C$ with a teaching set of size at most*

$$d2^{d+3}(\log(4e^2) + \log\log|C|).$$

It follows that the recursive teaching dimension of concept classes of VC-dimension d is at most $d2^{d+3}(\log(4e^2) + \log\log|C|)$ as well.

Subsequent to this paper, Chen, Cheng, and Tang [9] proved that the RT-dimension is at most $\exp(d)$. Their proof is based on ideas from this work, in particular they follow and improve the argument from the proof of Lemma 1.7.

[3] That is, C satisfies Sauer–Shelah–Perles Lemma with equality.

1.3 Sample Compression Schemes

A fundamental and well known statement in learning theory says that if the VC-dimension of a concept class C is small, then any consistent[4] algorithm successfully PAC learns concepts from C after seeing just a few labelled examples [7, 44]. In practice, however, a major challenge one has to face when designing a learning algorithm is the construction of an hypothesis that is consistent with the examples seen. Many learning algorithms share the property that the output hypothesis is constructed using a small subset of the examples. For example, in support vector machines, only the set of support vectors is needed to construct the separating hyperplane [11]. Sample compression schemes provide a formal meaning for this algorithmic property.

Before giving the formal definition of compression schemes, let us consider a simple illustrative example. Assume we are interested in learning the concept class of intervals on the real line. We get a collection of 100 samples of the form $(x, c_I(x))$ where $x \in \mathbb{R}$ and $c_I(x) \in \{0, 1\}$ indicates[5] if x is in the interval $I \subset \mathbb{R}$. Can we remember just a few of the samples in a way that allows to recover all the 100 samples? In this case, the answer is affirmative and in fact it is easy to do so. Just remember two locations, those of the left most 1 and of the right most 1 (if there are no 1s, just remember one of the 0s). From this data, we can reconstruct the value of c_I on all the other 100 samples.

The formal definition Littlestone and Warmuth (Relating data compression and learnability. Unpublished, 1986) formally defined sample compression schemes as follows. Let $C \subseteq \{0, 1\}^X$ with $|X| = n$. Let

$$L_C(k_1, k_2) = \{(Y, y) : Y \subseteq X,\ k_1 \leq |Y| \leq k_2,\ y \in C|_Y\},$$

the set of labelled samples from C, of sizes between k_1 and k_2. A k-sample compression scheme for C with information Q, consists of two maps κ, ρ for which the following hold:

(κ) The *compression map*

$$\kappa : L_C(1, n) \to L_C(0, k) \times Q$$

takes (Y, y) to $((Z, z), q)$ with $Z \subseteq Y$ and $y|_Z = z$.

(ρ) The *reconstruction map*

$$\rho : L_C(0, k) \times Q \to \{0, 1\}^X$$

[4] An algorithm that outputs an hypothesis in C that is consistent with the input examples.

[5] That is $c_I(x) = 1$ iff $x \in I$.

is so that for all (Y, y) in $L_C(1, n)$,

$$\rho(\kappa(Y, y))|_Y = y.$$

The *size* of the scheme is $k + \log |Q|$.

Intuitively, the compression map takes a long list of samples (Y, y) and encodes it as a short sub-list of samples (Z, z) together with some small amount of side information $q \in Q$, which helps in the reconstruction phase. The reconstruction takes a short list of samples (Z, z) and decodes it using the side information q, without any knowledge of (Y, y), to an hypothesis in a way that essentially inverts the compression. Specifically, the following property must always hold: if the compression of $(Y, c|_Y)$ is the same as that of $(Y', c'|_{Y'})$ then $c|_{Y \cap Y'} = c'|_{Y \cap Y'}$.

A different perspective of the side information is as a list decoding in which the small set of labelled examples (Z, z) is mapped to the set of hypothesis $\{\rho((Z, z), q) : q \in Q\}$, one of which is correct.

We note that it is not necessarily the case that the reconstructed hypothesis belongs to the original class C. All it has to satisfy is that for any $(Y, y) \in L_C(1, n)$ such that $h = \rho(\kappa(Y, y))$ we have that $h|_Y = y$. Thus, h has to be consistent only on the sampled coordinates that were compressed and not elsewhere.

Let us consider a simple example of a sample compression scheme, to help digest the definition. Let C be a concept class and let r be the rank over, say, $\mathbb{R}$ of the matrix whose rows correspond to the concepts in C. We claim that there is an r-sample compression scheme for C with no side information. Indeed, for any $Y \subseteq X$, let Z_Y be a set of at most r columns that span the columns of the matrix $C|_Y$. Given a sample (Y, y) compress it to $\kappa(Y, y) = (Z_Y, z)$ for $z = y|_{Z_Y}$. The reconstruction maps ρ takes (Z, z) to any concept $h \in C$ so that $h|_Z = z$. This sample compression scheme works since if $(Z, z) = \kappa(Y, y)$ then every two different rows in $C|_Y$ must disagree on Z.

Connections to learning Sample compression schemes are known to yield practical learning algorithms (see e.g. [33]), and allow learning for multi labelled concept classes [40].

They can also be interpreted as a formal manifestation of Occam's razor. Occam's razor is a philosophical principle attributed to William of Ockham from the late middle ages. It says that in the quest for an explanation or an hypothesis, one should prefer the simplest one which is consistent with the data. There are many works on the role of Occam's razor in learning theory, a partial list includes (Littlestone and Warmuth, Relating data compression and learnability. Unpublished, 1986) [7, 13, 16, 17, 26, 36]. In the context of sample compression schemes, simplicity is captured by the size of the compression scheme. Interestingly, this manifestation of Occam's razor is provably useful (Littlestone and Warmuth, Relating data compression and learnability. Unpublished, 1986): Sample compression schemes imply PAC learnability.

Theorem 1.5 (Littlestone–Warmuth) *Let $C \subseteq \{0,1\}^X$, and $c \in C$. Let μ be a distribution on X, and $x_1, \ldots, x_m$ be m independent samples from μ. Let $Y = (x_1, \ldots, x_m)$ and $y = c|_Y$. Let κ, ρ be a k-sample compression scheme for C with additional information Q. Let $h = \rho(\kappa(Y, y))$. Then,*

$$\Pr_{\mu^m}(\mathsf{dist}_\mu(h, c) > \epsilon) < |Q| \sum_{j=0}^{k} \binom{m}{j} (1-\epsilon)^{m-j}.$$

Proof sketch There are $\sum_{j=0}^{k} \binom{m}{j}$ subsets T of $[m]$ of size at most k. There are $|Q|$ choices for $q \in Q$. Each choice of T, q yields a function $h_{T,q} = \rho((T, y_T), q)$ that is measurable with respect to $x_T = (x_t : t \in T)$. The function h is one of the functions in $\{h_{T,q} : |T| \leq k, q \in Q\}$. For each $h_{T,q}$, the coordinates in $[m] - T$ are independent, and so if $\mathsf{dist}_\mu(h_{T,q}, c) > \epsilon$ then the probability that all these $m - |T|$ samples agree with c is less than $(1-\epsilon)^{m-|T|}$. The union bound completes the proof. □

The sample complexity of PAC learning is essentially the VC-dimension. Thus, from Theorem 1.5 we expect the VC-dimension to bound from below the size of sample compression schemes. Indeed, [17] proved that there are concept classes of VC-dimension d for which any sample compression scheme has size at least d.

This is part of the motivation for the following basic question that was asked by Littlestone and Warmuth (Relating data compression and learnability. Unpublished, 1986) nearly 30 years ago: Does a concept class of VC-dimension d have a sample compression scheme of size depending only on d (and not on the universe size)?

In fact, unlike the VC-dimension, the definition of sample compression schemes as well as the fact that they imply PAC learnability naturally generalizes to multi-class classification problems [40]. Thus, Littlestone and Warmuth's question above can be seen as the boolean instance of a much broader question: Is it true that the size of an optimal sample compression scheme for a given concept class (not necessarily binary-labeled) is the sample complexity of PAC learning of this class?

Previous constructions Floyd [16] and Floyd and Warmuth [17] constructed sample compression schemes of size $\log |C|$. The construction in [17] uses a transformation that converts certain online learning algorithms to compression schemes. Helmbold and Warmuth [26] and Freund [18] showed how to compress a sample of size m to a sample of size $O(\log(m))$ using some side information for classes of constant VC-dimension (the implicit constant in the $O(\cdot)$ depends on the VC-dimension).

In a long line of works, several interesting compression schemes for special cases were constructed. A partial list includes Helmbold et al. [25], Floyd and Warmuth [17], Ben-David and Litman [6], Chernikov and Simon [10], Kuzmin and Warmuth [31], Rubinstein et al. [37], Rubinstein and Rubinstein [38], Livni and Simon [32] and more. These works provided connections between compression schemes and geometry, topology and model theory.

Our contribution Here we make the first quantitive progress on this question, since the work of Floyd [16]. The following theorem shows that low VC-dimension implies the existence of relatively efficient compression schemes. The constructive proof is provided in Sect. 4.

Theorem 1.6 (Sample compression scheme) *If C has VC-dimension d then it has a k-sample compression scheme with additional information Q where $k = O(d2^d \log\log|C|)$ and $\log|Q| \leq O(k\log(k))$.*

Subsequent to this paper, the first and the last authors improved this bound [34], showing that any concept class of VC-dimension d has a sample compression scheme of size at most $\exp(d)$. The techniques used in [34] differ from the techniques we use in this paper. In particular, our scheme relies on Haussler's Packing Lemma (Theorem 1.3) and recursion, while the scheme in [34] relies on von Neumann's minimax theorem [35] and the ϵ-approximation theorem [24, 44], which follow from the double-sampling argument of [44]. Thus, despite the fact that our scheme is weaker than the one in [34], it provides a different angle on sample compression, which may be useful in further improving the exponential dependence on the VC-dimension to an optimal linear dependence, as conjectured by Floyd and Warmuth [17, 45].

1.4 Discussion and Open Problems

This work provides relatively efficient constructions of teaching sets and sample compression schemes. However, the exact relationship between VC-dimension, sample compression scheme size, and the RT-dimension remains unknown. Is there always a concept with a teaching set of size depending only on the VC-dimension? (The interesting case is finite concept classes, as mentioned above.) Are there always sample compression schemes of size linear (or even polynomial) in the VC-dimension?

The simplest case that is still open is VC-dimension 2. One can refine this case even further. VC-dimension 2 means that on any three coordinates $x, y, z \in X$, the projection $C|_{\{x,y,z\}}$ has at most 7 patterns. A more restricted family of classes is $(3, 6)$ concept classes, for which on any three coordinates there are at most 6 patterns. We can show that the recursive teaching dimension of $(3, 6)$ classes is at most 3.

Lemma 1.7 *Let C be a finite $(3, 6)$ concept class. Then there exists some $c \in C$ with a teaching set of size at most 3.*

Proof Assume that $C \subseteq \{0, 1\}^X$ with $X = [n]$. If C has VC-dimension 1 then there exists $c \in C$ with a teaching set of size 1 (see [1, 30]). Therefore, assume that the VC-dimension of C is 2. Every shattered pair $\{x, x'\} \subseteq X$ partitions C to 4 nonempty sets:

$$C^{x,x'}_{b,b'} = \{c \in C : c(x) = b, c(x') = b'\},$$

for $b, b' \in \{0,1\}$. Pick a shattered pair $\{x_*, x'_*\}$ and b_*, b'_* for which the size of $C^{x_*,x'_*}_{b_*,b'_*}$ is minimal. Without loss of generality assume that $\{x_*, x'_*\} = \{1,2\}$ and that $b_* = b'_* = 0$. To simplify notation, we denote $C^{1,2}_{b,b'}$ simply by $C_{b,b'}$.

We prove below that $C_{0,0}$ has VC-dimension 1. This completes the proof since then there is some $c \in C_{0,0}$ and some $x \in [n] \setminus \{1,2\}$ such that $\{x\}$ is a teaching set for c in $C_{0,0}$. Therefore, $\{1,2,x\}$ is a teaching set for c in C.

First, a crucial observation is that since C is a $(3,6)$ class, no pair $\{x,x'\} \subseteq [n] \setminus \{1,2\}$ is shattered by both $C_{0,0}$ and $C \setminus C_{0,0}$. Indeed, if $C \setminus C_{0,0}$ shatters $\{x,x'\}$ then either $C_{1,0} \cup C_{1,1}$ or $C_{0,1} \cup C_{1,1}$ has at least 3 patterns on $\{x,x'\}$. If in addition $C_{0,0}$ shatters $\{x,x'\}$ then C has at least 7 patterns on $\{1,x,x'\}$ or $\{2,x,x'\}$, contradicting the assumption that C is a $(3,6)$ class.

Now, assume towards contradiction that $C_{0,0}$ shatters $\{x,x'\}$. Thus, $\{x,x'\}$ is not shattered by $C \setminus C_{0,0}$ which means that there is some pattern $p \in \{0,1\}^{\{x,x'\}}$ so that $p \notin (C \setminus C_{0,0})|_{\{x,x'\}}$. This implies that $C^{x,x'}_{p(x),p(x')}$ is a proper subset of $C_{0,0}$, contradicting the minimality of $C_{0,0}$. □

2 The Dual Class

We shall repeatedly use the dual concept class to C and its properties. The dual concept class $C^* \subseteq \{0,1\}^C$ of C is defined by $C^* = \{c_x : x \in X\}$, where $c_x : C \to \{0,1\}$ is the map so that $c_x(c) = 1$ iff $c(x) = 1$. If we think of C as a binary matrix whose rows are the concepts in C, then C^* corresponds to the distinct rows of the transposed matrix (so it may be that $|C^*| < |n(C)|$).

We use the following well known property (see [4]).

Claim 2.1 (Assouad) *If the VC-dimension of C is d then the VC-dimension of C^* is at most 2^{d+1}.*

Proof sketch If the VC-dimension of C^* is 2^{d+1} then in the matrix representing C there are 2^{d+1} rows that are shattered, and in these rows there are $d+1$ columns that are shattered. □

We also define the dual approximating set (recall the definition of $A_\mu(C,\epsilon)$ from Sect. 1.1). Denote by $A^*(C,\epsilon)$ the set $A_U(C^*,\epsilon)$, where U is the uniform distribution on C^*.

3 Teaching Sets

In this section we prove Theorem 1.4. The high level idea is to use Theorem 1.3 and Claim 2.1 to identify two distinct x, x' in X so that the set of $c \in C$ so that $c(x) \neq c(x')$ is much smaller than $|C|$, add x, x' to the teaching set, and continue inductively.

Proof of Theorem 1.4 For classes with VC-dimension 1 there is $c \in C$ with a teaching set of size 1, see e.g. [12]. We may therefore assume that $d \geq 2$.

We show that if $|C| > (4e^2)^{d \cdot 2^{d+2}}$, then there exist $x \neq x'$ in X such that

$$0 < |\{c \in C : c(x) = 0 \text{ and } c(x') = 1\}| \leq |C|^{1-\frac{1}{d2^{d+2}}}. \tag{1}$$

From this the theorem follows, since if we iteratively add such x, x' to the teaching set and restrict ourselves to $\{c \in C : c(x) = 0 \text{ and } c(x') = 1\}$, then after at most $d2^{d+2} \log\log|C|$ iterations, the size of the remaining class is reduced to less than $(4e^2)^{d \cdot 2^{d+2}}$. At this point we can identify a unique concept by adding at most $\log((4e^2)^{d \cdot 2^{d+2}})$ additional indices to the teaching set, using the halving argument of [12, 46]. This gives a teaching set of size at most $2d2^{d+2}\log\log|C| + d2^{d+2}\log(4e^2)$ for some $c \in C$, as required.

In order to prove (1), it is enough to show that there exist $c_x \neq c_y$ in C^* such that the normalized hamming distance between c_x, c_y is at most $\epsilon := |C|^{-\frac{1}{d2^{d+2}}}$. Assume towards contradiction that the distance between every two concepts in C^* is more than ϵ, and assume without loss of generality that $n(C) = |C^*|$ (that is, all the columns in C are distinct). By Claim 2.1, the VC-dimension of C^* is at most 2^{d+1}. Theorem 1.3 thus implies that

$$n(C) = |C^*| \leq \left(\frac{4e^2}{\epsilon}\right)^{2^{d+1}} < \left(\frac{1}{\epsilon}\right)^{2^{d+2}}, \tag{2}$$

where the last inequality follows from the definition of ϵ and the assumption on the size of C. Therefore, we arrive at the following contradiction:

$$\begin{aligned} |C| &\leq (n(C))^d && \text{(by Theorem 1.1, since } VC(C) \geq 2\text{)} \\ &< \left(\frac{1}{\epsilon}\right)^{d \cdot 2^{d+2}} && \text{(by Equation 2 above)} \\ &= |C|. && \text{(by definition of } \epsilon\text{)} \end{aligned}$$

□

4 Sample Compression Schemes

In this section we prove Theorem 1.6. The theorem statement and the definition of sample compression schemes appear in Sect. 1.3.

While the details are somewhat involved, due to the complexity of the definitions, the high level idea may be (somewhat simplistically) summarized as follows.

For an appropriate choice of ϵ, we pick an ϵ-approximating set A^* of the dual class C^*. It is helpful to think of A^* as a subset of the domain X. Now, either A^* faithfully represents the sample (Y, y) or it does not (we do not formally define "faithfully represents" here). We identify the following win-win situation: In both cases, we can reduce the compression task to that in a much smaller set of concepts of size at most $\epsilon|C| \approx |C|^{1-2^{-d}}$, similarly to as for teaching sets in Sect. 3. This yields the same double-logarithmic behavior.

In the case that A^* faithfully represents (Y, y), Case 2 below, we recursively compress in the small class $C|_{A^*}$. In the unfaithful case, Case 1 below, we recursively compress in a (small) set of concepts for which disagreement occurs on some point of Y, just as in Sect. 3. In both cases, we have to extend the recursive solution, and the cost is adding one sample point to the compressed sample (and some small amount of additional information by which we encode whether Case 1 or 2 occurred).

The compression we describe is inductively defined, and has the following additional structure. Let $((Z, z), q)$ be in the image of κ. The information q is of the form $q = (f, T)$, where $T \geq 0$ is an integer so that $|Z| \leq T + O(d \cdot 2^d)$, and $f : \{0, 1, \ldots, T\} \to Z$ is a partial one-to-one function.[6]

The rest of this section is organized as follows. In Sect. 4.1 we define the compression map κ. In Sect. 4.2 we give the reconstruction map ρ. The proof of correctness is given in Sect. 4.3 and the upper bound on the size of the compression is calculated in Sect. 4.4.

4.1 Compression Map: Defining κ

Let C be a concept class. The compression map is defined by induction on $n = n(C)$. For simplicity of notation, let $d = VC(C) + 2$.

In what follows we shall routinely use $A^*(C, \epsilon)$. There are several ϵ-approximating sets and so we would like to fix one of them, say, the one obtained by greedily adding columns to $A^*(C, \epsilon)$ starting from the first[7] column (recall that we can think of C as a matrix whose rows correspond to concepts in C and whose columns are concepts in the dual class C^*). To keep notation simple, we shall use $A^*(C, \epsilon)$ to denote both the approximating set in C^* and the subset of X composed of columns that give rise to $A^*(C, \epsilon)$. This is a slight abuse of notation but the relevant meaning will always be clear from the context.

Induction base The base of the induction applies to all concept classes C so that $|C| \leq (4e^2)^{d \cdot 2^d + 1}$. In this case, we use the compression scheme of Floyd and Warmuth [16, 17] which has size $\log(|C|) = O(d \cdot 2^d)$. This compression scheme has

[6] That is, it is defined over a subset of $\{0, 1, \ldots, T\}$ and it is injective on its domain.

[7] We shall assume w.l.o.g. that there is some well known order on X.

no additional information. Therefore, to maintain the structure of our compression scheme we append to it redundant additional information by setting $T = 0$ and f to be empty.

Induction step Let C be so that $|C| > (4e^2)^{d\cdot 2^d+1}$. Let $0 < \epsilon < 1$ be so that

$$\epsilon|C| = \left(\frac{1}{\epsilon}\right)^{d\cdot 2^d}. \tag{3}$$

This choice balances the recursive size. By Claim 2.1, the VC-dimension of C^* is at most 2^{d-1} (recall that $d = VC(C) + 2$). Theorem 1.3 thus implies that

$$|A^*(C,\epsilon)| \leq \left(\frac{4e^2}{\epsilon}\right)^{2^{d-1}} < \left(\frac{1}{\epsilon}\right)^{2^d} < n(C). \tag{4}$$

(Where the second inequality follows from the definition of ϵ and the assumption on the size of C and the last inequality follows from the definition of ϵ and Theorem 1.1.)

Let $(Y, y) \in L_C(1, n)$. Every $x \in X$ has a rounding[8] $r(x)$ in $A^*(C, \epsilon)$. We distinguish between two cases:

Case 1: There exist $x \in Y$ and $c \in C$ such that $c|_Y = y$ and $c(r(x)) \neq c(x)$. This is the unfaithful case in which we recurse as in Sect. 3. Let

$$C' = \{c'|_{X-\{x,r(x)\}} : c' \in C, c'(x) = c(x), c'(r(x)) = c(r(x))\},$$
$$Y' = Y - \{x, r(x)\},$$
$$y' = y|_{Y'}.$$

Apply recursively κ on C' and the sample $(Y', y') \in L_{C'}(1, n(C'))$. Let $((Z', z'), (f', T'))$ be the result of this compression. Output $((Z, z), (f, T))$ defined as[9]

$$Z = Z' \cup \{x\},$$
$$z|_{Z'} = z',\ z(x) = y(x),$$
$$T = T' + 1,$$
$$f|_{\{0,\ldots,T-1\}} = f'|_{\{0,\ldots,T-1\}},$$
$$f(T) = x \qquad (f \text{ is defined on } T\text{, marking that Case 1 occurred})$$

[8] The choice of $r(x)$ also depends on C, ϵ, but to simplify the notation we do not explicitly mention it.

[9] Remember that f is a partial function.

Case 2: For all $x \in Y$ and $c \in C$ such that $c|_Y = y$, we have $c(x) = c(r(x))$. This is the faithful case, in which we compress by restricting C to A^*. Consider $r(Y) = \{r(y') : y' \in Y\} \subseteq A^*(C, \epsilon)$. For each $x' \in r(Y)$, pick[10] $s(x') \in Y$ to be an element such that $r(s(x')) = x'$. Let

$$\begin{aligned} C' &= C|_{A^*(C,\epsilon)}, \\ Y' &= r(Y), \\ y'(x') &= y(s(x'))\ \forall x' \in Y'. \end{aligned}$$

By (4), we know $|A^*(C, \epsilon)| < n(C)$. Therefore, we can recursively apply κ on C' and $(Y', y') \in L_{C'}(1, n(C'))$ and get $((Z', z'), (f', T'))$. Output $((Z, z), (f, T))$ defined as

$$\begin{aligned} &Z = \{s(x') : x' \in Z'\}, \\ &z(x) = z'(r(x))\ \forall x \in Z, && (r(x) \in Z') \\ &T = T' + 1, \\ &f = f'. && (f \text{ is not defined on } T, \text{ marking that Case 2 occurred}) \end{aligned}$$

The following lemma summarizes two key properties of the compression scheme. The correctness of this lemma follows directly from the definitions of Cases 1 and 2 above.

Lemma 4.1 *Let $(Y, y) \in L_C(1, n(C))$ and $((Z, z), (T, f))$ be the compression of (Y, y) described above, where $T \geq 1$. The following properties hold:*

1. *f is defined on T and $f(T) = x$ iff $x \in Y$ and there exists $c \in C$ such that $c|_Y = y$ and $c(r(x)) \neq c(x)$.*
2. *f is not defined on T iff for all $x \in Y$ and $c \in C$ such that $c|_Y = y$, it holds that $c(x) = c(r(x))$.*

4.2 Reconstruction Map: Defining ρ

The reconstruction map is similarly defined by induction on $n(C)$. Let C be a concept class and let $((Z, z), (f, T))$ be in the image[11] of κ with respect to C. Let $\epsilon = \epsilon(C)$ be as in (3).

[10] The function s can be thought of as the inverse of r. Since r is not necessarily invertible we use a different notation than r^{-1}.

[11] For $((Z, z), (f, T))$ not in the image of κ we set $\rho((Z, z), (f, T))$ to be some arbitrary concept.

Induction base The induction base here applies to the same classes like the induction base of the compression map. This is the only case where $T = 0$, and we apply the reconstruction map of Floyd and Warmuth [16, 17]

Induction step Distinguish between two cases:

Case 1: f is defined on T.
Let $x = f(T)$. Denote

$$\begin{aligned}
X' &= X - \{x, r(x)\},\\
C' &= \{c'|_{X'} : c' \in C, c'(x) = z(x), c'(r(x)) = 1 - z(x)\},\\
Z' &= Z - \{x, r(x)\},\\
z' &= z|_{Z'},\\
T' &= T - 1,\\
f' &= f|_{\{0,\ldots,T'\}}.
\end{aligned}$$

Apply recursively ρ on $C', ((Z', z'), (f', T'))$. Let $h' \in \{0, 1\}^{X'}$ be the result. Output h where

$$\begin{aligned}
h|_{X'} &= h',\\
h(x) &= z(x),\\
h(r(x)) &= 1 - z(x).
\end{aligned}$$

Case 2: f is not defined on T.
Consider $r(Z) = \{r(x) : x \in Z\} \subseteq A^*(C, \epsilon)$. For each $x' \in r(Z)$, pick $s(x') \in Z$ to be an element such that $r(s(x')) = x'$. Let

$$\begin{aligned}
X' &= A^*(C, \epsilon),\\
C' &= C|_{X'},\\
Z' &= r(Z),\\
z'(x') &= z(s(x')) \ \forall x' \in Z',\\
T' &= T - 1,\\
f' &= f|_{\{0,\ldots,T'\}}.
\end{aligned}$$

Apply recursively ρ on $C', ((Z', z'), (f', T'))$ and let $h' \in \{0, 1\}^{X'}$ be the result. Output h satisfying

$$h(x) = h'(r(x)) \ \forall x \in X.$$

4.3 Correctness

The following lemma yields the correctness of the compression scheme.

Lemma 4.2 *Let C be a concept class, $(Y, y) \in L_C(1, n)$, $\kappa(Y, y) = ((Z, z), (f, T))$ and $h = \rho(\kappa(Y, y))$. Then,*

1. *$Z \subseteq Y$ and $z|_Z = y|_Z$, and*
2. *$h|_Y = y|_Y$.*

Proof We proceed by induction on $n(C)$. In the base case, $|C| \leq (4e^2)^{d\cdot 2^d+1}$ and the lemma follows from the correctness of Floyd and Warmuth's compression scheme (this is the only case in which $T = 0$). In the induction step, assume $|C| > (4e^2)^{d\cdot 2^d+1}$. We distinguish between two cases:

Case 1: f is defined on T.

Let $x = f(T)$. This case corresponds to Case 1 in the definitions of κ and Case 1 in the definition of ρ. By Item 1 of Lemma 4.1, $x \in Y$ and there exists $c \in C$ and $x \in Y$ such that $c|_Y = y$ and $c(r(x)) \neq c(x)$. Let C', (Y', y') be the class defined in Case 1 in the definition of κ. Since $n(C') < n(C)$, we know that κ, ρ on C' satisfy the induction hypothesis. Let

$$((Z', z'), (f', T')) = \kappa(C', (Y', y')),$$
$$h' = \rho(C', ((Z', z'), (f', T'))),$$

be the resulting compression and reconstruction. Since we are in Case 1 in the definitions of κ and Case 1 in the definition of ρ, $((Z, z), (f, T))$ and h have the following form:

$$Z = Z' \cup \{x\},$$
$$z|_{Z'} = z', \ z(x) = y(x),$$
$$T = T' + 1,$$
$$f|_{\{0,\ldots,T-1\}} = f'|_{\{0,\ldots,T-1\}},$$
$$f(T) = x,$$

and

$$h|_{X-\{x,r(x)\}} = h',$$
$$h(x) = z(x) = y(x) = c(x),$$
$$h(r(x)) = 1 - z(x) = 1 - y(x) = 1 - c(x) = c(r(x)).$$

Consider item 1 in the conclusion of the lemma. By the definition of Y' and x,

$$\begin{aligned} Y' \cup \{x\} &\subseteq Y, && \text{(by the definition of } Y'\text{)} \\ Z' &\subseteq Y'. && \text{(by the induction hypothesis)} \end{aligned}$$

Therefore, $Z = Z' \cup \{x\} \subseteq Y$.
Consider item 2 in the conclusion of the lemma. By construction and induction,

$$h|_{Y\cap\{x,r(x)\}} = c|_{Y\cap\{x,r(x)\}} = y|_{Y\cap\{x,r(x)\}} \text{ and } h|_{Y'} = h'|_{Y'} = y'.$$

Thus, $h|_Y = y$.

Case 2: f is not defined on T.
This corresponds to Case 2 in the definitions of κ and Case 2 in the definition of ρ. Let $C', (Y', y')$ be the result of Case 2 in the definition of κ. Since $n(C') < n(C)$, we know that κ, ρ on C' satisfy the induction hypothesis. Let

$$\begin{aligned} &((Z', z'), (f', T')) = \kappa(C', (Y', y')), \\ &h' = \rho(C', ((Z', z'), (f', T'))), \\ &s : Y' \to Y, \end{aligned}$$

as defined in Case 2 in the definitions of κ and Case 2 in the definition of ρ. By construction, $((Z, z), (f, T))$ and h have the following form:

$$\begin{aligned} &Z = \{s(x') : x' \in Z'\}, \\ &z(x) = z'(r(x)) \ \forall x \in Z, \\ &T = T' + 1, \\ &f = f', \end{aligned}$$

and

$$h(x) = h'(r(x)) \ \forall x \in X.$$

Consider item 1 in the conclusion of the lemma. Let $x \in Z$. By the induction hypothesis, $Z' \subseteq Y'$. Thus, $x = s(x')$ for some $x' \in Z' \subseteq Y'$. Since the range of s is Y, it follows that $x \in Y$. This shows that $Z \subseteq Y$.

Consider item 2 in the conclusion of the lemma. For $x \in Y$,

$$\begin{aligned} h(x) &= h'(r(x)) && \text{(by the definition of } h\text{)} \\ &= y'(r(x)) && \text{(by the induction hypothesis)} \\ &= y(s(r(x))) && \text{(by the definition of } y' \text{ in Case 2 of } \kappa\text{)} \\ &= y(x), \end{aligned}$$

where the last equality holds due to item 2 of Lemma 4.1: Indeed, let $c \in C$ be so that $c|_Y = y$. Since f is not defined on T, for all $x \in Y$ we have $c(x) = c(r(x))$. In addition, for all $x \in Y$ it holds that $r(s(r(x))) = r(x)$ and $s(r(x)) \in Y$. Hence, if $y(s(r(x))) \neq y(x)$ then one of them is different than $c(r(x))$, contradicting the assumption that we are in Case 2 of κ. □

4.4 The Compression Size

Consider a concept class C which is not part of the induction base (i.e. $|C| > (4e^2)^{d\cdot 2^d+1}$). Let $\epsilon = \epsilon(C)$ be as in (3). We show the effect of each case in the definition of κ on either $|C|$ or $n(C)$:

1. Case 1 in the definition of κ: Here the size of C' becomes smaller

   $$|C'| \leq \epsilon |C|.$$

   Indeed, this holds as in the dual set system C^*, the normalized hamming distance between c_x and $c_{r(x)}$ is at most ϵ and therefore the number of $c \in C$ such that $c(x) \neq c(r(x))$ is at most $\epsilon|C|$.
2. Case 2 in the definition of κ: here $n(C')$ becomes smaller as

   $$n(C') = |A^*(C, \epsilon)| \leq \left(\frac{1}{\epsilon}\right)^{2^d}.$$

We now show that in either cases, $|C'| \leq |C|^{1-\frac{1}{d\cdot 2^d+1}}$, which implies that after

$$O((d \cdot 2^d + 1) \log\log |C|)$$

iterations, we reach the induction base.
In Case 1:

$$|C'| \leq \epsilon |C| = |C|^{1-\frac{1}{d\cdot 2^d+1}}. \qquad \text{(by the definition of } \epsilon\text{)}$$

In Case 2:

$$\begin{aligned}
|C'| &\leq (n(C'))^d && \text{(by Theorem 1.1, since } VC(C') \leq d-2\text{)}\\
&\leq \left(\frac{1}{\epsilon}\right)^{d\cdot 2^d} && \text{(by Theorem 1.3, since } n(C') = |A^*(C,\epsilon)|\text{)}\\
&= |C|^{1-\frac{1}{d\cdot 2^d+1}}. && \text{(by definition of } \epsilon\text{)}
\end{aligned}$$

Remark Note the similarity between the analysis of the cases above, and the analysis of the size of a teaching set in Sect. 3. Case 1 corresponds to the rate of the progress performed in each iteration of the construction of a teaching set. Case 2 corresponds to the calculation showing that in each iteration significant progress can be made.

Thus, the compression map κ performs at most

$$O((d \cdot 2^d + 1) \log\log |C|)$$

iterations. In every step of the recursion the sizes of Z and T increase by at most 1. In the base of the recursion, T is 0 and the size of Z is at most $O(d \cdot 2^d)$. Hence, the total size of the compression satisfies

$$\begin{aligned}
|Z| &\leq k = O(2^d d \log\log |C|),\\
\log(|Q|) &\leq O(k \log(k)).
\end{aligned}$$

This completes the proof of Theorem 1.6.

Acknowledgements We thank Noga Alon and Gillat Kol for helpful discussions in various stages of this work.

Appendix: Double Sampling

Here we provide our version of the double sampling argument from [8] that upper bounds the sample complexity of PAC learning for classes of constant VC-dimension. We use the following simple general lemma.

Lemma A.1 *Let* $(\Omega, \mathcal{F}, \mu)$ *and* $(\Omega', \mathcal{F}', \mu')$ *be countable*[12] *probability spaces. Let*

$$F_1, F_2, F_3, \ldots \in \mathcal{F},\ F'_1, F'_2, F'_3, \ldots \in \mathcal{F}'$$

[12] A similar statement holds in general.

be so that $\mu'(F_i') \geq 1/2$ *for all i. Then*

$$\mu \times \mu'\left(\bigcup_i F_i \times F_i'\right) \geq \frac{1}{2}\mu\left(\bigcup_i F_i\right),$$

where $\mu \times \mu'$ *is the product measure.*

Proof Let $F = \bigcup_i F_i$. For every $\omega \in F$, let $F'(\omega) = \bigcup_{i:\omega \in F_i} F_i'$. As there exists i such that $\omega \in F_i$ it holds that $F_i' \subseteq F'(\omega)$ and hence $\mu'(F'(\omega)) \geq 1/2$. Thus,

$$\mu \times \mu'\left(\bigcup_i F_i \times F_i'\right) = \sum_{\omega \in F} \mu(\{\omega\}) \cdot \mu'(F'(\omega)) \geq \sum_{\omega \in F} \mu(\{\omega\})/2 = \mu(F)/2.$$

□

We now give a proof of Theorem 1.2. To ease the reading we repeat the statement of the theorem.

Theorem *Let X be a set and* $C \subseteq \{0,1\}^X$ *be a concept class of VC-dimension d. Let* μ *be a distribution over X. Let* $\epsilon, \delta > 0$ *and m an integer satisfying* $2(2m+1)^d(1-\epsilon/4)^m < \delta$. *Let* $c \in C$ *and* $Y = (x_1, \ldots, x_m)$ *be a multiset of m independent samples from* μ. *Then, the probability that there is* $c' \in C$ *so that* $c|_Y = c'|_Y$ *but* $\mu(\{x : c(x) \neq c'(x)\}) > \epsilon$ *is at most* δ.

Proof of Theorem 1.2 Let $Y' = (x_1', \ldots, x_m')$ be another m independent samples from μ, chosen independently of Y. Let

$$H = \{h \in C : \mathsf{dist}_\mu(h,c) > \epsilon\}.$$

For $h \in C$, define the event

$$F_h = \{Y : c|_Y = h|_Y\},$$

and let $F = \bigcup_{h \in H} F_h$. Our goal is thus to upper bound $\Pr(F)$. For that, we also define the independent event

$$F_h' = \{Y' : \mathsf{dist}_{Y'}(h,c) > \epsilon/2\}.$$

We first claim that $\Pr(F_h') \geq 1/2$ for all $h \in H$. This follows from Chernoff's bound, but even Chebyshev's inequality suffices: For every $i \in [m]$, let V_i be the indicator variables of the event $h(x_i') \neq c(x_i')$ (i.e., $V_i = 1$ if and only if $h(x_i') \neq c(x_i')$). The event F_h' is equivalent to $V = \sum_i V_i/m > \epsilon/2$. Since $h \in H$, we have $p := \mathbb{E}[V] > \epsilon$. Since elements of Y' are chosen independently, it follows that $\mathrm{Var}(V) = p(1-p)/m$. Thus, the probability of the complement of F_h' satisfies

$$\Pr((F_h')^c) \leq \Pr(|V-p| \geq p - \epsilon/2) \leq \frac{p(1-p)}{(p-\epsilon/2)^2 m} < \frac{4}{\epsilon m} \leq 1/2.$$

We now give an upper bound on $\Pr(F)$. We note that

$$\Pr(F) \leq 2\Pr\left(\bigcup_{h\in H} F_h \times F'_h\right). \qquad \text{(Lemma A.1)}$$

Let $S = Y \cup Y'$, where the union is as multisets. Conditioned on the value of S, the multiset Y is a uniform subset of half of the elements of S. Thus,

$$\begin{aligned} 2\Pr\left(\bigcup_{h\in H} F_h \times F'_h\right) &= 2\mathop{\mathbb{E}}_{S}\left[\mathbb{E}\left[\mathbb{1}_{\{\exists h\in H: h|_Y = c|_Y,\ \mathsf{dist}_{Y'}(h,c)>\epsilon/2\}}\middle|S\right]\right] \\ &= 2\mathop{\mathbb{E}}_{S}\left[\mathbb{E}\left[\mathbb{1}_{\{\exists h'\in H|_S: h'|_Y = c|_Y,\ \mathsf{dist}_{Y'}(h',c)>\epsilon/2\}}\middle|S\right]\right] \\ &\leq 2\mathop{\mathbb{E}}_{S}\left[\sum_{h'\in H|_S}\mathbb{E}\left[\mathbb{1}_{\{h'|_Y = c|_Y,\ \mathsf{dist}_{Y'}(h',c)>\epsilon/2\}}\middle|S\right]\right]. \\ &\qquad \text{(by the union bound)} \end{aligned}$$

Notice that if $\mathsf{dist}_{Y'}(h', c) > \epsilon/2$ then $\mathsf{dist}_S(h', c) > \epsilon/4$, hence the probability that we choose Y such that $h'|_Y = c|_Y$ is at most $(1 - \epsilon/4)^m$. Using Theorem 1.1 we get

$$\Pr(F) \leq 2\mathop{\mathbb{E}}_{S}\left[\sum_{h'\in H|_S}(1-\epsilon/4)^m\right] \leq 2(2m+1)^d(1-\epsilon/4)^m.$$

□

References

1. N. Alon, S. Moran, A. Yehudayoff, Sign rank, VC dimension and spectral gaps. Electronic Colloquium on Computational Complexity (ECCC) vol. 21, no. 135 (2014)
2. D. Angluin, M. Krikis, Learning from different teachers. Mach. Learn. **51**(2), 137–163 (2003)
3. M. Anthony, G. Brightwell, D.A. Cohen, J. Shawe-Taylor. On exact specification by examples, in *COLT*, 1992, pp. 311–318
4. P. Assouad, Densite et dimension. Ann. Inst. Fourier **3**, 232–282 (1983)
5. F. Balbach, Models for algorithmic teaching. PhD thesis, University of Lübeck, 2007
6. S. Ben-David, A. Litman, Combinatorial variability of Vapnik–Chervonenkis classes with applications to sample compression schemes. Discret. Appl. Math. **86**(1), 3–25 (1998)
7. A. Blumer, A. Ehrenfeucht, D. Haussler, M.K. Warmuth, Occam's razor. Inf. Process. Lett. **24**(6), 377–380 (1987)
8. A. Blumer, A. Ehrenfeucht, D. Haussler, M.K. Warmuth, Learnability and the Vapnik–Chervonenkis dimension. J. Assoc. Comput. Mach. **36**(4), 929–965 (1989)
9. X. Chen, Y. Cheng, B. Tang, A note on teaching for VC classes. Electronic Colloquium on Computational Complexity (ECCC), vol. 23, no. 65 (2016)

10. A. Chernikov, P. Simon, Externally definable sets and dependent pairs. Isr. J. Math. **194**(1), 409–425 (2013)
11. N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods* (Cambridge University Press, Cambridge, 2000)
12. T. Doliwa, H.-U. Simon, S. Zilles, Recursive teaching dimension, learning complexity, and maximum classes, in *ALT*, 2010, pp. 209–223
13. P. Domingos, The role of Occam's razor in knowledge discovery. Data Min. Knowl. Discov. **3**(4), 409–425 (1999)
14. R.M. Dudley, Central limit theorems for empirical measures. Ann. Probab. **6**, 899–929 (1978)
15. Z. Dvir, A. Rao, A. Wigderson, A. Yehudayoff, Restriction access, in *Innovations in Theoretical Computer Science*, Cambridge, 8–10, Jan 2012, pp. 19–33
16. S. Floyd, Space-bounded learning and the Vapnik–Chervonenkis dimension, in *COLT*, 1989, pp. 349–364
17. S. Floyd, M.K. Warmuth, Sample compression, learnability, and the Vapnik–Chervonenkis dimension. Mach. Learn. **21**(3), 269–304 (1995)
18. Y. Freund, Boosting a weak learning algorithm by majority. Inf. Comput. **121**(2), 256–285 (1995)
19. S.A. Goldman, M. Kearns, On the complexity of teaching. J. Comput. Syst. Sci. **50**(1), 20–31 (1995)
20. S.A. Goldman, H.D. Mathias, Teaching a smarter learner. J. Comput. Syst. Sci. **52**(2), 255–267 (1996)
21. S.A. Goldman, R.L. Rivest, R.E. Schapire, Learning binary relations and total orders. SIAM J. Comput. **22**(5), 1006–1034 (1993)
22. S. Hanneke, Teaching dimension and the complexity of active learning, in *COLT*, 2007, pp. 66–81
23. D. Haussler, Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik–Chervonenkis dimension. J. Comb. Theory Ser. A **69**(2), 217–232 (1995)
24. D. Haussler, E. Welzl, epsilon-nets and simplex range queries. Discret. Comput. Geom. **2**, 127–151 (1987)
25. D.P. Helmbold, R.H. Sloan, M.K. Warmuth, Learning integer lattices. SIAM J. Comput. **21**(2), 240–266 (1992)
26. D.P. Helmbold, M.K. Warmuth, On weak learning. J. Comput. Syst. Sci. **50**(3), 551–573 (1995)
27. J.C. Jackson, A. Tomkins, A computational model of teaching, in *COLT*, 1992, pp. 319–326
28. M. Kearns, U.V. Vazirani, *An Introduction to Computational Learning Theory* (MIT Press, Cambridge, 1994)
29. H. Kobayashi, A. Shinohara, Complexity of teaching by a restricted number of examples, in *COLT*, 2009
30. C. Kuhlmann, On teaching and learning intersection-closed concept classes, in *EuroCOLT*, 1999, pp. 168–182
31. D. Kuzmin, M.K. Warmuth, Unlabeled compression schemes for maximum classes. J. Mach. Learn. Res. **8**, 2047–2081 (2007)
32. R. Livni, P. Simon, Honest compressions and their application to compression schemes, in *COLT*, 2013, pp. 77–92
33. M. Marchand, J. Shawe-Taylor, The set covering machine. J. Mach. Learn. Res. **3**, 723–746 (2002)
34. S. Moran, A. Yehudayoff. Sample compression for VC classes. Electronic Colloquium on Computational Complexity (ECCC), vol. 22, no. 40 (2015)
35. J. von Neumann, Zur theorie der gesellschaftsspiele. Mathematische Annalen **100**, 295–320 (1928)
36. J.R. Quinlan, R.L. Rivest, Inferring decision trees using the minimum description length principle. Inf. Comput. **80**(3), 227–248 (1989)
37. B.I.P. Rubinstein, P.L. Bartlett, J.H. Rubinstein, Shifting: one-inclusion mistake bounds and sample compression. J. Comput. Syst. Sci. **75**(1), 37–59 (2009)

38. B.I.P. Rubinstein, J.H. Rubinstein, A geometric approach to sample compression. J. Mach. Learn. Res. **13**, 1221–1261 (2012)
39. R. Samei, P. Semukhin, B. Yang, S. Zilles, Algebraic methods proving Sauer's bound for teaching complexity. Theor. Comput. Sci. **558**, 35–50 (2014)
40. R. Samei, P. Semukhin, B. Yang, S. Zilles, Sample compression for multi-label concept classes, in *COLT*, vol. 35, 2014, pp. 371–393
41. N. Sauer, On the density of families of sets. J. Comb. Theory Ser. A **13**, 145–147 (1972)
42. A. Shinohara, S. Miyano, Teachability in computational learning, in *ALT*, 1990, pp. 247–255
43. L.G. Valiant, A theory of the learnable. Commun. ACM **27**, 1134–1142 (1984)
44. V.N. Vapnik, A.Ya. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities. Theory Probab. Appl. **16**, 264–280 (1971)
45. M.K. Warmuth, Compressing to VC dimension many points, in *COLT/Kernel*, 2003, pp. 743–744
46. A. Wigderson, A. Yehudayoff, Population recovery and partial identification, in *FOCS*, 2012, pp. 390–399
47. S. Zilles, S. Lange, R. Holte, M. Zinkevich, Models of cooperative teaching and learning. J. Mach. Learn. Res. **12**, 349–384 (2011)

Restricted Invertibility Revisited

Assaf Naor and Pierre Youssef

Dedicated to Jirka Matoušek

Abstract Suppose that $m, n \in \mathbb{N}$ and that $A : \mathbb{R}^m \to \mathbb{R}^n$ is a linear operator. It is shown here that if $k, r \in \mathbb{N}$ satisfy $k < r \leqslant \mathbf{rank}(A)$ then there exists a subset $\sigma \subseteq \{1, \dots, m\}$ with $|\sigma| = k$ such that the restriction of A to $\mathbb{R}^\sigma \subseteq \mathbb{R}^m$ is invertible, and moreover the operator norm of the inverse $A^{-1} : A(\mathbb{R}^\sigma) \to \mathbb{R}^m$ is at most a constant multiple of the quantity $\sqrt{mr/((r-k)\sum_{i=r}^m \mathsf{s}_i(A)^2)}$, where $\mathsf{s}_1(A) \geqslant \dots \geqslant \mathsf{s}_m(A)$ are the singular values of A. This improves over a series of works, starting from the seminal Bourgain–Tzafriri Restricted Invertibility Principle, through the works of Vershynin, Spielman–Srivastava and Marcus–Spielman–Srivastava. In particular, this directly implies an improved restricted invertibility principle in terms of Schatten–von Neumann norms.

1 Introduction

Given $m, n \in \mathbb{N}$, the rank of a linear operator $A : \mathbb{R}^m \to \mathbb{R}^n$ equals the largest possible dimension of a linear subspace $V \subseteq \mathbb{R}^m$ on which A is injective, i.e., the inverse $A^{-1} : A(V) \to V$ exists. The *restricted invertibility problem* asks for conditions on A that ensure a strengthening of this basic fact from linear algebra

A. N. was supported by BSF grant 2010021, the Packard Foundation and the Simons Foundation.

A. Naor (✉)
Mathematics Department, Princeton University Fine Hall, Washington Road, 08544-1000
Princeton, NJ, USA
e-mail: naor@math.princeton.edu

P. Youssef
Laboratoire de Probabilités et de Modèles Aléatoires, Université Paris-Diderot, 5 rue Thomas
Mann, 75205 CEDEX 13 Paris, France
e-mail: youssef@math.univ-paris-diderot.fr

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_27

in two ways, corresponding to additional *structural information* on the subspace $V \subseteq \mathbb{R}^m$ on which A is injective, as well as *quantitative information* on the behavior of the inverse $A^{-1} : A(V) \to V$. Firstly, the goal is to find a large dimensional *coordinate subspace* on which A is invertible, i.e., we wish to find a large subset $\sigma \subseteq \{1, \dots, m\}$ such that A is injective on $\mathbb{R}^\sigma \subseteq \mathbb{R}^m$. Secondly, rather than being satisfied with mere invertibility we ask for A to be *quantitatively invertible* on $\mathbb{R}^\sigma$ in the sense that the operator norm of the inverse $A^{-1} : A(\mathbb{R}^\sigma) \to \mathbb{R}^\sigma$ is not too large. Obviously, additional assumptions on A are required for such conclusions to hold true.

The following theorem, which is known as the Bourgain–Tzafriri Restricted Invertibility Principle [5, 6, 8], is a seminal result that addressed the above question and had major influence on subsequent research, with a variety of interesting applications to several areas. Throughout what follows, for $m \in \mathbb{N}$ the standard coordinate basis of $\mathbb{R}^m$ will be denoted by $e_1, \dots, e_m \in \mathbb{R}^m$.

Theorem 1 (Bourgain–Tzafriri) *There exist two universal constants $c, C \in (0, \infty)$ with the following property. Suppose that $m \in \mathbb{N}$ and that $A : \mathbb{R}^m \to \mathbb{R}^m$ is a linear operator such that the Euclidean norm of the vector $Ae_j \in \mathbb{R}^m$ equals 1 for every $j \in \{1, \dots, m\}$. Letting $\|A\|$ denote the operator norm of A, there exists a subset $\sigma \subseteq \{1, \dots, m\}$ with $|\sigma| \geqslant cm/\|A\|^2$ such that A is injective on $\mathbb{R}^\sigma$ and the operator norm of the inverse $A^{-1} : A(\mathbb{R}^\sigma) \to \mathbb{R}^\sigma$ is at most C.*

In what follows, for $p \in [1, \infty]$ and $m \in \mathbb{N}$ the ℓ_p norm of a vector $x \in \mathbb{R}^m$ will be denoted as usual by $\|x\|_p$. Thus $\|x\|_2$ is the Euclidean norm of x. We shall also denote (as usual) by ℓ_p^m the normed space $\mathbb{R}^m$ equipped with the ℓ_p norm. The standard scalar product on $\mathbb{R}^m$ will be denoted $\langle \cdot, \cdot \rangle$. For $k, m, n \in \mathbb{N}$ and a k-dimensional subspace $V \subseteq \mathbb{R}^m$, the Schatten–von Neumann p norm of a linear operator $A : V \to \mathbb{R}^n$ will be denoted below by $\|A\|_{\mathsf{S}_p}$. Thus

$$\|A\|_{\mathsf{S}_p} \stackrel{\mathrm{def}}{=} \left(\mathbf{Tr}(A^*A)^{\frac{p}{2}}\right)^{\frac{1}{p}} = \left(\sum_{j=1}^{k} \mathsf{s}_j(A)^p\right)^{\frac{1}{p}},$$

where $\mathsf{s}_1(A) \geqslant \mathsf{s}_2(A) \geqslant \dots \geqslant \mathsf{s}_k(A)$ denote the singular values of A, i.e., they are the (decreasing rearrangement of the) eigenvalues of the positive semidefinite operator $\sqrt{A^*A} : V \to V^*$. Thus $\|A\|_{\mathsf{S}_\infty} = \mathsf{s}_1(A)$ is the operator norm of A. Also, $\|A\|_{\mathsf{S}_2}$ is the Hilbert–Schmidt norm of A, i.e., for every orthonormal basis $u_1, \dots, u_k$ of V we have $\|A\|_{\mathsf{S}_2}^2 = \sum_{i=1}^k \sum_{j=1}^n \langle Au_i, e_j \rangle^2 = \sum_{i=1}^k \|Ae_i\|_2^2$. Below it will sometimes be convenient to denote the smallest singular value of A by $\mathsf{s}_{\min}(A) = \mathsf{s}_k(A)$. Thus A is injective if and only if $\mathsf{s}_{\min}(A) > 0$, in which case $\|A^{-1}\|_{\mathsf{S}_\infty} = 1/\mathsf{s}_{\min}(A)$.

Given $m \in \mathbb{N}$ and $\sigma \subseteq \{1, \dots, m\}$ it will be convenient to denote the formal identity from $\mathbb{R}^\sigma$ to $\mathbb{R}^m$ by $J_\sigma : \mathbb{R}^\sigma \to \mathbb{R}^m$, i.e., $J_\sigma((a_j)_{j\in\sigma}) = \sum_{j\in\sigma} a_j e_j$ for every $(a_j)_{j\in\sigma} \in \mathbb{R}^\sigma$. With this notation, given an operator $A : \mathbb{R}^m \to \mathbb{R}^n$ that is injective on $\mathbb{R}^\sigma$ we can consider the operator $(AJ_\sigma)^{-1} : A(\mathbb{R}^\sigma) \to \mathbb{R}^\sigma$. We shall sometimes drop the need to mention explicitly that A is injective on $\mathbb{R}^\sigma$ by adhering to the convention that if A is not injective on $\mathbb{R}^\sigma$ then $\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} = \infty$.

Using the above notation, Theorem 1 asserts that if $A : \mathbb{R}^m \to \mathbb{R}^m$ is a linear operator that satisfies $\|Ae_j\|_2 = 1$ for all $j \in \{1, \ldots, m\}$ then there exists $\sigma \subseteq \{1, \ldots, m\}$ with $|\sigma| \gtrsim m/\|A\|_{\mathsf{S}_\infty}$ such that $\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} \lesssim 1$, or equivalently $\mathsf{s}_{\min}(AJ_\sigma) \gtrsim 1$. Here, and in what follows, we use the following standard asymptotic notation. Given two quantities $K, L \in \mathbb{R}$ the notation $K \lesssim L$ (respectively $K \gtrsim L$) means that there exists a universal constant $c \in (0, \infty)$ such that $K \leqslant cL$ (respectively $K \geqslant cL$). The notation $K \asymp L$ means that both $K \lesssim L$ and $K \gtrsim L$ hold true.

The following theorem is a useful strengthening of the Bourgain–Tzafriri Restricted Invertibility Principle that was discovered by Vershynin in [33].

Theorem 2 (Vershynin) *There exists a universal constant $c \in (0, \infty)$ with the following property. Fix $k, m, n \in \mathbb{N}$. Let $A : \mathbb{R}^m \to \mathbb{R}^n$ be a linear operator with $\|Ae_j\|_2 = 1$ for all $j \in \{1, \ldots, m\}$. Also, let $\Delta : \mathbb{R}^n \to \mathbb{R}^n$ be a positive definite diagonal operator, i.e., there exist $d_1, \ldots, d_n \in (0, \infty)$ such that $\Delta x = (d_1 x_1, \ldots, d_n x_n)$ for every $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Suppose that $k < \|A\Delta\|_{\mathsf{S}_2}^2 / \|A\Delta\|_{\mathsf{S}_\infty}^2$ and write $k = (1 - \varepsilon)\|A\Delta\|_{\mathsf{S}_2}^2 / \|A\Delta\|_{\mathsf{S}_\infty}^2$ where $\varepsilon \in (0, 1)$ (thus $\varepsilon = 1 - k\|A\Delta\|_{\mathsf{S}_\infty}^2 / \|A\Delta\|_{\mathsf{S}_2}^2$). Then there exists a subset $\sigma \subseteq \{1, \ldots, m\}$ with $|\sigma| = k$ such that $\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} \leqslant \varepsilon^{-c\log(1/\varepsilon)}$.*

For a linear operator $T : \mathbb{R}^m \to \mathbb{R}^n$, the quantity $\|T\|_{\mathsf{S}_2}^2 / \|T\|_{\mathsf{S}_\infty}^2$ is often called the *stable rank* of T, though this terminology sometimes also refers to the quantity $\|T\|_{\mathsf{S}_1} / \|T\|_{\mathsf{S}_\infty}$. In both cases, the use of the term ‘stable’ in this context expresses the fact that the quantity in question is a robust replacement for the rank of T in the sense that the rank of T could be large due to the fact that T has many positive but nevertheless very small singular values, while if the stable rank of T is large then its singular values are large on average. Below we shall use the terminology ‘stable rank’ exclusively for the quantity $\|T\|_{\mathsf{S}_2}^2 / \|T\|_{\mathsf{S}_\infty}^2$, which we denote by $\mathbf{srank}(T) = \|T\|_{\mathsf{S}_2}^2 / \|T\|_{\mathsf{S}_\infty}^2$.

Theorem 1 coincides with the special case $\varepsilon = \frac{1}{2}$ and $\Delta = I_n$ of Theorem 2, where I_n is the identity operator on $\mathbb{R}^n$. However, Theorem 2 improves over Theorem 1 in three ways that are important for geometric applications. Firstly, Theorem 2 treats rectangular matrices while Theorem 1 treats only the case $m = n$. Secondly, even in the special case $\Delta = I_n$ of Theorem 2 the size of the subset $\sigma \subseteq \{1, \ldots, m\}$ is allowed to be arbitrarily close to $\mathbf{srank}(A)$, while in Theorem 1 it can only be taken to be a constant multiple of $\mathbf{srank}(A)$. Lastly, Theorem 2 actually allows for the size of the subset $\sigma \subseteq \{1, \ldots, m\}$ to be arbitrarily close to the supremum of $\mathbf{srank}(A\Delta)$ over all positive definite diagonal operators $\Delta : \mathbb{R}^m \to \mathbb{R}^m$, a quantity that could be much larger than $\mathbf{srank}(A)$.

Remark 3 Theorem 2 is often stated in the literature as a subset selection principle for John decompositions of the identity. Namely, suppose that $k, m, n \in \mathbb{N}$ and $x_1, \ldots, x_m \in \mathbb{R}^n \smallsetminus \{0\}$ satisfy $\sum_{j=1}^m \langle x_j, y\rangle^2 = \|y\|_2^2$ for all $y \in \mathbb{R}^n$. Equivalently, we have $\sum_{j=1}^m x_j \otimes x_j = I_n$, where for $x, y \in \mathbb{R}^n$ the rank-one operator $x \otimes y : \mathbb{R}^n \to \mathbb{R}^n$

is defined as usual by setting $(x \otimes y)(z) = \langle x, z\rangle y$ for every $z \in \mathbb{R}^n$. Suppose that $T : \mathbb{R}^n \to \mathbb{R}^n$ is a linear operator satisfying $Tx_1, \dots, Tx_m \neq 0$, and that $k = (1-\varepsilon)\mathbf{srank}(T)$ for some $\varepsilon \in (0,1)$. Then there exists $\sigma \subseteq \{1,\dots,m\}$ with $|\sigma| = k$ such that

$$\forall\, \{a_j\}_{j\in\sigma} \subseteq \mathbb{R}, \qquad \Big\| \sum_{j\in\sigma} \frac{a_j}{\|Tx_j\|_2} Tx_j \Big\|_2 \geqslant \varepsilon^{c\log(1/\varepsilon)} \Big(\sum_{j\in\sigma} a_j^2 \Big)^{\frac{1}{2}}.$$

The above formulation is equivalent to Theorem 2 as stated in terms of rectangular matrices by considering the operator $A : \mathbb{R}^m \to \mathbb{R}^n$ that is given by $Ae_j = Tx_j/\|Tx_j\|_2$ for every $j \in \{1,\dots,m\}$.

A recent breakthrough of Spielman–Srivastava [26], that relies nontrivially on a remarkable method for sparsifying quadratic forms that was developed by Batson–Spielman–Srivastava [2] (see also the survey [20]), yielded the following improved restricted invertibility principle, via techniques that are entirely different from those used by Bourgain–Tzafriri and Vershynin.

Theorem 4 (Spielman–Srivastava) *Suppose that $k, m, n \in \mathbb{N}$ and let $A : \mathbb{R}^m \to \mathbb{R}^n$ be a linear operator such that $k < \mathbf{srank}(A)$. Write $k = (1-\varepsilon)\mathbf{srank}(A)$ where $\varepsilon \in (0,1)$. Then there exists a subset $\sigma \subseteq \{1,\dots,m\}$ with $|\sigma| = k$ such that*

$$\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} \leqslant \frac{1}{1-\sqrt{1-\varepsilon}} \cdot \frac{\sqrt{m}}{\|A\|_{\mathsf{S}_2}} \leqslant \frac{2\sqrt{m}}{\varepsilon \|A\|_{\mathsf{S}_2}}.$$

In the setting of Theorem 4, since $\|A\|_{\mathsf{S}_2} = \sqrt{m}$ when the columns of A have unit Euclidean norm, Theorem 1 is a special case of Theorem 4. As in the case $\Delta = I_n$ of Theorem 2, the statement of Theorem 4 has the additional feature that the subset $\sigma \subseteq \{1,\dots,m\}$ can have size arbitrarily close to $\mathbf{srank}(A)$. Moreover, in Theorem 4 the columns of A need not have unit Euclidean norm, and the upper bound on $\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty}$ in terms of ε is much better in Theorem 4 than the corresponding bound in the case $\Delta = I_n$ of Theorem 2; in fact this bound is asymptotically sharp [3] as $\varepsilon \to 0$. An additional feature of Theorem 4 is that its proof in [26] yields a deterministic polynomial time algorithm for finding the subset σ, while previous to [26] only a randomized polynomial time algorithm was available [32]. Theorem 2 does have a feature that Theorem 4 does not, namely the size of the subset $\sigma \subseteq \{1,\dots,m\}$ can be taken to be arbitrarily close to the supremum of $\mathbf{srank}(A\Delta)$ over all positive definite diagonal operators $\Delta : \mathbb{R}^m \to \mathbb{R}^m$, albeit with worse dependence on ε. However, in [34] it was shown how to combine the features of Theorems 2 and 4 so as to yield this stronger guarantee with the better dependence on ε that is asserted in Theorem 4. This improvement is important for certain geometric applications [34]. The new results that are presented below have this stronger "weighted" feature, but for the sake of simplicity of the initial discussion in the Introduction we shall first present all the ensuing statements in their "unweighted" form that corresponds to the way Theorem 4 is stated above.

A different proof of Theorem 4 in the special case $AA^* = I_n$ was found by Marcus, Spielman and Srivastava in [16], using their powerful method of interlacing polynomials [17, 18]. In fact, their forthcoming work [19] obtains Theorem 5 below, which yields for the first time a restricted invertibility principle for subsets that can be asymptotically larger than the stable rank, with their size depending on the ratio of the Hilbert–Schmidt norm and the Schatten–von Neumann 4 norm. This result was announced by Srivastava in his talk at the conference *Banach Spaces: Geometry and Analysis* (Hebrew University, May 2013), and it is actually a precursor to the outstanding subsequent work [18]. Its proof will appear for the first time in the forthcoming preprint [19], but we confirmed with the authors that they obtain Theorem 5 as stated below.

Theorem 5 (Marcus–Spielman–Srivastava) *Suppose that $k, m, n \in \mathbb{N}$ and let $A : \mathbb{R}^m \to \mathbb{R}^n$ be a linear operator such that $k < \frac{1}{4}(\|A\|_{\mathsf{S}_2}/\|A\|_{\mathsf{S}_4})^4$. Define $\varepsilon \in (3/4, 1)$ by $k = (1-\varepsilon)\|A\|_{\mathsf{S}_2}^4/\|A\|_{\mathsf{S}_4}^4$. Then there exists a subset $\sigma \subseteq \{1, \ldots, m\}$ with $|\sigma| = k$ such that*

$$\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} \leqslant \frac{1}{\sqrt{1 - 2\sqrt{1-\varepsilon}}} \cdot \frac{\sqrt{m}}{\|A\|_{\mathsf{S}_2}}. \tag{1}$$

Theorem 5 can be much better than the previously known restricted invertibility principles at detecting large well-invertible sub-matrices. To state a concrete example, suppose that the singular values of A are $\mathsf{s}_1(A) \asymp \sqrt[4]{m}$ and $\mathsf{s}_2(A) \asymp \mathsf{s}_3(A) \asymp \ldots \asymp \mathsf{s}_m(A) = 1$. Then Theorem 4 yields a subset $\sigma \subseteq \{1, \ldots, m\}$ of size of order $\sqrt{m}$ for which the operator norm of the inverse of AJ_σ is $O(1)$, while Theorem 5 yields such a subset whose size is at least a constant multiple of m.

The restriction $k < \frac{1}{4}(\|A\|_{\mathsf{S}_2}/\|A\|_{\mathsf{S}_4})^4$ in Theorem 5 ensures that $\varepsilon > 3/4$, so that the quantity appearing under the square root in (1) is positive. Thus, in the statement of Theorem 5, k cannot be arbitrarily close to the "modified stable rank" $\|A\|_{\mathsf{S}_2}^4/\|A\|_{\mathsf{S}_4}^4$, but this will be remedied below.

It is important to note that the quantity $\|A\|_{\mathsf{S}_2}^4/\|A\|_{\mathsf{S}_4}^4$ is always at least $\mathbf{srank}(A)$. More generally, given $p \in (2, \infty]$, if we define the p-stable rank of A to be the quantity

$$\mathbf{srank}_p(A) \stackrel{\mathrm{def}}{=} \left(\frac{\|A\|_{\mathsf{S}_2}}{\|A\|_{\mathsf{S}_p}}\right)^{\frac{2p}{p-2}}, \tag{2}$$

then in particular $\mathbf{srank}_4(A) = \|A\|_{\mathsf{S}_2}^4/\|A\|_{\mathsf{S}_4}^4$ and $\mathbf{srank}_\infty(A) = \mathbf{srank}(A)$. We claim that

$$p \geqslant q > 2 \implies \mathbf{srank}_p(A) \leqslant \mathbf{srank}_q(A), \tag{3}$$

Indeed, by direct application of Hölder's inequality we have

$$\|A\|_{\mathsf{S}_q} \leqslant \|A\|_{\mathsf{S}_2}^{\frac{2(p-q)}{q(p-2)}} \cdot \|A\|_{\mathsf{S}_p}^{\frac{p(q-2)}{q(p-2)}},$$

which simplifies to give (3). The limit as $p \to 2^+$ of $\mathbf{srank}_p(A)$ can be computed explicitly, yielding the quantity below, denoted $\mathbf{Entrank}(A)$, which we naturally call the entropic stable rank of A.

$$\mathbf{Entrank}(A) \overset{\mathrm{def}}{=} \lim_{p\to 2+} \mathbf{srank}_p(A) = \exp\left(\log \sum_{j=1}^{m} \mathsf{s}_j(A)^2 - \frac{2\sum_{j=1}^{m} \mathsf{s}_j(A)^2 \log \mathsf{s}_j(A)}{\sum_{j=1}^{m} \mathsf{s}_j(A)^2}\right)$$

$$= \exp\left(\frac{\mathbf{Tr}(A^*A)\log \mathbf{Tr}(A^*A) - \mathbf{Tr}(A^*A\log(A^*A))}{\mathbf{Tr}(A^*A)}\right) = \|A\|_{\mathsf{S}_2}^2 \prod_{j=1}^{m} \mathsf{s}_j(A)^{-\frac{2\mathsf{s}_j(A)^2}{\|A\|_{\mathsf{S}_2}^2}}.$$

As we shall explain in the next section, here we obtain an improved restricted invertibility theorem that in particular yields a strengthening of Theorem 5 that allows one to make use of the p-stable rank of A for every $p > 2$, thus producing well-invertible sub-matrices of A of size that can be any integer that is less than the entropic stable rank of A.

1.1 Restricted Invertibility in Terms of Rank

Our main new result is the following theorem.

Theorem 6 *Suppose that $k, m, n \in \mathbb{N}$. Let $A : \mathbb{R}^m \to \mathbb{R}^n$ be a linear operator with* $\mathbf{rank}(A) > k$. *Then there exists a subset $\sigma \subseteq \{1, \ldots, m\}$ with $|\sigma| = k$ such that*

$$\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} \lesssim \min_{r\in\{k+1,\ldots,\mathbf{rank}(A)\}} \sqrt{\frac{mr}{(r-k)\sum_{i=r}^{m} \mathsf{s}_i(A)^2}}. \tag{4}$$

Example 7 To illustrate the relation between Theorems 4, 5 and 6, consider a linear operator $A : \mathbb{R}^m \to \mathbb{R}^n$ with $\mathsf{s}_j(A) \asymp 1/\sqrt{j}$ for every $j \in \{1, \ldots, m\}$. Thus $\mathbf{rank}(A) = m$, $\mathbf{srank}(A) \asymp \log m$ and $\mathbf{srank}_4(m) \asymp (\log m)^2$. Since $\sqrt{m}/\|A\|_{\mathsf{S}_2} \asymp \sqrt{m/\log m}$, Theorem 4 yields $\sigma \subseteq \{1, \ldots, m\}$ with $|\sigma| \asymp \log m$ and $\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} \lesssim \sqrt{m/\log m}$, Theorem 5 yields such a subset with $|\sigma| \asymp (\log m)^2$, and Theorem 6 yields such a subset with $|\sigma| \gtrsim \sqrt{m}$. In fact, for every $\varepsilon \in (0, 1)$, Theorem 6 yields $\sigma \subseteq \{1, \ldots, m\}$ with $|\sigma| \gtrsim m^{1-\varepsilon}$ such that $\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} \lesssim \frac{1}{\sqrt{\varepsilon}}\sqrt{m/\log m}$.

Theorem 6 has the feature that it asserts the existence of a coordinate subspace of dimension arbitrarily close to the rank of the given operator on which it is invertible, with quantitative control on the operator norm of the inverse. The rank is not a stable quantity, but it is simple to deduce stable consequences of Theorem 6 that are stronger than Theorem 5. Indeed, continuing with the notations of Theorem 6, for every $p \in (2,\infty)$ we can apply Hölder's inequality to deduce that

$$\begin{aligned}\|A\|_{\mathsf{S}_2}^2 &= \sum_{i=1}^{r-1} \mathsf{s}_i(A)^2 + \sum_{i=r}^{m} \mathsf{s}_i(A)^2\\ &\leqslant (r-1)^{1-\frac{2}{p}}\Bigg(\sum_{i=1}^{r-1}\mathsf{s}_i(A)^p\Bigg)^{\frac{2}{p}} + \sum_{i=r}^{m}\mathsf{s}_i(A)^2 \leqslant (r-1)^{1-\frac{2}{p}}\|A\|_{\mathsf{S}_p}^2 + \sum_{i=r}^{m}\mathsf{s}_i(A)^2.\end{aligned}$$

Hence,

$$\sum_{i=r}^{m}\mathsf{s}_i(A)^2 \geqslant \|A\|_{\mathsf{S}_2}^2 - (r-1)^{1-\frac{2}{p}}\|A\|_{\mathsf{S}_p}^2 \overset{(2)}{=} \|A\|_{\mathsf{S}_2}^2\left(1-\left(\frac{r-1}{\mathbf{srank}_p(A)}\right)^{1-\frac{2}{p}}\right). \qquad (5)$$

A substitution of (5) into (4) yields the following estimate.

$$\mathsf{s}_{\min}(AJ_\sigma)^2 \gtrsim \max_{r\in\{k+1,\ldots,\mathbf{srank}_p(A)\}}\left(1-\frac{k}{r}\right)\left(1-\left(\frac{r-1}{\mathbf{srank}_p(A)}\right)^{1-\frac{2}{p}}\right)\cdot\frac{\|A\|_{\mathsf{S}_2}^2}{m}. \qquad (6)$$

The estimate (6) is nontrivial only when $k < \mathbf{srank}_p(A)$, so write $k = (1-\varepsilon)\mathbf{srank}_p(A)$ for some $\varepsilon \in (0,1)$. One checks that the following choice of $r \in \{k+1,\ldots,\mathbf{srank}_p(A)\}$ attains the maximum in the right hand side of (6), up to universal constant factors. If ε is bounded away from 1, say $\varepsilon \in (0,1/2]$, choose $r \asymp (1-\varepsilon/2)\mathbf{srank}_p(A)$. If $1/2 < \varepsilon \leqslant 1-e^{-p/(p-2)}$ then choose $r \asymp \log(1/(1-\varepsilon))\cdot\mathbf{srank}_p(A)$. If $1-e^{-p/(p-2)} < \varepsilon < 1$ then choose $r \asymp e^{-p/(p-2)}\mathbf{srank}_p(A)$. Thus,

$$\begin{aligned}0<\varepsilon\leqslant\frac{1}{2} &\implies \|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} \lesssim \sqrt{\frac{p}{p-2}}\cdot\frac{\sqrt{m}}{\varepsilon\|A\|_{\mathsf{S}_2}},\\ \frac{1}{2}<\varepsilon\leqslant 1-e^{-\frac{p}{p-2}} &\implies \|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} \lesssim \sqrt{\frac{p}{p-2}}\cdot\frac{\sqrt{m}}{\log\left(1/(1-\varepsilon)\right)\|A\|_{\mathsf{S}_2}},\\ 1-e^{-\frac{p}{p-2}}<\varepsilon<1 &\implies \|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} \lesssim \frac{\sqrt{m}}{\|A\|_{\mathsf{S}_2}}.\end{aligned}$$

A more concise way to write these estimates is as follows.

$$\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} \lesssim \left(1+\frac{p}{(p-2)\left|\log\left(1-\varepsilon^2\right)\right|}\right)^{\frac{1}{2}}\frac{\sqrt{m}}{\|A\|_{\mathsf{S}_2}}.$$

For ease of future reference, we record the above corollary of Theorem 6 as Theorem 8 below.

Theorem 8 (Restricted invertibility in terms of Schatten–von Neumann norms) *Suppose that $k, m, n \in \mathbb{N}$, $\varepsilon \in (0,1)$ and $p \in (2,\infty)$. Let $A : \mathbb{R}^m \to \mathbb{R}^n$ be a linear operator that satisfies $k \leqslant (1-\varepsilon)\mathbf{srank}_p(A)$. Then there exists a subset $\sigma \subseteq \{1,\ldots,m\}$ with $|\sigma| = k$ such that*

$$\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} \lesssim \left(1 + \frac{p}{(p-2)\big|\log(1-\varepsilon^2)\big|}\right)^{\frac{1}{2}} \frac{\sqrt{m}}{\|A\|_{\mathsf{S}_2}}.$$

Equivalently, if $k < \mathbf{Entrank}(A)$ then there exists $\sigma \subseteq \{1,\ldots,m\}$ with $|\sigma| = k$ such that

$$\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} \lesssim \inf_{p>2} \psi_p\left(1 - \frac{k}{\mathbf{srank}_p(A)}\right) \frac{\sqrt{m}}{\|A\|_{\mathsf{S}_2}},$$

where $\psi_p : \mathbb{R} \to [0,\infty]$ is defined by $\psi_p(\varepsilon) = \infty$ if $\varepsilon \leqslant 0$, $\psi_p(x) = (\sqrt{p/(p-2)})/\varepsilon$ if $0 < \varepsilon < 1/2$, $\psi_p(\varepsilon) = (\sqrt{p/(p-2)})/\log(1/(1-\varepsilon))$ if $1/2 < \varepsilon \leqslant 1 - e^{-p/(p-2)}$ and $\psi_p(\varepsilon) = 1$ if $\varepsilon > 1 - e^{-p/(p-2)}$.

The case $p = 4$ of Theorem 8 implies (up to constant factors) the conclusion of Theorem 5, though now treating any $\varepsilon \in (0,1)$, i.e., k arbitrarily close to $\mathbf{srank}_4(A)$, while Theorem 5 applies only when $\varepsilon > 3/4$. Theorem 8 can detect the well-invertibility of A on coordinate subspaces that are much larger than those detected by Theorem 5. For example suppose that the singular values of A are $\mathsf{s}_1(A) \asymp \sqrt[3]{m}$ and $\mathsf{s}_2(A) \asymp \mathsf{s}_3(A) \asymp \ldots \asymp \mathsf{s}_m(A) \asymp 1$. Then Theorem 5 yields a subset $\sigma \subseteq \{1,\ldots,m\}$ of size of order $m^{2/3}$ for which the operator norm of the inverse of AJ_σ is $O(1)$, while (the case $p = 3$ of) Theorem 8 yields such a subset whose size is proportional to m.

We shall prove Theorem 6 through an application of Theorem 9 below, which is a restricted invertibility statement of independent interest, in combination with a volumetric argument that leads to Lemma 10 below. Throughout what follows, given $n \in \mathbb{N}$ and a linear subspace $F \subseteq \mathbb{R}^n$, we shall denote the orthogonal projection from $\mathbb{R}^n$ onto F by $\mathsf{Proj}_F : \mathbb{R}^n \to F$.

Theorem 9 *Fix $k, m, n \in \mathbb{N}$ and a linear operator $A : \mathbb{R}^m \to \mathbb{R}^n$ satisfying $\mathbf{rank}(A) > k$. Let $\omega \subseteq \{1,\ldots,m\}$ be any subset with $|\omega| = \mathbf{rank}(A)$ such that the vectors $\{Ae_i\}_{i\in\omega} \subseteq \mathbb{R}^n$ are linearly independent. For every $j \in \omega$ let $F_j \subseteq \mathbb{R}^n$ be the orthogonal complement of the span of $\{Ae_i\}_{i\in\omega\smallsetminus\{j\}} \subseteq \mathbb{R}^n$, i.e.,*

$$F_j \stackrel{\mathrm{def}}{=} \left(\mathbf{span}\,\{Ae_i\}_{i\in\omega\smallsetminus\{j\}}\right)^{\perp}. \tag{7}$$

Then there exists a subset $\sigma \subseteq \omega$ with $|\sigma| = k$ such that

$$\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} \lesssim \frac{\sqrt{\mathbf{rank}(A)}}{\sqrt{\mathbf{rank}(A) - k}} \cdot \max_{j\in\omega} \frac{1}{\|\mathsf{Proj}_{F_j} Ae_j\|_2}. \tag{8}$$

The link between Theorems 9 and 6 is furnished through the following lemma.

Lemma 10 *Fix $r, m, n \in \mathbb{N}$. Let $A : \mathbb{R}^m \to \mathbb{R}^n$ be a linear operator with $\mathbf{rank}(A) \geqslant r$. For every $\tau \subseteq \{1, \dots, m\}$ let $E_\tau \subseteq \mathbb{R}^n$ be the orthogonal complement of the span of $\{Ae_j\}_{j\in\tau} \subseteq \mathbb{R}^n$, i.e.,*[1]

$$E_\tau \stackrel{\mathrm{def}}{=} \left(\mathbf{span}\left\{Ae_j\right\}_{j\in\tau}\right)^{\perp}. \tag{9}$$

Then there exists a subset $\tau \subseteq \{1, \dots, m\}$ with $|\tau| = r$ such that

$$\forall\, j \in \tau, \qquad \left\|\mathsf{Proj}_{E_{\tau\smallsetminus\{j\}}} Ae_j\right\|_2 \geqslant \frac{1}{\sqrt{m}}\left(\sum_{i=r}^{m} \mathsf{s}_i(A)^2\right)^{\frac{1}{2}}. \tag{10}$$

The deduction of Theorem 6 from Theorem 9 and Lemma 10 is simple. Indeed, in the setting of Theorem 6, take $r \in \{k+1, \dots, \mathbf{rank}(A)\}$ and apply Lemma 10 to obtain a subset $\tau \subseteq \{1, \dots, m\}$ with $|\tau| = r$ that satisfies (10). This implies in particular that $\{Ae_j\}_{j\in\tau}$ are linearly independent, hence the operator $AJ_\tau : \mathbb{R}^\tau \to \mathbb{R}^n$ has rank r. By Theorem 9 applied with A replaced by AJ_τ, $m = r = \mathbf{rank}(A)$ and $\omega = \tau$, we obtain a further subset $\sigma \subseteq \tau$ with $|\sigma| = k$ such that

$$\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} \overset{(8)\wedge(10)}{\lesssim} \sqrt{\frac{mr}{(r-k)\sum_{i=r}^{m} \mathsf{s}_i(A)^2}}.$$

This is precisely the assertion of Theorem 6.

In Sect. 5 we shall prove the following variant of Theorem 9.

Theorem 11 *Fix $k, m, n \in \mathbb{N}$ and a linear operator $A : \mathbb{R}^m \to \mathbb{R}^n$ satisfying $\mathbf{rank}(A) > k$. Then there exists a subset $\sigma \subseteq \{1, \dots, m\}$ with $|\sigma| = k$ such that*

$$\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} \leqslant \frac{\sqrt{m}}{\sqrt{\mathbf{rank}(A)} - \sqrt{k}}\left(\frac{1}{\mathbf{rank}(A)}\sum_{i=1}^{\mathbf{rank}(A)} \frac{1}{\mathsf{s}_i(A)^2}\right)^{\frac{1}{2}}. \tag{11}$$

To explain how Theorem 11 relates to Theorem 6, note that in the setting of Theorem 6 we have

$$\sum_{j\in\omega} \frac{1}{\|\mathsf{Proj}_{F_j} Ae_j\|_2^2} = \sum_{i=1}^{\mathbf{rank}(A)} \frac{1}{\mathsf{s}_i(AJ_\omega)^2}. \tag{12}$$

The simple linear-algebraic justification of (12) appears in Sect. 2.1 below. For simplicity suppose that $\omega = \{1, \dots, m\}$, so $\mathbf{rank}(A) = m$, and write $k = (1-\varepsilon)m$

[1] Comparing (7) and (9) we see that $F_j = E_{\omega\smallsetminus\{j\}}$ for every $j \in \omega$.

for some $\varepsilon \in (0, 1)$. Then Theorem 6 yields a subset $\sigma \subseteq \{1, \dots, m\}$ with $|\sigma| = k$ such that

$$\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} \lesssim \frac{1}{\sqrt{\varepsilon}} \cdot \max_{j \in \{1,\dots,m\}} \frac{1}{\|\mathsf{Proj}_{F_j} Ae_j\|_2}, \tag{13}$$

while, due to (12), Theorem 11 yields a subset $\sigma \subseteq \{1, \dots, m\}$ with $|\sigma| = k$ such that

$$\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} \leqslant \frac{1}{1 - \sqrt{1-\varepsilon}} \left(\frac{1}{m} \sum_{j=1}^{m} \frac{1}{\|\mathsf{Proj}_{F_j} Ae_j\|_2^2} \right)^{\frac{1}{2}} \asymp \frac{1}{\varepsilon} \left(\frac{1}{m} \sum_{j=1}^{m} \frac{1}{\|\mathsf{Proj}_{F_j} Ae_j\|_2^2} \right)^{\frac{1}{2}}. \tag{14}$$

The estimates (13) and (14) are incomparable since (13) yields a dependence on ε that is better than that of (14) as $\varepsilon \to 0$, while the bound in (14) is in terms of the average of the quantities $\left\{1/\|\mathsf{Proj}_{F_j} Ae_j\|_2^2\right\}_{j=1}^{m}$ rather than their maximum. It remains an interesting open question whether one could obtain a restricted invertibility theorem that combines the best terms in (13) and (14).

Remark 12 Theorem 9 is best possible, up to constant factors. Indeed, fix $k, m \in \mathbb{N}$ with $k < m$ and let B be the m by m matrix all of whose diagonal entries equal m and all of whose off-diagonal entries equal -1. Then B is positive definite (diagonal-dominant) and we choose $A = \sqrt{B}$. We are thus in the setting of Theorem 9 with $m = n = \mathbf{rank}(A)$ and $\omega = \{1, \dots, m\}$. The quantity $1/\|\mathsf{Proj}_{F_j} Ae_j\|_2^2$ is equal to the j'th diagonal entry of $(A^*A)^{-1} = B^{-1}$; see equation (16) in Sect. 2.1 below for a simple justification of this fact. The matrix B is an invertible circulant matrix, and as such B^{-1} is also a circulant matrix whose diagonal entries equal $2/(m + 1)$; see [9, 15] for more on the explicit evaluation of basic quantities related to circulant matrices, including their inverses and eigenvalues, which we use here. Therefore $1/\|\mathsf{Proj}_{F_j} Ae_j\|_2 = \sqrt{2/(m + 1)}$ for every $j \in \{1, \dots, m\}$, so that the right hand side of (8) equals $\sqrt{2m/((m + 1)(m - k))} \asymp 1/\sqrt{m-k}$. At the same time, take any $\sigma \subseteq \{1, \dots, m\}$ with $|\sigma| = k$. Then $(AJ_\sigma)^*(AJ_\sigma) = J_\sigma^* B J_\sigma$ corresponds to a k by k matrix whose diagonal entries equal m and whose off-diagonal entries equal -1. This is again a circulant matrix whose eigenvalues equal to $m + 1$ with multiplicity $k-1$ and $m+1-k$ with multiplicity 1. Thus $\mathsf{s}_1(AJ_\sigma) = \dots = \mathsf{s}_{k-1}(AJ_\sigma) = \sqrt{m+1}$ and $\mathsf{s}_k(AJ_\sigma) = \mathsf{s}_{\min}(AJ_\sigma) = 1/\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} = \sqrt{m+1-k}$. This shows that $\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} \asymp 1/\sqrt{m-k}$, so that (8) is sharp up to constant factors.

1.2 Remarks on the Proofs

The original proof of Bourgain and Tzafriri of Theorem 1 consists of a beautiful combination of probabilistic, combinatorial and analytic arguments. It proceeds roughly along three steps. Firstly, using random selectors one finds a large collection of columns of A that is "well separated." In the second step one uses the Sauer–Shelah lemma [24, 25] to find a further subset of the columns such that the inverse of the restriction of A to this subset, when viewed as an operator from ℓ_2 to ℓ_1, has small norm; the Sauer–Shelah lemma is discussed in Sect. 2.4 below, since it plays an important role here as well. The third step of the Bourgain–Tzafriri proof uses tools from functional analysis, specifically the Little Grothendieck's Inequality [14] and the Pietsch Domination Theorem [22], to control the desired Hilbertian operator norm; these analytic tools are used here as well, and are explained in detail in Sects. 2.2 and 2.3 below.

Vershynin's proof of Theorem 2 uses the Bourgain–Tzafriri restricted invertibility theorem as a "black box," alongside with (unpublished) work of Kashin and Tzafriri (see Theorem 2.5 in [33]). A key contribution of Verhynin was the idea to work with the Hilbert–Schmidt norm so as to allow for an iterative argument. As we stated earlier, the proof of Spielman and Srivastava of Theorem 4 is entirely different from the previously used methods in this context, relying on the 'sparsification method' of Batson–Spielman–Srivastava [2]. This refreshing approach led to many important developments, and it was subsequently augmented by the powerful 'method of interlacing polynomials' of Marcus–Spielman–Srivastava, which they used to prove Theorem 5, showing that one could use higher Schatten–von Neumann norms to address the restricted invertibility problem.

Our starting point here was the realization that one could use ideas and techniques that predate the works of Vershynin, Spielman–Srivastava and Marcus–Spielman–Srivastava to obtain asymptotically sharp results such as Theorem 4, and even to strengthen the statement in terms of higher Schatten–von Neumann norms that is contained in Theorem 5. These later results were based on the discovery of powerful new techniques, leading to many additional applications (crowned by the solution of the Kadison–Singer problem [18]) that are not covered here, but the present work shows how to apply classical methods to improve over the best known bounds on the restricted invertibility problem. Specifically, we rely on the beautiful work of Giannopoulos [13], which treats a seemingly unrelated geometric question (see also [12]), though it is partially inspired by the work of Bourgain–Tzafriri [5] itself, as well as the works of Bourgain–Szarek [7] and Szarek–Talagrand [30] (see also [29]). The key step is to use Giannopoulos' clever iterative application of the Sauer–Shelah lemma (Bourgain–Tzafriri used the Sauer–Shelah lemma only once in their original argument) in the proof of Theorem 9. In fact, one could use a geometric statement of Giannopoulos [13] as a "black box" so as to obtain a shorter proof of Theorem 9; this is carried out in Sect. 4.1 below, but only after we present a self-contained argument in Sect. 4.

Theorem 11 is of a different nature, since its proof uses the Marcus–Spielman–Srivastava method of interlacing polynomials. We do not see how to prove it using the classical analytic techniques that are utilized elsewhere in this article, and in fact we do not need it for the applications that are obtained here (as we explained earlier, Theorem 11 is incomparable to Theorem 9, being weaker in terms of the dependence on certain parameters and stronger in other respects). Nevertheless, Theorem 11 certainly belongs to the family of restricted invertibility results that we study here.

Among the interesting questions that arise naturally from the present work, we ask whether Theorems 6, 8, 9 and 11 can be made to be algorithmic. Our current proofs do not yield a polynomial time algorithm that finds the desired coordinate subspace, due to various reasons, including (but not limited to) the use of the Sauer–Shelah lemma (in Theorems 6, 8 and 9) and the use of the method of interlacing polynomials (in Theorem 11).

1.3 Roadmap

While this article is primarily devoted to new results, it also has an expository component due to the fact that we are using tools and ideas from diverse fields, with which some readers may not be familiar. Being very much inspired by Matoušek's exceptionally clear style of mathematical exposition, we also made an effort for the ensuing arguments to be self-contained by including quick explanations of classical results that are being used. It seems impossible to fully achieve a Matoušek-style exposition, but hopefully his influence helped us to make an important area of mathematics and a collection of powerful and versatile tools accessible to a wider audience.

Section 2 describes auxiliary statements that will be used in the subsequent proofs. These include classical results of major importance to several fields, and we include brief deductions of what we need so as to make this article self-contained. Section 3 contains the proof of Lemma 10. A self-contained proof of Theorem 6, using a clever iterative procedure of Giannopoulus [13], appears in Sect. 4. This is followed by Sect. 4.1, where it is shown that Theorem 6 is equivalent to a geometric theorem of Giannopulos [13], thus yielding a shorter (but not self-contained) proof of Theorem 6. Section 5 contains the proof of Theorem 11.

2 Preliminaries

In this section we shall describe several tools that will be used in the ensuing arguments, and derive certain corollaries of them in forms that will be easy to quote as the need arises later.

2.1 A Bit of Linear Algebra

We shall start with elementary linear algebraic reasoning that clarifies the meaning of some of the quantities that were discussed in the Introduction. In particular, we shall see why the identity (12) holds true.

We work here in the setting of Theorem 9, namely we are given $k, m, n \in \mathbb{N}$ and a linear operator $A : \mathbb{R}^m \to \mathbb{R}^n$ satisfying $\mathbf{rank}(A) > k$. We are also fixing any subset $\omega \subseteq \{1, \ldots, m\}$ with $|\omega| = \mathbf{rank}(A)$ such that the vectors $\{Ae_i\}_{i\in\omega} \subseteq \mathbb{R}^n$ are linearly independent. For $j \in \omega$ we consider the linear subspace $F_j \subseteq \mathbb{R}^n$ that is defined in (7), namely F_j is the orthogonal complement of the span of $\{Ae_i\}_{i\in\omega\smallsetminus\{j\}} \subseteq \mathbb{R}^n$. For every $j \in \omega$ define a vector $\mathsf{v}_j \in \mathbb{R}^n$ as follows.

$$\mathsf{v}_j \stackrel{\mathrm{def}}{=} \frac{\mathsf{Proj}_{F_j} Ae_j}{\|\mathsf{Proj}_{F_j} Ae_j\|_2^2} \in \mathbb{R}^n. \tag{15}$$

For every $j \in \omega$, since $I_n - \mathsf{Proj}_{F_j}$ is the orthogonal projection onto $\mathbf{span}(\{Ae_i\}_{i\in\omega\smallsetminus\{j\}}) \subseteq \mathbb{R}^n$, we know that $I_n - \mathsf{Proj}_{F_j} Ae_j \in \mathbf{span}(\{Ae_i\}_{i\in\omega\smallsetminus\{j\}})$. So, $\{\mathsf{Proj}_{F_j} Ae_j\}_{j\in\omega} \subseteq \mathbf{span}(\{Ae_i\}_{i\in\omega})$, and therefore $\{\mathsf{v}_j\}_{j\in\omega} \subseteq \mathbf{span}(\{Ae_i\}_{i\in\omega})$. For $j \in \omega$ we have $\langle \mathsf{Proj}_{F_j} Ae_j, Ae_j\rangle = \|\mathsf{Proj}_{F_j} Ae_j\|_2^2$, so $\langle \mathsf{v}_j, Ae_j\rangle = 1$. Also, because $\mathsf{Proj}_{F_j} Ae_j$ is orthogonal to $\{Ae_i\}_{i\in\omega\smallsetminus\{j\}}$, we have $\langle \mathsf{v}_j, Ae_i\rangle = 0$ for every $i \in \omega \smallsetminus \{j\}$. Since $\{Ae_i\}_{i\in\omega}$ is a basis of $\mathbf{span}(\{Ae_i\}_{i\in\omega})$ and $\{\mathsf{v}_j\}_{j\in\omega} \subseteq \mathbf{span}(\{Ae_i\}_{i\in\omega})$, this means that $\{\mathsf{v}_j\}_{j\in\omega}$ is the *unique* dual basis of $\{Ae_i\}_{i\in\omega}$ in $\mathbf{span}(\{Ae_i\}_{i\in\omega})$.

The operator $(AJ_\omega)^*(AJ_\omega) : \mathbb{R}^\omega \to \mathbb{R}^\omega$ has rank $|\omega| = \mathbf{rank}(A)$, hence it is invertible. For every $j \in \omega$ we may therefore consider the vector

$$\mathsf{w}_j \stackrel{\mathrm{def}}{=} (AJ_\omega)\big((AJ_\omega)^*(AJ_\omega)\big)^{-1} e_j \in \mathbf{span}(\{Ae_i\}_{i\in\omega}).$$

Observe that for every $i, j \in \omega$ we have

$$\begin{aligned}\langle \mathsf{w}_j, Ae_i\rangle &= \Big\langle (AJ_\omega)\big((AJ_\omega)^*(AJ_\omega)\big)^{-1} e_j, (AJ_\omega)e_i\Big\rangle \\ &= \Big\langle (AJ_\omega)^*(AJ_\omega)\big((AJ_\omega)^*(AJ_\omega)\big)^{-1} e_j, e_i\Big\rangle = \langle e_j, e_i\rangle.\end{aligned}$$

By the uniqueness of the dual basis of $\{Ae_i\}_{i\in\omega}$ in $\mathbf{span}(\{Ae_i\}_{i\in\omega})$, we conclude that $\mathsf{v}_j = \mathsf{w}_j$ for every $j \in \omega$. This implies in particular that for every $j \in \omega$ we have

$$\begin{aligned}\frac{1}{\|\mathsf{Proj}_{F_j} Ae_j\|_2^2} &= \|\mathsf{v}_j\|_2^2 = \langle \mathsf{w}_j, \mathsf{w}_j\rangle = \Big\langle (AJ_\omega)\big((AJ_\omega)^*(AJ_\omega)\big)^{-1} e_j, (AJ_\omega)\big((AJ_\omega)^*(AJ_\omega)\big)^{-1} e_j\Big\rangle \\ &= \Big\langle \big((AJ_\omega)^*(AJ_\omega)\big)^{-1} e_j, (AJ_\omega)^*(AJ_\omega)\big((AJ_\omega)^*(AJ_\omega)\big)^{-1} e_j\Big\rangle = \Big\langle \big((AJ_\omega)^*(AJ_\omega)\big)^{-1} e_j, e_j\Big\rangle.\end{aligned} \tag{16}$$

Consequently,

$$\sum_{j\in\omega} \frac{1}{\|\mathsf{Proj}_{F_j} Ae_j\|_2^2} = \sum_{j\in\omega} \Big\langle \big((AJ_\omega)^*(AJ_\omega)\big)^{-1} e_j, e_j\Big\rangle = \mathbf{Tr}\Big(\big((AJ_\omega)^*(AJ_\omega)\big)^{-1}\Big) = \sum_{i=1}^{\mathbf{rank}(A)} \frac{1}{\mathsf{s}_i(AJ_\omega)^2}.$$

This is precisely the identity (12). The above discussion, and in particular the auxiliary vectors (15) and their properties that were derived above, will play a role in later arguments as well.

2.2 Grothendieck

We shall use later the following important theorem of Grothendieck [14].

Theorem 13 (Little Grothendieck Inequality) *Fix $k, m, n \in \mathbb{N}$. Suppose that $T : \mathbb{R}^m \to \mathbb{R}^n$ is a linear operator. Then for every $x_1, \dots, x_k \in \mathbb{R}^m$ there exists $i \in \{1, \dots, m\}$ such that*

$$\sum_{r=1}^{k} \|Tx_r\|_2^2 \leqslant \frac{\pi}{2} \|T\|^2_{\ell_\infty^m \to \ell_2^n} \sum_{r=1}^{k} x_{ri}^2. \tag{17}$$

Here $\|T\|_{\ell_\infty^m \to \ell_2^n} \overset{\mathrm{def}}{=} \max_{x \in [-1,1]^m} \|Tx\|_2$ is the operator norm of T when it is viewed as an operator from ℓ_∞^m to ℓ_2^n, and $x_{ri} = \langle x_r, e_i \rangle$ is the i'th coordinate of $x_r \in \mathbb{R}^m$.

To see the significance of Theorem 13, note that the definition of the operator norm of T when it is viewed as an operator from ℓ_∞^m to ℓ_2^n is nothing more than the smallest $C \geqslant 0$ such that for every $x \in \mathbb{R}^m$ there exists $i \in \{1, \dots, m\}$ for which $\|Tx\|_2^2 \leqslant C^2 x_i^2$. So, the case $k = 1$ of (17) without the factor $\pi/2$ in the right hand side is a tautology. Theorem (13) asserts that the case $k = 1$ of (17) automatically "upgrades" to (17) for general $k \in \mathbb{N}$ at the cost of a loss of the constant factor $\pi/2$.

The literature contains clear expositions of Theorem 13 and its various useful generalizations and equivalent formulations; see e.g. [10, 23]. Nevertheless, for the sake of completeness we shall now quickly explain why Theorem 13 holds true, following (a specialization of) the standard proofs of this fact [10, 23]. We note that the factor $\pi/2$ in (17) is sharp; see e.g. the remark immediately following the proof of Theorem 5.4 in [23].

To prove Theorem 13, by rescaling both T and $(x_1, \dots, x_k)$ we may assume without loss of generality that $\|T\|_{\ell_\infty^m \to \ell_2^n} = 1$ and $\sum_{r=1}^{k} \|Tx_r\|_2^2 = 1$. With this normalization, we claim that

$$\sum_{j=1}^{m} \left(\sum_{r=1}^{k} (T^* T x_r)_j^2 \right)^{\frac{1}{2}} \leqslant \sqrt{\frac{\pi}{2}}. \tag{18}$$

Once proven, (18) implies the desired estimate (17) via the following application of Cauchy–Schwarz.

$$1 = \sum_{r=1}^{k} \|Tx_r\|_2^2 = \sum_{r=1}^{k} \langle x_r, T^*Tx_r\rangle = \sum_{j=1}^{m}\sum_{r=1}^{k} x_{rj}(T^*Tx_r)_j \leqslant \sum_{j=1}^{m}\left(\sum_{r=1}^{k} x_{rj}^2\right)^{\frac{1}{2}}\left(\sum_{r=1}^{k}(T^*Tx_r)_j^2\right)^{\frac{1}{2}}$$

$$\leqslant \max_{i\in\{1,\ldots,m\}}\left(\sum_{r=1}^{k} x_{ri}^2\right)^{\frac{1}{2}}\sum_{j=1}^{m}\left(\sum_{r=1}^{k}(T^*Tx_r)_j^2\right)^{\frac{1}{2}} \overset{(18)}{\leqslant} \sqrt{\frac{\pi}{2}}\cdot \max_{i\in\{1,\ldots,m\}}\left(\sum_{r=1}^{k} x_{ri}^2\right)^{\frac{1}{2}}.$$

To prove (18), let $\{g_r\}_{r=1}^k$ be i.i.d. standard Gaussian random variables. For every $j \in \{1,\ldots,m\}$ the random variable $\sum_{r=1}^k g_r(T^*Tx_r)_j$ is Gaussian with mean 0 and variance $\sum_{r=1}^k (T^*Tx_r)_j^2$. So,

$$\mathbb{E}\left[\sum_{j=1}^{m}\left|\left(T^*\sum_{r=1}^{k} g_rTx_r\right)_j\right|\right] = \mathbb{E}\left[\sum_{j=1}^{m}\left|\sum_{r=1}^{k} g_r(T^*Tx_r)_j\right|\right] = \sum_{j=1}^{m}\mathbb{E}\left[\left|\sum_{r=1}^{k} g_r(T^*Tx_r)_j\right|\right]$$

$$= \mathbb{E}\big[|g_1|\big]\sum_{j=1}^{m}\left(\sum_{r=1}^{k}(T^*Tx_r)_j^2\right)^{\frac{1}{2}} = \sqrt{\frac{2}{\pi}}\sum_{j=1}^{m}\left(\sum_{r=1}^{k}(T^*Tx_r)_j^2\right)^{\frac{1}{2}}. \tag{19}$$

Let $z \in \{-1,1\}^m$ be the random vector given by $z_j \overset{\mathrm{def}}{=} \mathbf{sign}\Big(\big(T^*\sum_{r=1}^k g_rTx_r\big)_j\Big)$. Then

$$\sum_{j=1}^{m}\left|\left(T^*\sum_{r=1}^{k} g_rTx_r\right)_j\right| = \left\langle z, T^*\sum_{r=1}^{k} g_rTx_r\right\rangle = \left\langle Tz, \sum_{r=1}^{k} g_rTx_r\right\rangle$$

$$\leqslant \|Tz\|_2\cdot\left\|\sum_{r=1}^{k} g_rTx_r\right\|_2 \leqslant \|T\|_{\ell_\infty^m\to\ell_2^n}\cdot\|z\|_\infty\cdot\left\|\sum_{r=1}^{k} g_rTx_r\right\|_2 = \left\|\sum_{r=1}^{k} g_rTx_r\right\|_2. \tag{20}$$

By taking expectations in (20) we see that

$$\sqrt{\frac{2}{\pi}}\sum_{j=1}^{m}\left(\sum_{r=1}^{k}(T^*Tx_r)_j^2\right)^{\frac{1}{2}} \overset{(19)}{=} \mathbb{E}\left[\sum_{j=1}^{m}\left|\left(T^*\sum_{r=1}^{k} g_rTx_r\right)_j\right|\right]$$

$$\overset{(20)}{\leqslant} \mathbb{E}\left[\left\|\sum_{r=1}^{k} g_rTx_r\right\|_2\right] \leqslant \left(\mathbb{E}\left[\left\|\sum_{r=1}^{k} g_rTx_r\right\|_2^2\right]\right)^{\frac{1}{2}} = \sum_{r=1}^{k}\|Tx_r\|_2^2 = 1,$$

This is precisely the desired estimate (18), thus completing the proof of Theorem 13. □

2.3 Pietsch

Another classical tool that will be used later (together with the Little Grothendieck Inequality) is the Pietsch Domination Theorem [22].

Theorem 14 (Pietsch Domination) *Fix $m, n \in \mathbb{N}$ and $M \in (0,\infty)$. Suppose that $T : \mathbb{R}^m \to \mathbb{R}^n$ is a linear operator such that for every $k \in \mathbb{N}$ and $x_1,\ldots,x_k \in \mathbb{R}^m$ there exists $i \in \{1,\ldots,m\}$ with $\sum_{r=1}^k \|Tx_r\|_2^2 \leqslant M^2 \sum_{r=1}^k x_{ri}^2$. Then there exist $\mu_1,\ldots,\mu_m \in [0,1]$ with $\sum_{i=1}^m \mu_i = 1$ such that*

$$\forall\, w = (w_1,\ldots,w_m) \in \mathbb{R}^m, \qquad \|Tw\|_2^2 \leqslant M^2 \sum_{i=1}^m \mu_i w_i^2.$$

Observe in passing that the conclusion of Theorem 14 immediately implies its assumption. Indeed, by applying this conclusion with $w = x_r$ for each $r \in \{1,\ldots,k\}$, and then summing the resulting inequalities over $r \in \{1,\ldots,k\}$, we get that $\sum_{r=1}^k \|Tx_r\|_2^2 \leqslant \sum_{i=1}^m \mu_i (M^2 \sum_{r=1}^k x_{ri}^2)$, so the existence of the desired index $i \in \{1,\ldots,m\}$ follows from the fact that $(\mu_1,\ldots,\mu_m)$ is a probability measure. The main point here is therefore the reverse implication, as stated in Theorem 14.

In Banach space theoretic terminology, the assumption on the operator T in Theorem 14 says that T has 2-*summing norm at most M* when it is viewed as an operator from ℓ_∞^m to ℓ_2^n. We refer to the monographs [10, 31] for much more on this topic, as well as proofs of (more general versions of) the Pietsch Domination Theorem. As before, for the sake of completeness we shall now explain why Theorem 14 holds true, following (a specialization of) the standard proofs [10, 31] of this fact, which amount to an application of the separation theorem (equivalently, Hahn–Banach or duality of linear programming) to appropriately chosen convex sets.

Let $\mathsf{K} \subseteq \mathbb{R}^m$ be the set of all those vectors $y \in \mathbb{R}^m$ for which there exists $k \in \mathbb{N}$ and $x_1,\ldots,x_k \in \mathbb{R}^m$ such that $y_i = \sum_{r=1}^k \|Tx_r\|_2^2 - M^2 \sum_{r=1}^k x_{ri}^2$ for every $i \in \{1,\ldots,m\}$. It is immediate to check that K is convex, and the assumption on T can be restated as saying that $\mathsf{K} \cap (0,\infty)^m = \emptyset$. By the separation theorem there exists $\mu = (\mu_1,\ldots,\mu_m) \in \mathbb{R}^m$ such that $\sum_{i=1}^m \mu_i y_i < \sum_{i=1}^m \mu_i z_i$ for every $y \in \mathsf{K}$ and $z \in (0,\infty)^m$. In particular, $\mu \neq 0$ and $\inf_{z\in(0,\infty)^m} \langle z, \mu\rangle > -\infty$, so necessarily $\mu_i \geqslant 0$ for all $i \in \{1,\ldots,m\}$. We may rescale so that $\sum_{i=1}^m \mu_i = 1$. If $w \in \mathbb{R}^m$ then $(\|Tw\|_2^2 - M^2 w_i^2)_{i=1}^m \in \mathsf{K}$, so $\|Tw\|_2^2 - M^2 \sum_{i=1}^m \mu_i w_i^2 = \sum_{i=1}^m \mu_i (\|Tw\|_2^2 - M^2 w_i^2) \leqslant \inf_{z\in(0,\infty)^m} \sum_{i=1}^m \mu_i z_i = 0$. □

The following lemma is a combination of the Little Grothendieck Inequality and the Pietsch Domination Theorem; this is how Theorems 13 and 14 will be used in what follows.

Lemma 15 *Fix $m, n \in \mathbb{N}$ and $\varepsilon \in (0,1)$. Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a linear operator. Then there exists a subset $\sigma \subseteq \{1,\ldots,m\}$ with $|\sigma| \geqslant (1-\varepsilon)m$ such that*

$$\|\mathsf{Proj}_{\mathbb{R}^\sigma} T\|_{\mathsf{S}_\infty} \leqslant \sqrt{\frac{\pi}{2\varepsilon m}} \cdot \|T\|_{\ell_2^n \to \ell_1^m}. \qquad (21)$$

Proof Since we have $\|T^*\|_{\ell_\infty^m\to\ell_2^n} = \|T\|_{\ell_2^n\to\ell_1^m}$, an application of Theorem 13 to $T^* : \mathbb{R}^m \to \mathbb{R}^n$ shows that the assumption of Theorem 14 holds true with T replaced by T^* and $M = \sqrt{\pi/2}\cdot\|T\|_{\ell_2^n\to\ell_1^m}$. Hence, Theorem 14 shows that there exists $\mu \in [0,1]^m$ with $\sum_{i=1}^m \mu_i = 1$ such that

$$\forall\, y \in \mathbb{R}^m, \qquad \|T^*y\|_2^2 \leqslant \frac{\pi}{2}\|T\|^2_{\ell_2^n\to\ell_1^m}\sum_{i=1}^m \mu_i y_i^2. \tag{22}$$

Define

$$\sigma \overset{\mathrm{def}}{=} \left\{ i \in \{1,\ldots,m\} :\ \mu_i \leqslant \frac{1}{m\varepsilon}\right\}. \tag{23}$$

Since μ is a probability measure on $\{1,\ldots,m\}$, by Markov's inequality we have $|\sigma| \geqslant (1-\varepsilon)m$.

Take $x \in \mathbb{R}^n$ and choose $y \in \mathbb{R}^m$ such that $\|y\|_2 = 1$ and $\|\mathsf{Proj}_{\mathbb{R}^\sigma}Tx\|_2 = \langle y, \mathsf{Proj}_{\mathbb{R}^\sigma}Tx\rangle$. Then,

$$\begin{aligned}&\|\mathsf{Proj}_{\mathbb{R}^\sigma}Tx\|_2^2 = \langle y, \mathsf{Proj}_{\mathbb{R}^\sigma}Tx\rangle^2 = \langle T^*\mathsf{Proj}_{\mathbb{R}^\sigma}y, x\rangle^2 \leqslant \|T^*\mathsf{Proj}_{\mathbb{R}^\sigma}y\|_2^2\cdot\|x\|_2^2\\ &\overset{(22)}{\leqslant} \frac{\pi}{2}\|T\|^2_{\ell_2^n\to\ell_1^m}\cdot\|x\|_2^2\sum_{i\in\sigma}\mu_i y_i^2 \overset{(23)}{\leqslant} \frac{\pi}{2m\varepsilon}\|T\|^2_{\ell_2^n\to\ell_1^m}\cdot\|x\|_2^2\cdot\|y\|_2^2 = \frac{\pi}{2m\varepsilon}\|T\|^2_{\ell_2^n\to\ell_1^m}\cdot\|x\|_2^2.\end{aligned} \tag{24}$$

Since (24) holds true for every $x \in \mathbb{R}^n$, this completes the proof of the desired estimate (21). □

2.4 Sauer–Shelah

The Sauer–Shelah lemma [24, 25] is a fundamental combinatorial principle of wide applicability that will be used crucially later.

Lemma 16 (Sauer–Shelah) *Fix $m,n \in \mathbb{N}$. Suppose that $\Omega \subseteq \{-1,1\}^n$ satisfies $|\Omega| > \sum_{k=0}^{m-1}\binom{n}{k}$. Then there exists a subset $\sigma \subseteq \{1,\ldots,n\}$ with $|\sigma| \geqslant m$ such that $\mathsf{Proj}_{\mathbb{R}^\sigma}\Omega = \{-1,1\}^\sigma$, i.e., for every $\varepsilon \in \{-1,1\}^\sigma$ there exists $\delta \in \Omega$ such that $\delta_j = \varepsilon_j$ for every $j \in \sigma$. In particular, if $|\Omega| > 2^{n-1}$ then such a subset $\sigma \subseteq \{1,\ldots,n\}$ exists with $|\sigma| \geqslant \lceil (n+1)/2\rceil \geqslant n/2$.*

It is simple to prove Lemma 16 by induction on n when one strengthens the inductive hypothesis as follows. Denoting $\mathbf{sh}(\Omega) = \Big\{\sigma \subseteq \{1,\ldots,n\} :\ \mathsf{Proj}_{\mathbb{R}^\sigma}\Omega = \{-1,1\}^\sigma\Big\}$, we claim that $|\mathbf{sh}(\Omega)| \geqslant |\Omega|$; this would imply Lemma 16 since the number of subsets of $\{1,\ldots,n\}$ of size at most $m-1$ equals $\sum_{k=0}^{m-1}\binom{n}{k}$. This stronger statement is due to Pajor [21], and the resulting very short inductive proof which we shall now sketch for completeness appears as Theorem 1.1 in [1].

The case $n = 1$ holds trivially (here we use the convention that $\{-1,1\}^{\emptyset} = \emptyset$ and $\mathsf{Proj}_{\mathbb{R}^{\emptyset}}\Omega = \emptyset$). Assuming the validity of the above statement for n, take $\Omega \subseteq \{-1,1\}^{n+1} = \{-1,1\}^n \times \{-1,1\}$ and denote $\Omega_1 = \{x \in \{-1,1\}^n : (x,1) \in \Omega\}$ and $\Omega_{-1} = \{x \in \{-1,1\}^n : (x,-1) \in \Omega\}$. Then $|\Omega_1| + |\Omega_{-1}| = |\Omega|$ and by the inductive hypothesis we have $|\mathbf{sh}(\Omega_1)| \geqslant |\Omega_1|$ and $|\mathbf{sh}(\Omega_{-1})| \geqslant |\Omega_{-1}|$. By our definitions we have $\mathbf{sh}(\Omega) \supseteq (\mathbf{sh}(\Omega_1) \cup \mathbf{sh}(\Omega_{-1})) \uplus \{\sigma \cup \{n+1\} : \sigma \in \mathbf{sh}(\Omega_1) \cap \mathbf{sh}(\Omega_{-1})\}$, so $|\mathbf{sh}(\Omega)| \geqslant |\mathbf{sh}(\Omega_1) \cup \mathbf{sh}(\Omega_{-1})| + |\mathbf{sh}(\Omega_1) \cap \mathbf{sh}(\Omega_{-1})| = |\mathbf{sh}(\Omega_1)| + |\mathbf{sh}(\Omega_{-1})| \geqslant |\Omega_1| + |\Omega_{-1}| = |\Omega|$. □

2.5 Fan and Hilbert–Schmidt

We record for ease of future use the following lemma that controls the influence of multiplication by an orthogonal projection on the Hilbert–Schmidt norm of a linear operator. Its proof is a simple consequence of the classical *Fan Maximum Principle* [11], but we couldn't locate a reference where it is stated explicitly in the form that we will use later.

Lemma 17 *Fix $m, n \in \mathbb{N}$ and $r \in \{1, \ldots, n\}$. Let $A : \mathbb{R}^m \to \mathbb{R}^n$ be a linear operator and let $\mathsf{P} : \mathbb{R}^n \to \mathbb{R}^n$ be an orthogonal projection of rank r. Then*

$$\|\mathsf{P}A\|_{\mathsf{S}_2} \geqslant \Bigg(\sum_{i=n-r+1}^{m} \mathsf{s}_i(A)^2\Bigg)^{\frac{1}{2}}.$$

Proof Since $I_n - \mathsf{P}$ is an orthogonal projection of rank $n - r$, by a classical result of Fan [11],

$$\mathbf{Tr}(AA^*(I_n - \mathsf{P})) \leqslant \sum_{i=1}^{n-r} \mathsf{s}_i(AA^*) = \sum_{i=1}^{n-r} \mathsf{s}_i(A)^2. \tag{25}$$

The proof of (25) is simple; see e.g. [28, Lemma 8.1.8] for a short proof and [4, Chapter III] for more general variational principles along these lines. Now, since P is an orthogonal projection,

$$\begin{aligned}\|\mathsf{P}A\|_{\mathsf{S}_2}^2 &= \mathbf{Tr}((\mathsf{P}A)^*(\mathsf{P}A)) = \mathbf{Tr}(A^*\mathsf{P}A) = \mathbf{Tr}(AA^*\mathsf{P}) = \mathbf{Tr}(AA^*) - \mathbf{Tr}(AA^*(I_n - \mathsf{P}))\\ &= \sum_{i=1}^{m} \mathsf{s}_i(A)^2 - \mathbf{Tr}(AA^*(I_n - \mathsf{P})) \overset{(25)}{\geqslant} \sum_{i=1}^{m} \mathsf{s}_i(A)^2 - \sum_{i=1}^{n-r} \mathsf{s}_i(A)^2 = \sum_{i=n-r+1}^{m} \mathsf{s}_i(A)^2.\end{aligned}$$

□

3 Proof of Lemma 10

In this section, we shall prove Lemma 10 in a more general weighted form that corresponds to the renormalization step in Vershynin's Theorem, i.e., Theorem 2. Using this weighted version of Lemma 10, one can directly deduce weighted versions of Theorems 6 and 8 as well, by combining Lemma 18 below with Theorem 9, exactly as we did in the Introduction.

Lemma 18 (weighted version of Lemma 10) *Fix $r, m, n \in \mathbb{N}$. Let $A : \mathbb{R}^m \to \mathbb{R}^n$ be a linear operator with* $\mathbf{rank}(A) \geqslant r$*. For every $\tau \subseteq \{1, \ldots, m\}$ let $E_\tau \subseteq \mathbb{R}^n$ be defined as in (9), i.e., it is the orthogonal complement of the span of $\{Ae_j\}_{j\in\tau} \subseteq \mathbb{R}^n$. Then for every $d_1, \ldots, d_m \in (0, \infty)$ there exists a subset $\tau \subseteq \{1, \ldots, m\}$ with $|\tau| = r$ such that*

$$\forall j \in \tau, \qquad \left\|\mathsf{Proj}_{E_{\tau\smallsetminus\{j\}}} Ae_j\right\|_2 \geqslant \frac{d_j}{\sqrt{\sum_{i=1}^m d_i^2}} \left(\sum_{i=r}^m \mathsf{s}_i(A)^2\right)^{\frac{1}{2}}. \tag{26}$$

Proof For every $\tau \subseteq \{1, \ldots, m\}$ let $K_\tau \subseteq \mathbb{R}^n$ be the convex hull of the vectors $\{\pm Ae_j/d_j\}_{j\in\tau}$, i.e.,

$$K_\tau \stackrel{\mathrm{def}}{=} \mathbf{conv}\left(\left\{\frac{1}{d_j}Ae_j : j \in \tau\right\} \cup \left\{-\frac{1}{d_j}Ae_j : j \in \tau\right\}\right). \tag{27}$$

The desired subset $\tau \subseteq \{1, \ldots, m\}$ will be chosen so as to maximize the r-dimensional volume of the convex hull of K_σ over all those subsets σ of $\{1, \ldots, m\}$ of size r. Namely, we shall fix from now on a subset $\tau \subseteq \{1, \ldots, m\}$ with $|\tau| = r$ such that

$$\mathbf{vol}_r(K_\tau) = \max_{\substack{\sigma\subseteq\{1,\ldots,m\}\\ |\sigma|=r}} \mathbf{vol}_r(K_\sigma). \tag{28}$$

Take any $\beta \subseteq \{1, \ldots, m\}$ with $|\beta| = r - 1$ and fix $i \in \{1, \ldots, m\} \smallsetminus \beta$. Then by the definition (27) we have $K_{\beta\cup\{i\}} = \mathbf{conv}(\{\pm Ae_i/d_i\} \cup K_\beta)$, i.e., $K_{\beta\cup\{i\}}$ is the union of the two cones with base K_β and apexes at $\pm Ae_i/d_i$. Recalling (9), note that $K_\beta \subseteq \mathbf{span}(K_\beta) = E_\beta^\perp$. Hence, the height of these two cones equals the Euclidean length of the orthogonal projection of Ae_i/d_i onto E_β. Therefore,

$$\mathbf{vol}_r\left(K_{\beta\cup\{i\}}\right) = \frac{2\left\|\mathsf{Proj}_{E_\beta} Ae_i\right\|_2 \mathbf{vol}_{r-1}(K_\beta)}{rd_i}. \tag{29}$$

Returning to the subset τ that was chosen in (28), we see that if $j \in \tau$ and $i \in \{1,\ldots,m\}$ then

$$\frac{2\left\|\mathsf{Proj}_{E_{\tau\smallsetminus\{j\}}}Ae_j\right\|_2\mathbf{vol}_{r-1}\left(K_{\tau\smallsetminus\{j\}}\right)}{rd_j} \overset{(29)}{=} \mathbf{vol}_r(K_\tau)$$
$$\overset{(28)}{\geqslant} \mathbf{vol}_r\left(K_{(\tau\smallsetminus\{j\})\cup\{i\}}\right) \overset{(29)}{=} \frac{2\left\|\mathsf{Proj}_{E_{\tau\smallsetminus\{j\}}}Ae_i\right\|_2\mathbf{vol}_{r-1}\left(K_{\tau\smallsetminus\{j\}}\right)}{rd_i}. \tag{30}$$

Since we are assuming that $r \leqslant \mathbf{rank}(A)$, we know that $\mathbf{vol}_r(K_\tau) > 0$. It therefore follows from (30) that also $\mathbf{vol}_{r-1}\left(K_{\tau\smallsetminus\{j\}}\right) > 0$, so me may cancel the quantity $2\mathbf{vol}_{r-1}\left(K_{\tau\smallsetminus\{j\}}\right)/r$ from both sides of (30). Since the resulting estimate holds true for every $i \in \{1,\ldots,m\}$, we conclude that

$$\forall\, j \in \tau, \qquad \frac{\left\|\mathsf{Proj}_{E_{\tau\smallsetminus\{j\}}}Ae_j\right\|_2}{d_j} = \max_{i\in\{1,\ldots,m\}} \frac{\left\|\mathsf{Proj}_{E_{\tau\smallsetminus\{j\}}}Ae_i\right\|_2}{d_i}. \tag{31}$$

Consequently, for every $j \in \tau$ we have

$$\frac{\left\|\mathsf{Proj}_{E_{\tau\smallsetminus\{j\}}}Ae_j\right\|_2^2}{d_j^2}\left(\sum_{i=1}^m d_i^2\right) \overset{(31)}{\geqslant} \sum_{i=1}^m \left\|\mathsf{Proj}_{E_{\tau\smallsetminus\{j\}}}Ae_i\right\|_2^2 = \left\|\mathsf{Proj}_{E_{\tau\smallsetminus\{j\}}}A\right\|_{\mathsf{S}_2}^2.$$

Equivalently,

$$\forall\, j \in \tau, \qquad \left\|\mathsf{Proj}_{E_{\tau\smallsetminus\{j\}}}Ae_j\right\|_2 \geqslant \frac{d_j}{\sqrt{\sum_{i=1}^m d_i^2}}\left\|\mathsf{Proj}_{E_{\tau\smallsetminus\{j\}}}A\right\|_{\mathsf{S}_2}. \tag{32}$$

Recalling (9), since $|\tau| = r$ we know that $\dim(E_{\tau\smallsetminus\{j\}}) = n-(r-1)$ for every $j \in \tau$. Consequently, $\mathsf{Proj}_{E_{\tau\smallsetminus\{j\}}} : \mathbb{R}^n \to \mathbb{R}^n$ is an orthogonal projection of rank $n-(r-1)$, so that the desired inequality (26) follows from (32) and Lemma 17. □

4 Giannopoulos

In this section we shall prove Theorem 9, following the lines of a clever iterative procedure that was devised by Giannopoulos in [13]. Throughout the ensuing discussion, we may assume in the setting of Theorem 9 that $\omega = \{1,\ldots,m\}$, in which case $\mathbf{rank}(A) = m$. Indeed, there is no loss of generality by doing so because for general $\omega \subseteq \{1,\ldots,m\}$ we could then consider the restricted operator $AJ_\omega : \mathbb{R}^\omega \to \mathbb{R}^n$ in order to obtain Theorem 9 as stated in the Introduction.

Proof overview The overall strategy of the ensuing proof can be explained in broad strokes given the tools that were already presented in Sect. 2. The ultimate goal of Theorem 9 is to obtain an upper bound on the operator norm $\|\cdot\|_{\mathsf{S}_\infty}$ of a certain m by n matrix (the inverse of an appropriate coordinate restriction of the given n by m matrix A), while we have already seen in Lemma 15 that if one does not mind composing with a further coordinate projection then such a bound follows automatically from a weaker upper estimate on the operator norm $\|\cdot\|_{\ell_2^n\to\ell_1^m}$. The latter quantity can be controlled using the Sauer–Shelah lemma due to the following reasoning.

Let $\{\mathsf{v}_j\}_{j=1}^m$ be the dual basis of $\{Ae_j\}_{j=1}^m$ that is given in (15). Consider the subset Ω of the hypercube $\{-1,1\}^m$ consisting of all those sign vectors $\varepsilon=(\varepsilon_1,\dots,\varepsilon_m)$ for which the Euclidean norm $\|\sum_{j=1}^m\varepsilon_j\mathsf{v}_j\|_2$ is not too large, with the precise meaning of "not too large" here to be specified in the proof of Lemma 19 below; see (37). The parallelogram identity says that if $\varepsilon\in\{-1,1\}^m$ is chosen uniformly at random then the expectation of $\|\sum_{j=1}^m\varepsilon_j\mathsf{v}_j\|_2^2$ equals $\sum_{j=1}^m\|\mathsf{v}_j\|_2^2$. So, by Markov's inequality, an appropriate setting of the parameters would yield that the cardinality of Ω is greater than $2^{m-1}=|\{-1,1\}^m|/2$. The Sauer–Shelah lemma would then furnish a coordinate subset $\beta\subseteq\{1,\dots,m\}$ with the property that every sign pattern $(\varepsilon_j)_{j\in\beta}\in\{-1,1\}^\beta$ can be completed to a full dimensional sign vector $\varepsilon\in\{-1,1\}^m$ such that $\sum_{j=1}^m\varepsilon_j\mathsf{v}_j$ is "short" in the Euclidean norm.

The above conclusion implies an upper bound on the operator norm of the inverse of the restriction of A to $\mathbb{R}^\beta$, when it is viewed as an operator from ℓ_2^β to ℓ_1^m. Indeed, given an arbitrary vector $(a_j)_{j\in\beta}\in\mathbb{R}^\beta$, the goal is to bound $\sum_{j\in\beta}|a_j|$ in terms of $\|\sum_{j\in\beta}a_jAe_j\|_2$. The sign pattern to be considered is then the signs of the coefficients $(a_j)_{j\in\beta}\in\mathbb{R}^\beta$, i.e., set $\varepsilon_j=\mathbf{sign}(a_j)$ for every $j\in\beta$. The (Sauer–Shelah) subset $\beta\subseteq\{1,\dots,m\}$ was constructed so that this sign vector can be completed to a full dimensional sign vector $\varepsilon\in\{-1,1\}^m$ with control on the Euclidean length of $\sum_{j=1}^m\varepsilon_j\mathsf{v}_j$. But $\{\mathsf{v}_j\}_{j=1}^m$ is a dual basis of $\{Ae_j\}_{j=1}^m$, so by the definition of $(\varepsilon_j)_{j\in\beta}$ the quantity $\sum_{j\in\beta}|a_j|$ is equal to the scalar product of $\sum_{j\in\beta}a_jAe_j$ with the "short" vector $\sum_{j=1}^m\varepsilon_j\mathsf{v}_j$. By Cauchy–Schwarz this scalar product is bounded from above by the Euclidean length of $\sum_{j\in\beta}a_jAe_j$ times the Euclidean length of $\sum_{j=1}^m\varepsilon_j\mathsf{v}_j$, with the latter quantity being bounded above by design.

By Lemma 15 we can now pass to a further subset of β and compose the resulting inverse matrix with the coordinate projection onto that subset so as to "upgrade" this control on the operator norm from ℓ_2^β to ℓ_1^m to a better upper bound on $\|\cdot\|_{\mathsf{S}_\infty}$. Complications arise when one examines the above strategy from the quantitative perspective. The Sauer–Shelah lemma can at best produce a coordinate subset of size $m/2$, while we desire to obtain restricted invertibility on a potentially larger subset. Moreover, in the above procedure the Sauer–Shelah subset is further reduced in size due to the subsequent use of Lemma 15. Since we desire to extract larger coordinate subsets, one can attempt to apply this reasoning iteratively, i.e., start by using the Sauer–Shelah lemma to obtain a coordinate subset, followed by an application of Lemma 15 to pass to a further subset $\beta'\subseteq\{1,\dots,m\}$. Now apply the same double selection procedure to $\{1,\dots,m\}\smallsetminus\beta'$, thus obtaining a subset $\beta''\subseteq$

$\{1,\ldots,m\}\smallsetminus\beta'$, and iterate this procedure by now considering $\{1,\ldots,m\}\smallsetminus(\beta'\cup\beta'')$ and so forth. To make this strategy work, one needs to formulate a stronger inductive hypothesis so as to allow one to "glue" the local information on the subsets that are extracted in each step of the iteration into global information on their union, while ensuring that the end result is a sufficiently large coordinate subset. This is the reason why the assumptions of Lemma 19 below are more complicated. The technical details that implement the above strategy are explained in the remainder of this section.

Lemma 19 *Fix $n\in\mathbb{N}$ and $m\in\{1,\ldots,n\}$. Let $A:\mathbb{R}^m\to\mathbb{R}^n$ be a linear operator such that the vectors $\{Ae_j\}_{j=1}^m\subseteq\mathbb{R}^n$ are linearly independent. Suppose that $k\in\mathbb{N}\cup\{0\}$ and $\sigma\subseteq\{1,\ldots,m\}$. For $j\in\{1,\ldots,m\}$ recall the definition of the subspace $F_j\subseteq\mathbb{R}^n$ in (7) (with $\omega=\{1,\ldots,m\}$), i.e,*

$$F_j=\left(\mathbf{span}\,\{Ae_i\}_{i\in\{1,\ldots,m\}\smallsetminus\{j\}}\right)^{\perp}.$$

Then there exists $\tau\subseteq\sigma$ with $|\tau|\geqslant(1-2^{-k})|\sigma|$ such that for every $\vartheta\subseteq\{1,\ldots,m\}$ that satisfies $\vartheta\supseteq\tau$ and every $a=(a_1,\ldots,a_m)\in\mathbb{R}^m$ there exists an index $j\in\{1,\ldots,m\}$ for which

$$\sum_{i\in\tau}|a_i|\leqslant\frac{\sqrt{|\sigma|}\sum_{r=1}^k 2^{\frac{r}{2}}}{\|\mathsf{Proj}_{F_j}Ae_j\|_2}\Big\|\sum_{i\in\vartheta}a_iAe_i\Big\|_2+(2^k-1)\sum_{i\in\vartheta\cap(\sigma\smallsetminus\tau)}|a_i|. \tag{33}$$

Proof It will be convenient to introduce the following notation.

$$M\overset{\mathrm{def}}{=}\max_{j\in\{1,\ldots,m\}}\frac{1}{\|\mathsf{Proj}_{F_j}Ae_j\|_2}\qquad\text{and}\qquad\alpha_k\overset{\mathrm{def}}{=}\sum_{r=1}^k 2^{\frac{r}{2}}. \tag{34}$$

Throughout we adhere to the convention that an empty sum vanishes, thus in particular $\alpha_0=0$.

Under the notation (34), our goal becomes to show that there exists $\tau\subseteq\sigma$ with $|\tau|\geqslant(1-2^{-k})|\sigma|$ such that for every $\vartheta\subseteq\{1,\ldots,m\}$ that satisfies $\vartheta\supseteq\tau$ and every $a\in\mathbb{R}^m$ we have

$$\sum_{i\in\tau}|a_i|\leqslant\alpha_k M\sqrt{|\sigma|}\Big\|\sum_{i\in\vartheta}a_iAe_i\Big\|_2+(2^k-1)\sum_{i\in\vartheta\cap(\sigma\smallsetminus\tau)}|a_i|. \tag{35}$$

We shall prove this statement by induction on k. The case $k=0$ holds vacuously by taking $\tau=\emptyset$. Assuming the validity of this statement for k, we shall proceed to deduce its validity for $k+1$.

We are given $\tau\subseteq\sigma$ with $|\tau|\geqslant(1-2^{-k})|\sigma|$ such that for every $\vartheta\subseteq\{1,\ldots,m\}$ that satisfies $\vartheta\supseteq\tau$ we know that (35) holds true for every $a\in\mathbb{R}^m$. Observe that if $\tau=\sigma$ then τ itself would satisfy the required statement for $k+1$, so we may assume from now on that $\sigma\smallsetminus\tau\neq\emptyset$.

For every $j \in \{1, \dots, m\}$ let v_j be given as in (15), i.e.,

$$\mathsf{v}_j \overset{\text{def}}{=} \frac{\mathsf{Proj}_{F_j} A e_j}{\|\mathsf{Proj}_{F_j} A e_j\|_2^2} \in \mathbb{R}^n. \tag{36}$$

Observe that the denominator in (36) (and also in (33) and (34)) does not vanish since we are assuming in Lemma 19 that $\{Ae_j\}_{j=1}^m$ are linearly independent. Define $\Omega \subseteq \{-1,1\}^{\sigma \smallsetminus \tau}$ as follows.

$$\Omega \overset{\text{def}}{=} \left\{ \varepsilon \in \{-1,1\}^{\sigma \smallsetminus \tau} : \left\| \sum_{i \in \sigma \smallsetminus \tau} \varepsilon_i \mathsf{v}_i \right\|_2 \leqslant M\sqrt{2|\sigma \smallsetminus \tau|} \right\}. \tag{37}$$

By the parallelogram identity we have

$$\begin{aligned} M^2|\sigma \smallsetminus \tau| &\overset{(34)}{\geqslant} \sum_{i \in \sigma \smallsetminus \tau} \frac{1}{\|\mathsf{Proj}_{F_i} A e_i\|_2^2} \overset{(36)}{=} \sum_{i \in \sigma \smallsetminus \tau} \|\mathsf{v}_i\|_2^2 = \frac{1}{2^{|\sigma \smallsetminus \tau|}} \sum_{\varepsilon \in \{-1,1\}^{\sigma \smallsetminus \tau}} \left\| \sum_{i \in \sigma \smallsetminus \tau} \varepsilon_i \mathsf{v}_i \right\|_2^2 \\ &\overset{(37)}{>} \frac{1}{2^{|\sigma \smallsetminus \tau|}} \sum_{\substack{\varepsilon \in \{-1,1\}^{\sigma \smallsetminus \tau} \\ \varepsilon \notin \Omega}} 2M^2|\sigma \smallsetminus \tau| = 2M^2|\sigma \smallsetminus \tau| \left(1 - \frac{|\Omega|}{2^{|\sigma \smallsetminus \tau|}}\right). \end{aligned} \tag{38}$$

Since $|\sigma \smallsetminus \tau| > 0$, it follows from (38) that $|\Omega| > 2^{|\sigma \smallsetminus \tau| - 1}$.

We can now apply the Sauer–Shelah lemma, i.e., Lemma 16, thus deducing that there exists a subset $\beta \subseteq \sigma \smallsetminus \tau$ with $|\beta| \geqslant |\sigma \smallsetminus \tau|/2$ such that $\mathsf{Proj}_{\mathbb{R}^\beta} \Omega = \{-1,1\}^\beta$. Defining $\tau^* = \tau \cup \beta$ we shall now proceed to show that τ^* satisfies the inductive hypothesis with k replaced by $k+1$.

Since $\beta \cap \tau = \emptyset$, $\tau \subseteq \sigma$ and $|\beta| \geqslant |\sigma \smallsetminus \tau|/2$ we have

$$|\tau^*| = |\tau| + |\beta| \geqslant |\tau| + \frac{|\sigma| - |\tau|}{2} = \frac{|\tau| + |\sigma|}{2} \geqslant \frac{(1 - 2^{-k})|\sigma| + |\sigma|}{2} = (1 - 2^{-k-1})|\sigma|. \tag{39}$$

Next, suppose that $\vartheta \subseteq \{1, \dots, m\}$ satisfies $\vartheta \supseteq \tau^*$. If $a \in \mathbb{R}^m$ then because $\mathsf{Proj}_{\mathbb{R}^\beta} \Omega = \{-1,1\}^\beta$ there exists $\varepsilon \in \Omega$ such that for every $j \in \beta$ we have $\varepsilon_j = \mathbf{sign}(a_j)$. The fact that $\varepsilon \in \Omega$ means that

$$\left\| \sum_{i \in \sigma \smallsetminus \tau} \varepsilon_i \mathsf{v}_i \right\|_2 \leqslant M\sqrt{2|\sigma \smallsetminus \tau|} \leqslant \frac{M\sqrt{2|\sigma|}}{2^{k/2}}, \tag{40}$$

where in the last step of (40) we used the fact that $|\tau| \geqslant (1 - 2^{-k})|\sigma|$.

The definition (36) of $\{\mathsf{v}_j\}_{j=1}^m$ implies that $\langle \mathsf{v}_i, Ae_j\rangle = \delta_{ij}$ for every $i,j \in \{1,\dots,m\}$. Hence,

$$\sum_{i\in\beta}|a_i| = \Big\langle \sum_{i\in\beta} a_iAe_i, \sum_{i\in\sigma\smallsetminus\tau}\varepsilon_i\mathsf{v}_i\Big\rangle = \Big\langle \sum_{i\in\vartheta} a_iAe_i, \sum_{i\in\sigma\smallsetminus\tau}\varepsilon_i\mathsf{v}_i\Big\rangle - \sum_{i\in(\vartheta\smallsetminus\beta)\cap(\sigma\smallsetminus\tau)}\varepsilon_i a_i$$
$$\leqslant \Big\|\sum_{i\in\vartheta} a_iAe_i\Big\|_2\Big\|\sum_{i\in\sigma\smallsetminus\tau}\varepsilon_i\mathsf{v}_i\Big\|_2 + \sum_{i\in\vartheta\cap(\sigma\smallsetminus\tau^*)}|a_i| \overset{(40)}{\leqslant} \frac{M\sqrt{2|\sigma|}}{2^{k/2}}\Big\|\sum_{i\in\vartheta} a_iAe_i\Big\|_2 + \sum_{i\in\vartheta\cap(\sigma\smallsetminus\tau^*)}|a_i|. \tag{41}$$

The penultimate step of (41) uses the Cauchy–Schwarz inequality and the fact that, by the definition of τ^*, we have $(\vartheta\smallsetminus\beta)\cap(\sigma\smallsetminus\tau) = \vartheta\cap(\sigma\smallsetminus\tau^*)$. Now,

$$\sum_{i\in\tau^*}|a_i| = \sum_{i\in\tau}|a_i| + \sum_{i\in\beta}|a_i| \overset{(35)}{\leqslant} \alpha_k M\sqrt{|\sigma|}\Big\|\sum_{i\in\vartheta}a_iAe_i\Big\|_2 + (2^k-1)\sum_{i\in\vartheta\cap(\sigma\smallsetminus\tau)}|a_i| + \sum_{i\in\beta}|a_i|$$
$$= \alpha_k M\sqrt{|\sigma|}\Big\|\sum_{i\in\vartheta}a_iAe_i\Big\|_2 + (2^k-1)\sum_{i\in\vartheta\cap(\sigma\smallsetminus\tau^*)}|a_i| + 2^k\sum_{i\in\beta}|a_i|, \tag{42}$$

where for the last step of (42) recall that $\vartheta\cap(\sigma\smallsetminus\tau) = (\vartheta\cap(\sigma\smallsetminus\tau^*))\,\dot\cup\,\beta$. It remains to combine (41) and (42) to deduce that

$$\sum_{i\in\tau^*}|a_i| \leqslant \left(\alpha_k + 2^{\frac{k+1}{2}}\right)M\sqrt{|\sigma|}\Big\|\sum_{i\in\vartheta}a_iAe_i\Big\|_2 + (2^{k+1}-1)\sum_{i\in\vartheta\cap(\sigma\smallsetminus\tau^*)}|a_i|. \tag{43}$$

Recalling the definition of α_k in (34), we have $\alpha_{k+1} = \alpha_k + 2^{(k+1)/2}$, so the validity of (39) and (43) completes the proof that τ^* satisfies the inductive hypothesis with k replaced by $k+1$. □

Lemma 20 *Fix $m,n,t\in\mathbb{N}$ and $\beta\subseteq\{1,\dots,m\}$. Let $A:\mathbb{R}^m\to\mathbb{R}^n$ be a linear operator such that the vectors $\{Ae_j\}_{j=1}^m\subseteq\mathbb{R}^n$ are linearly independent. Then there exist two subsets $\sigma,\tau\subseteq\beta$ satisfying $\sigma\subseteq\tau$, $|\tau|\geqslant(1-2^{-t})|\beta|$ and $|\tau\smallsetminus\sigma|\leqslant|\beta|/4$ such that if we denote $\vartheta=\tau\cup(\{1,\dots,m\}\smallsetminus\beta)$ then*

$$\big\|\mathsf{Proj}_{\mathbb{R}^\sigma}(AJ_\vartheta)^{-1}\big\|_{\mathsf{S}_\infty} \lesssim \max_{j\in\{1,\dots,m\}}\frac{2^{\frac{t}{2}}}{\|\mathsf{Proj}_{F_j}Ae_j\|_2},$$

where we recall that the definition of the subspace $F_j\subseteq\mathbb{R}^n$ is given in (7).

Proof An application of Lemma 19 with $\sigma=\beta$ and $k=t$ produces $\tau\subseteq\beta$ with $|\tau|\geqslant(1-2^{-t})|\beta|$ such that if we choose $\vartheta=\tau\cup(\{1,\dots,m\}\smallsetminus\beta)$ in (33) and continue with the notation in (34) then

$$\forall\, a \in \mathbb{R}^m, \qquad \sum_{i\in\tau} |a_i| \lesssim 2^{\frac{t}{2}} M\sqrt{|\beta|}\left\| \sum_{i\in\vartheta} a_i A e_i \right\|_2. \tag{44}$$

Note that the above choice of ϑ makes the second term in the right hand side of (33) vanish, and this is the only way by which (33) will be used here. However, the more complicated form of (33) was needed in Lemma 19 to allow for the inductive construction to go through.

A different way to state (44) is the following operator norm bound.

$$\left\| \mathsf{Proj}_{\mathbb{R}^\tau}(AJ_\vartheta)^{-1} \right\|_{\ell_2^\vartheta \to \ell_1^\tau} \lesssim 2^{\frac{t}{2}} M\sqrt{|\beta|}.$$

Since $|\tau| \geqslant (1-2^{-t})|\beta| \geqslant |\beta|/2$, if we set $\varepsilon \stackrel{\mathrm{def}}{=} |\beta|/(4|\tau|)$ then $\varepsilon \in (0,1/2)$. We are therefore in position to use Lemma 15, thus producing a subset $\sigma \subseteq \tau$ with $|\tau \smallsetminus \sigma| \leqslant \varepsilon|\tau| = |\beta|/4$ such that

$$\left\| \mathsf{Proj}_{\mathbb{R}^\sigma}(AJ_\vartheta)^{-1} \right\|_{\mathsf{S}_\infty} = \left\| \mathsf{Proj}_{\mathbb{R}^\sigma}\mathsf{Proj}_{\mathbb{R}^\tau}(AJ_\vartheta)^{-1} \right\|_{\mathsf{S}_\infty} \lesssim \frac{2^{\frac{t}{2}} M\sqrt{|\beta|}}{\sqrt{\varepsilon|\tau|}} \asymp 2^{\frac{t}{2}} M.$$

□

Proof of Theorem 9 Recall that, in the setting of Theorem 9, we are currently assuming without loss of generality that $\omega = \{1,\ldots,m\}$. Choose $r \in \mathbb{N}\cup\{0\}$ such that

$$\frac{1}{2^{2r+1}} \leqslant 1 - \frac{k}{m} \leqslant \frac{1}{2^{2r-1}}. \tag{45}$$

Denote $\tau_0 \stackrel{\mathrm{def}}{=} \{1,\ldots,m\}$ and $\sigma_0 \stackrel{\mathrm{def}}{=} \emptyset$. We shall construct by induction on $u \in \{0,\ldots,r+1\}$ two subsets $\sigma_u, \tau_u \subseteq \{1,\ldots,m\}$ such that if we denote

$$\beta_u \stackrel{\mathrm{def}}{=} \tau_u \smallsetminus \sigma_u \qquad \text{and} \qquad \forall\, u \in \{1,\ldots,r+1\}, \qquad \vartheta_u \stackrel{\mathrm{def}}{=} \tau_u \cup (\{1,\ldots,m\} \smallsetminus \beta_{u-1}), \tag{46}$$

then the following properties hold true for every $u \in \{1,\ldots,r+1\}$.

(a) $\sigma_u \subseteq \tau_u \subseteq \beta_{u-1}$.
(b) $|\tau_u| \geqslant (1-2^{-2r+u-4})|\beta_{u-1}|$ and $|\beta_u| \leqslant \frac{1}{4}|\beta_{u-1}|$.
(c) $\left\| \mathsf{Proj}_{\mathbb{R}^{\sigma_u}}(AJ_{\vartheta_u})^{-1} \right\|_{\mathsf{S}_\infty} \lesssim 2^{r-\frac{u}{2}} M$, where M is defined in (34).

Indeed, assuming inductively that σ_{u-1}, τ_{u-1} have been constructed, the existence of sets σ_u, τ_u with the desired properties follows from an application of Lemma 20 with $\beta = \beta_{u-1}$ and $t = 2r - u + 4$.

Recalling (46), by (a) we have $\beta_{u-1} = \beta_u \mathbin{\dot\cup} \sigma_u \mathbin{\dot\cup} (\beta_{u-1} \smallsetminus \tau_u)$ for every $u \in \{1, \ldots, r+1\}$. Hence,

$$|\sigma_u| = |\beta_{u-1}| - |\beta_u| - |\beta_{u-1} \smallsetminus \tau_u| \geqslant |\beta_{u-1}| - |\beta_u| - \frac{|\beta_{u-1}|}{2^{2r-u+4}}$$

$$\geqslant |\beta_{u-1}| - |\beta_u| - \frac{m}{2^{2r+u+2}}, \tag{47}$$

where the penultimate inequality in (47) uses the first assertion in (b) and the final inequality in (47) uses the fact that, by induction, the second assertion in (b) implies that $|\beta_{u-1}| \leqslant m/4^{u-1}$, since $\beta_0 = \{1, \ldots, m\}$. Observe that the sets $\{\sigma_u\}_{u=1}^{r+1}$ are pairwise disjoint, so if we denote

$$\sigma \stackrel{\mathrm{def}}{=} \bigcup_{u=1}^{r+1}{}^{\!\!\cdot}\ \sigma_u, \tag{48}$$

then

$$|\sigma| = \sum_{u=1}^{r+1} |\sigma_u| \overset{(47)}{\geqslant} |\beta_0| - |\beta_{r+1}| - \frac{m}{2^{2r+2}} \sum_{u=1}^{\infty} \frac{1}{2^u} \geqslant m - \frac{m}{4^{r+1}} - \frac{m}{2^{2r+2}} = m - \frac{m}{2^{2r+1}} \overset{(45)}{\geqslant} k. \tag{49}$$

Next, recalling the definition of ϑ_u in (46), observe that

$$\sigma \subseteq \bigcap_{u=1}^{r+1} \vartheta_u. \tag{50}$$

Indeed, in order to verify the validity of (50) note that due to (a) we have $\sigma_u, \sigma_{u+1}, \ldots, \sigma_{r+1} \subseteq \tau_u$ and $\sigma_1, \ldots, \sigma_{u-1} \subseteq \{1, \ldots, m\} \smallsetminus \beta_{u-1}$ for every $u \in \{1, \ldots, r+1\}$. It follows from (50) that if $a \in \mathbb{R}^\sigma$ then for every $u \in \{1, \ldots, r+1\}$ we have $J_\sigma a \in J_{\vartheta_u} \mathbb{R}^{\vartheta_u} \subseteq \mathbb{R}^m$. Consequently,

$$\mathsf{Proj}_{\mathbb{R}^{\sigma_u}} (AJ_{\vartheta_u})^{-1} (AJ_\sigma) a = \mathsf{Proj}_{\mathbb{R}^{\sigma_u}} J_\sigma a. \tag{51}$$

We therefore have the following estimate.

$$\|J_\sigma a\|_2^2 \overset{(48)}{=} \Big\| \sum_{u=1}^{r+1} \mathsf{Proj}_{\mathbb{R}^{\sigma_u}} J_\sigma a \Big\|_2^2 = \sum_{u=1}^{r+1} \big\|\mathsf{Proj}_{\mathbb{R}^{\sigma_u}} J_\sigma a\big\|_2^2 \overset{(51)}{=} \sum_{u=1}^{r+1} \big\|\mathsf{Proj}_{\mathbb{R}^{\sigma_u}} (AJ_{\vartheta_u})^{-1} (AJ_\sigma) a\big\|_2^2$$

$$\overset{(c)}{\lesssim} \sum_{u=1}^{r+1} 2^{2r-u} M^2 \, \|(AJ_\sigma)a\|_2^2 \asymp 2^{2r} M^2 \, \|(AJ_\sigma)a\|_2^2 \overset{(45)}{\asymp} \frac{mM^2}{m-k} \, \|(AJ_\sigma)a\|_2^2 . \tag{52}$$

Recalling the definition of M in (34), since (52) holds true for every $a \in \mathbb{R}^\sigma$ we conclude that

$$\big\|(AJ_\sigma)^{-1}\big\|_{\mathsf{S}_\infty} \lesssim \frac{\sqrt{m}}{\sqrt{m-k}} \cdot \max_{j \in \{1,\ldots,m\}} \frac{1}{\|\mathsf{Proj}_{F_j} A e_j\|_2}.$$

This is the desired estimate (8), which, together with (49), concludes the proof of Theorem 9. □

4.1 Geometric Interpretation of Theorem 9

Theorem 21 below is a result of Giannopoulos [13]. It can be viewed as a geometric analogue of the Sauer–Shelah lemma for ellipsoids. The (rough) analogy between the two results is that they both assert that certain "large" subsets of $\mathbb{R}^n$ must admit a large rank coordinate projection that contains a certain "canonical shape" (a full hypercube in the Sauer–Shelah case and a large Euclidean ball in Giannopoulos' case). A different geometric analogue of the Sauer–Shelah lemma was proved by Szarek and Talagrand in [30].

Theorem 21 (Giannopoulos) *There exists a universal constant $c \in (0,\infty)$ with the following property. Suppose that $m,n \in \mathbb{N}$ and $\varepsilon \in (0,1)$. Let $y_1,\dots,y_m \in \mathbb{R}^n$ be vectors that satisfy $\|y_i\|_2 \leqslant 1$ for every $i \in \{1,\dots,m\}$. Denote*

$$\mathcal{E} \stackrel{\mathrm{def}}{=} \left\{a=(a_1,\dots,a_m)\in\mathbb{R}^m;\ \left\|\sum_{j=1}^m a_jy_j\right\|_2 \leqslant 1\right\}. \tag{53}$$

Then there exists a subset $\sigma \subseteq \{1,\dots,m\}$ with $|\sigma| \geqslant (1-\varepsilon)m$ such that $\mathsf{Proj}_{\mathbb{R}^\sigma}(\mathcal{E}) \supseteq c\sqrt{\varepsilon}B_2^\sigma$, where $B_2^\sigma = \{x\in\mathbb{R}^\sigma:\ \|x\|_2\leqslant 1\}$ denotes the unit Euclidean ball in $\mathbb{R}^\sigma$.
In this section we shall show that Theorem 21 is equivalent to Theorem 9, thus in particular describing a shorter proof of Theorem 9 that relies on Theorem 21.

Let us first prove that Theorem 9 implies Theorem 21. Suppose that we are in the setting that is described in the statement of Theorem 21. It was observed in [13] that Theorem 21 with the additional assumption that $y_1,\dots,y_m$ are linearly independent formally implies Theorem 21 in the above stated generality. Indeed, this follows by applying (the linear independent case of) Theorem 21 to the linearly independent vectors $y_1+e_{n+1}, y_2+e_{n+2},\dots,y_m+e_{n+m}\in\mathbb{R}^{n+m}$. So, suppose that $y_1,\dots,y_m\in\mathbb{R}^n$ are linearly independent and let $x_1,\dots,x_m\in \mathbf{span}\{y_1,\dots,y_m\}$ be the corresponding dual basis, i.e.,

$$\forall\, i,j\in\{1,\dots,m\},\qquad \langle x_i,y_j\rangle=\delta_{ij}. \tag{54}$$

Define a linear operator $A:\mathbb{R}^m\to\mathbb{R}^n$ by setting $Ae_i=x_i$ for every $i\in\{1,\dots,m\}$. Continuing with the notation for the subspace $F_j\subseteq\mathbb{R}^n$ that is given in (7) (with $\omega=\{1,\dots,m\}$), we know by (54) that $y_j\in F_j$, so $\langle \mathsf{Proj}_{F_j}x_j,y_j\rangle=\langle x_j,y_j\rangle=1$. Since we are assuming in the setting of Theorem 21 that $\|y_j\|_2\leqslant 1$, this implies that $1=\langle \mathsf{Proj}_{F_j}x_j,y_j\rangle\leqslant\|y_j\|_2\cdot\|\mathsf{Proj}_{F_j}x_j\|_2\leqslant\|\mathsf{Proj}_{F_j}x_j\|_2$.

An application of Theorem 9 now shows that there exists $\sigma \subseteq \{1,\ldots,m\}$ with $|\sigma| \geqslant \lfloor (1-\varepsilon)m \rfloor$ and a universal constant $c \in (0,\infty)$ such that

$$\forall\, b \in \mathbb{R}^\sigma, \qquad \Big\| \sum_{j\in\sigma} b_j x_j \Big\|_2 \geqslant c\sqrt{\varepsilon}\Big(\sum_{j\in\sigma} b_j^2\Big)^{\frac12}. \tag{55}$$

We claim that (55) implies that $\mathsf{Proj}_{\mathbb{R}^\sigma}(\mathcal{E}) \supseteq c\sqrt{\varepsilon}B_2^\sigma$, where $\mathcal{E}$ is given in (53). Indeed, suppose that $a = \sum_{j\in\sigma} a_j e_j \in \mathbb{R}^\sigma$ satisfies

$$a \in c\sqrt{\varepsilon}B_2^\sigma \iff \Big(\sum_{j\in\sigma} a_j^2\Big)^{\frac12} \leqslant c\sqrt{\varepsilon}. \tag{56}$$

Since the vectors $\{x_j\}_{j\in\sigma} \cup \{y_j\}_{j\in\{1,\ldots,m\}\smallsetminus\sigma}$ form a basis of $\mathbf{span}\{y_1,\ldots,y_m\}$, there exists a vector $b=(b_1,\ldots,b_m)\in\mathbb{R}^m$ such that

$$\sum_{j\in\sigma} a_j y_j = \sum_{j\in\sigma} b_j x_j + \sum_{j\in\{1,\ldots,m\}\smallsetminus\sigma} b_j y_j. \tag{57}$$

Denote

$$a^* = (a_1^*,\ldots,a_m^*) \stackrel{\mathrm{def}}{=} \sum_{j\in\sigma} a_j e_j - \sum_{j\in\{1,\ldots,m\}\smallsetminus\sigma} b_j e_j \in \mathbb{R}^m. \tag{58}$$

Then $\mathsf{Proj}_{\mathbb{R}^\sigma} a^* = a$ and

$$\begin{aligned}\Big\|\sum_{j=1}^m a_j^* y_j\Big\|_2^2 &= \Big\langle \sum_{j=1}^m a_j^* y_j, \sum_{j=1}^m a_j^* y_j\Big\rangle \stackrel{(57)\wedge(58)}{=} \Big\langle \sum_{j=1}^m a_j^* y_j, \sum_{j\in\sigma} b_j x_j\Big\rangle \stackrel{(54)\wedge(58)}{=} \sum_{j\in\sigma} a_j b_j \\ &\leqslant \Big(\sum_{j\in\sigma} a_j^2\Big)^{\frac12}\Big(\sum_{j\in\sigma} b_j^2\Big)^{\frac12} \stackrel{(56)}{\leqslant} c\sqrt{\varepsilon}\Big(\sum_{j\in\sigma} b_j^2\Big)^{\frac12} \stackrel{(55)}{\leqslant} \Big\|\sum_{j\in\sigma} b_j x_j\Big\|_2 \stackrel{(57)\wedge(58)}{=} \Big\|\sum_{j=1}^m a_j^* y_j\Big\|_2.\end{aligned} \tag{59}$$

By cancelling $\big\|\sum_{j=1}^m a_j^* y_j\big\|_2$ from both sides of (59) and recalling (53), we conclude that $a^* \in \mathcal{E}$. Thus $a = \mathsf{Proj}_{\mathbb{R}^\sigma} a^* \in \mathsf{Proj}_{\mathbb{R}^\sigma}(\mathcal{E})$, as required.

Next, we shall prove the converse implication, i.e., that Theorem 21 implies Theorem 9. Suppose that we are in the setting of Theorem 9. As we explained in the beginning of Sect. 4, we may assume without loss of generality that $\omega = \{1,\ldots,m\}$, hence $\mathbf{rank}(A) = m$. Let $M \in (0,\infty)$ be defined as in (34), i.e., $M = \max_{j\in\{1,\ldots,m\}} \|\mathsf{Proj}_{F_j} Ae_j\|_2^{-1}$. Set

$$\forall\, i \in \{1,\ldots,m\}, \qquad y_i \stackrel{\mathrm{def}}{=} \frac{\mathsf{Proj}_{F_i} Ae_i}{\|\mathsf{Proj}_{F_i} Ae_i\|_2} \in \mathbb{R}^n.$$

Then by definition $\|y_i\|_2 = 1$ for every $j \in \{1, \dots, m\}$, and, by the same reasoning as in the beginning of Sect. 2.1, we know that $\langle y_j, Ae_j\rangle \geqslant 1/M$ and $\langle y_i, Ae_j\rangle = 0$ for every distinct $i, j \in \{1, \dots, m\}$. By Theorem 21 applied with $\varepsilon = 1-k/m$ there exists $\sigma \subseteq \{1, \dots, m\}$ of size $|\sigma| \geqslant (1-\varepsilon)m = k$ such that $\mathsf{Proj}_{\mathbb{R}^\sigma}(\mathcal{E}) \supseteq c\sqrt{\varepsilon}B_2^\sigma$, where $\mathcal{E}$ is defined in (53). Suppose that $a \in \mathbb{R}^\sigma \smallsetminus \{0\}$. Then $c\sqrt{\varepsilon}a/\|a\|_2 \in \mathsf{Proj}_{\mathbb{R}^\sigma}(\mathcal{E})$, which means that there exists $b \in \mathbb{R}^m$ such that $b_j = c\sqrt{\varepsilon}a_j/\|a\|_2$ for every $j \in \sigma$ and (by the definition of $\mathcal{E}$) we have $\|\sum_{i=1}^m b_i y_i\|_2 \leqslant 1$. So,

$$\Big\|\sum_{j\in\sigma} a_j Ae_j\Big\|_2 \geqslant \Big\|\sum_{j\in\sigma} a_j Ae_j\Big\|_2 \cdot \Big\|\sum_{j=1}^m b_j y_j\Big\|_2 \geqslant \Big\langle \sum_{j\in\sigma} a_j Ae_j, \sum_{j=1}^m b_j y_j \Big\rangle$$
$$= \sum_{j\in\sigma} a_j b_j \langle Ae_j, y_j\rangle = \sum_{j\in\sigma} \frac{c\sqrt{\varepsilon}a_j^2}{\|a\|_2}\langle Ae_j, y_j\rangle \geqslant \frac{c\sqrt{\varepsilon}}{M\|a\|_2}\sum_{j\in\sigma} a_j^2 = \frac{c\sqrt{m-k}}{M\sqrt{m}}\|a\|_2.$$

This is precisely the desired conclusion in Theorem 9. □

5 Marcus–Spielman–Srivastava

Our goal here is to prove Theorem 11. This section differs from the previous sections in that we shall use the method of interlacing polynomials of Marcus–Spielman–Srivastava without sketching the proofs of the tools that we quote. The reason for this is that the ideas of Marcus–Spielman–Srivastava are remarkable and deep, but nevertheless elementary and accessible, and their presentation in [17, 18] and especially in the beautiful survey [16] (which is the main reference in the present section) is already a perfect exposition for a wide mathematical audience.

Suppose that $A : \mathbb{R}^m \to \mathbb{R}^n$ is a linear operator. Let $\mathsf{j}_1, \dots, \mathsf{j}_k$ be i.i.d. random variables that are distributed uniformly over $\{1, \dots, m\}$. For every $t \in \{1, \dots, k\}$ consider the random vector

$$\mathsf{w}_t \stackrel{\mathrm{def}}{=} \sqrt{m}Ae_{\mathsf{j}_t}. \tag{60}$$

Then,

$$\mathbb{E}\big[\mathsf{w}_t \otimes \mathsf{w}_t\big] = \sum_{i=1}^m (Ae_i) \otimes (Ae_i) = AA^*. \tag{61}$$

Denote

$$\gamma \stackrel{\mathrm{def}}{=} \frac{\mathbf{rank}(A)\left(\sqrt{\mathbf{rank}(A)} - \sqrt{k}\right)^2}{\sum_{i=1}^{\mathbf{rank}(A)} \frac{1}{\mathsf{s}_i(A)^2}}. \tag{62}$$

With this notation, we shall prove below that

$$\Pr\left[\mathsf{s}_k\bigg(\sum_{t=1}^{k}\mathsf{w}_t\otimes\mathsf{w}_t\bigg)\geqslant\gamma\right]>0. \tag{63}$$

Recalling (60), we see that (62) and (63) imply that there exist $j_1,\ldots,j_k\in\{1,\ldots,m\}$ such that

$$\mathsf{s}_k\bigg(\sum_{t=1}^{k}(Ae_{j_t})\otimes(Ae_{j_t})\bigg)\geqslant\frac{\gamma}{m}=\frac{\mathbf{rank}(A)\left(\sqrt{\mathbf{rank}(A)}-\sqrt{k}\right)^2}{m\sum_{i=1}^{\mathbf{rank}(A)}\frac{1}{\mathsf{s}_i(A)^2}}. \tag{64}$$

The rank of the operator $B\overset{\mathrm{def}}{=}\sum_{t=1}^{k}(Ae_{j_t})\otimes(Ae_{j_t})$ is at most the cardinality of $\sigma\overset{\mathrm{def}}{=}\{j_1,\ldots,j_k\}$. At the same time, by (64) we know that $\mathsf{s}_k(B)>0$, because we are assuming that $k<\mathbf{rank}(A)$. Thus B has rank at least k, implying that the indices $j_1,\ldots,j_k$ are necessarily distinct, or equivalently that $|\sigma|=k$. Consequently $B=(AJ_\sigma)(AJ_\sigma)^*$ and $\mathsf{s}_k(B)=\mathsf{s}_{\min}(B)=\mathsf{s}_{\min}(AJ_\sigma)^2=1/\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty}^2$. Therefore (64) is the same as the desired restricted invertibility statement (11) of Theorem 11.

It remains to establish the validity of (63). Denote $Q\overset{\mathrm{def}}{=}AA^*:\mathbb{R}^n\to\mathbb{R}^n$ and let $\mathsf{q}:\mathbb{R}\to\mathbb{R}$ be the polynomial that is defined as follows.

$$\forall\, x\in\mathbb{R},\qquad \mathsf{q}(x)\overset{\mathrm{def}}{=}(I-\partial_y)^k\mathbf{det}(xI_n+yQ)\big|_{y=0},$$

where I denotes the identity operator on the space of polynomials and ∂_y is the differentiation operator with respect to the variable y (and, as before, I_n is the n by n identity matrix). By Theorem 4.1 in [16], the degree n polynomial q is the expectation of the characteristic polynomial of the random matrix $\sum_{t=1}^{k}\mathsf{w}_t\otimes\mathsf{w}_t$. By Theorem 4.5 in [16], all the roots of q are real, and we denote their decreasing rearrangement by $\rho_1\geqslant\rho_2\geqslant\ldots\geqslant\rho_n$. Thus, ρ_k is the k'th largest root of q. A combination of Theorem 1.7 in [16] and Theorem 4.1 in [16] shows that

$$\Pr\left[\mathsf{s}_k\bigg(\sum_{t=1}^{k}\mathsf{w}_t\otimes\mathsf{w}_t\bigg)\geqslant\rho_k\right]>0. \tag{65}$$

Consequently, in order to prove (63) it suffices to prove that $\rho_k\geqslant\gamma$, where γ is defined in (62).

Write $Q=U\Delta U^{-1}$, where $U:\mathbb{R}^n\to\mathbb{R}^n$ is an orthogonal matrix and $\Delta:\mathbb{R}^n\to\mathbb{R}^n$ is a diagonal matrix whose diagonal equals $(\mathsf{s}_1(A)^2,\ldots,\mathsf{s}_n(A)^2)\in\mathbb{R}^n$. Then for every $x,y\in\mathbb{R}$ we have

$$\mathbf{det}(xI_n+yQ)=\mathbf{det}\left(U(xI_n+y\Delta)U^{-1}\right)=\prod_{i=1}^{n}\left(x+y\mathsf{s}_i(A)^2\right)=x^{n-\mathbf{rank}(A)}\prod_{i=1}^{\mathbf{rank}(A)}\left(x+y\mathsf{s}_i(A)^2\right),$$

where we used the fact that $\mathsf{s}_i(A) = 0$ when $i > \mathbf{rank}(A)$. Consequently,

$$\mathsf{q}(x) = x^{n-\mathbf{rank}(A)}(I - \partial_y)^k \prod_{i=1}^{\mathbf{rank}(A)} \left(x + y\mathsf{s}_i(A)^2\right)\Big|_{y=0}. \tag{66}$$

We claim that if we denote by D the differentiation operator on the space of polynomials then

$$\mathsf{q}(x) = x^{n-k} \prod_{i=1}^{\mathbf{rank}(A)} \left(I - \mathsf{s}_i(A)^2\mathsf{D}\right) x^k. \tag{67}$$

The identity (67) is proven in the special case $\mathsf{s}_1(A) = \ldots = \mathsf{s}_{\mathbf{rank}(A)}(A) = 1$ in [16]. The validity of (67) in full generality follows from checking that the coefficients of the polynomials that appear in the right hand sides of (66) and (67) are equal to each other. Indeed, starting with (66),

$$\begin{aligned} &x^{n-\mathbf{rank}(A)}(I - \partial_y)^k \prod_{i=1}^{\mathbf{rank}(A)} \left(x + y\mathsf{s}_i(A)^2\right)\Big|_{y=0} \\ &\qquad = x^{n-\mathbf{rank}(A)} \sum_{u=0}^{k} \binom{k}{u} (-1)^u \partial_y^u \sum_{\Omega \subseteq \{1,\ldots,\mathbf{rank}(A)\}} x^{\mathbf{rank}(A)-|\Omega|} y^{|\Omega|} \prod_{i\in\Omega} \mathsf{s}_i(A)^2\Big|_{y=0} \\ &\qquad = \sum_{\substack{\Omega \subseteq \{1,\ldots,\mathbf{rank}(A)\} \\ |\Omega| \leqslant k}} \frac{(-1)^{|\Omega|} x^{n-|\Omega|} k!}{(k-|\Omega|)!} \prod_{i\in\Omega} \mathsf{s}_i(A)^2, \end{aligned} \tag{68}$$

since $\partial_y^u y^{|\Omega|}|_{y=0} = |\Omega|! \cdot \mathbf{1}_{\{|\Omega|=u\}}$ for every $(u, \Omega) \in \{0, \ldots, k\} \times \{1, \ldots, \mathbf{rank}(A)\}$. At the same time,

$$x^{n-k} \prod_{i=1}^{\mathbf{rank}(A)} \left(I - \mathsf{s}_i(A)^2\mathsf{D}\right) x^k = x^{n-k} \sum_{\Omega \subseteq \{1,\ldots,\mathbf{rank}(A)\}} (-1)^{|\Omega|} \left(\prod_{i\in\Omega} \mathsf{s}_i(A)^2\right) \mathsf{D}^{|\Omega|} x^k. \tag{69}$$

Since for every for every $(u, \Omega) \in \{0, \ldots, k\} \times \{1, \ldots, \mathbf{rank}(A)\}$ we have $\mathsf{D}^{|\Omega|}x^k = 0$ if $|\Omega| > k$ and $\mathsf{D}^{|\Omega|}x^k = x^{k-|\Omega|}k!/(k - |\Omega|)!$ if $|\Omega| \leqslant k$, the validity of (67) follows by comparing (68) and (69).

Having established the identity (67), we shall proceed to prove the desired estimate $\rho_k \geqslant \gamma$ by applying the barrier method of [2], reasoning along the lines of the argument that is presented in [16]. Following [2, 27], given a polynomial $f : \mathbb{R} \to \mathbb{R}$ and $\phi \in (0, \infty)$ we consider the corresponding "soft spectral edge" $\mathbf{smin}_\phi(f) \in \mathbb{R}$, which is defined as follows

$$\mathbf{smin}_\phi(f) \stackrel{\mathrm{def}}{=} \inf\left\{b \in \mathbb{R} : f'(b) = -\phi f(b)\right\}. \tag{70}$$

As explained in [16, Section 3.2], it is simple to check that for every $\phi \in (0,\infty)$ the smallest real root of f is at least the quantity $\mathbf{smin}_\phi(f)$. Hence, if we define

$$g(x) \stackrel{\mathrm{def}}{=} \prod_{i=1}^{\mathbf{rank}(A)} \big(I - \mathsf{s}_i(A)^2\mathsf{D}\big)\, x^k, \tag{71}$$

then it follows from the above discussion and the identity (67) that it suffices to prove that

$$\sup_{\phi\in(0,\infty)} \mathbf{smin}_\phi(g) \geqslant \gamma. \tag{72}$$

Indeed, by (67) the n real roots of q consist of 0 with multiplicity $n-k$ and also the k roots of g (which are therefore necessarily real). Since g has degree k, the validity of (72) would imply that the smallest root of g is at least $\gamma > 0$, so the k'th largest root of q would be at least γ as well.

To prove (72), recall that Lemma 3.8 of [16] asserts that for every polynomial $f:\mathbb{R}\to\mathbb{R}$ all of whose roots are real, and for every $\phi \in (0,\infty)$, we have

$$\mathbf{smin}_\phi\big((I-\mathsf{D})f\big) \geqslant \mathbf{smin}_\phi(f) + \frac{1}{1+\phi}. \tag{73}$$

For $\mathsf{s} \in (0,\infty)$ define $f_\mathsf{s}:\mathbb{R}\to\mathbb{R}$ by setting $f_\mathsf{s}(x) \stackrel{\mathrm{def}}{=} f(\mathsf{s}x)$ for every $x\in\mathbb{R}$. Observe that

$$\forall\, \mathsf{s}\in(0,\infty),\qquad (I-\mathsf{sD})f = ((I-\mathsf{D})f_\mathsf{s})_{1/\mathsf{s}} \qquad \text{and} \qquad \mathbf{smin}_\phi(f_\mathsf{s}) \stackrel{(70)}{=} \frac{\mathbf{smin}_{\phi/\mathsf{s}}(f)}{\mathsf{s}}. \tag{74}$$

Consequently, for every real-rooted polynomial f and every $\mathsf{s},\phi\in(0,\infty)$ we have

$$\begin{aligned}\mathbf{smin}_\phi\big((I-\mathsf{sD})f\big) &\stackrel{(74)}{=} \mathbf{smin}_\phi\big(((I-\mathsf{D})f_\mathsf{s})_{1/\mathsf{s}}\big) \stackrel{(74)}{=} \mathsf{s}\cdot\mathbf{smin}_{\mathsf{s}\phi}\big((I-\mathsf{D})f_\mathsf{s}\big)\\ &\stackrel{(73)}{\geqslant} \mathsf{s}\left(\mathbf{smin}_{\mathsf{s}\phi}(f_\mathsf{s}) + \frac{1}{1+\mathsf{s}\phi}\right) \stackrel{(74)}{=} \mathbf{smin}_\phi(f) + \frac{1}{\frac{1}{\mathsf{s}}+\phi}.\end{aligned} \tag{75}$$

By iterating (75) we see that

$$\begin{aligned}\mathbf{smin}_\phi(g) &\geqslant \mathbf{smin}_\phi(x^k) + \sum_{i=1}^{\mathbf{rank}(A)} \frac{1}{\frac{1}{\mathsf{s}_i(A)^2}+\phi}\\ &\stackrel{(70)}{=} -\frac{k}{\phi} + \sum_{i=1}^{\mathbf{rank}(A)} \frac{1}{\frac{1}{\mathsf{s}_i(A)^2}+\phi} \geqslant -\frac{k}{\phi} + \frac{\mathbf{rank}(A)}{\phi + \frac{1}{\mathbf{rank}(A)}\sum_{i=1}^{\mathbf{rank}(A)}\frac{1}{\mathsf{s}_i(A)^2}},\end{aligned} \tag{76}$$

where the last step of (76) holds true due to the convexity of the function $x \mapsto 1/(\phi + x)$ on $(0, \infty)$. One can check that the value of ϕ that maximizes the right hand side of (76) is

$$\phi_{\max} \stackrel{\text{def}}{=} \frac{\sqrt{k}}{\sqrt{\mathbf{rank}(A)} - \sqrt{k}} \left(\frac{1}{\mathbf{rank}(A)} \sum_{i=1}^{\mathbf{rank}(A)} \frac{1}{\mathsf{s}_i(A)^2} \right).$$

The right hand side of (76) equals γ when $\phi = \phi_{\max}$, so $\rho_k \geqslant \mathbf{smin}_{\phi_{\max}}(g) \geqslant \gamma$, as required. □

Remark 22 The above argument actually yields a subset $\sigma \subseteq \{1, \ldots, m\}$ with $|\sigma| = k$ such that

$$\mathsf{s}_{\min}(AJ_\sigma)^2 = \mathsf{s}_k(AJ_\sigma)^2 \geqslant \frac{1}{m} \sup \left\{ -\frac{k}{\phi} + \sum_{i=1}^{\mathbf{rank}(A)} \frac{\mathsf{s}_i(A)^2}{1 + \phi \mathsf{s}_i(A)^2} : \phi \in (0, \infty) \right\}. \tag{77}$$

Indeed, continuing with the above notation, we explained why $\rho_k \geqslant \sup_{\phi \in (0,\infty)} \mathbf{smin}_\phi(g)$, so (77) follows from (65) and the penultimate step in (76).

The estimate (77) is more complicated than the assertion of Theorem 11, but it is sometimes significantly stronger. One such instance is the matrix A of Example 7. In that case, a somewhat tedious but straightforward computation allows one to obtain sharp estimates on the right hand side of (77), yielding bounds that coincide (up to constant factors) with those that are stated in Example 7 as a consequence of Theorem 6, while Theorem 11 yields much weaker bounds. There are also situations in which (77) yields worse bounds than those that follow from Theorem 9, e.g. when $\mathsf{s}_1(A) \asymp \ldots \asymp \mathsf{s}_m(A) \asymp 1$ and $k = (1-\varepsilon)m$ the bound on $\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty}$ that follows from (77) is $O(1/\varepsilon)$ while in the same situation Theorem 9 yields the bound $\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} \lesssim 1/\sqrt{\varepsilon}$.

5.1 Added in Proof

It turns out that the following statement is a formal consequence of Theorem 9. Fix $k, m, n \in \mathbb{N}$ with $k < m$ and write $k = (1-\varepsilon)m$ for some $\varepsilon \in (0,1)$. Let $A : \mathbb{R}^m \to \mathbb{R}^n$ be a linear operator of rank m. Then there exists a subset $\sigma \subseteq \{1, \ldots, m\}$ with $|\sigma| = k$ such that

$$\|(AJ_\sigma)^{-1}\|_{\mathsf{S}_\infty} \lesssim \sqrt{\frac{\log\left(\frac{2}{\varepsilon}\right)}{\varepsilon}} \left(\frac{1}{m} \sum_{j=1}^{m} \frac{1}{\|\mathsf{Proj}_{F_j} Ae_j\|_2^2} \right)^{\frac{1}{2}},$$

where $F_j = (\{Ae_i\}_{i \in \{1,\ldots,n\} \setminus \{j\}})^\perp$. This follows by grouping the vectors $\{e_j\}_{j=1}^m$ according to the consecutive powers of 2 between which each of the numbers

$\left\{\|\mathsf{Proj}_{F_j} Ae_j\|_2^2\right\}_{j=1}^m$ lies, and applying Theorem 9 to each of these groups; the full details of the (short) derivation of this statement are omitted and will appear elsewhere, where further applications will be explored. Note that this almost answers the question that we posed in the paragraph that immediately follows (14), up to the term $\sqrt{\log(2/\varepsilon)}$. This also improves over (14) and shows that the method of interlacing polynomials is not needed for any of the restricted invertibility theorems that are stated here (and, in fact, the method of interlacing polynomials yields inferior results).

Acknowledgements We thank Bill Johnson for helpful discussions. This work was initiated while we were participating in the workshop *Beyond Kadison–Singer: paving and consequences* at the American Institute of Mathematics. We thank the organizers for the excellent working conditions.

References

1. R. P. Anstee, L. Rónyai, A. Sali, Shattering news. Graphs Combin. **18**(1), 59–73 (2002)
2. J. Batson, D. A. Spielman, N. Srivastava, Twice-Ramanujan sparsifiers. SIAM J. Comput. **41**(6), 1704–1721 (2012)
3. K. Berman, H. Halpern, V. Kaftal, G. Weiss, Matrix norm inequalities and the relative Dixmier property. Integr. Equ. Oper. Theory **11**(1), 28–48 (1988)
4. R. Bhatia, in *Matrix Analysis*. Graduate Texts in Mathematics, vol. 169 (Springer, New York, 1997). ISBN:0-387-94846-5. doi:10.1007/978-1-4612-0653-8
5. J. Bourgain, L. Tzafriri, Invertibility of "large" submatrices with applications to the geometry of Banach spaces and harmonic analysis. Isr. J. Math. **57**(2), 137–224 (1987)
6. J. Bourgain, L. Tzafriri, On a problem of Kadison and Singer. J. Reine Angew. Math. **420**, 1–43 (1991)
7. J. Bourgain, S. J. Szarek, The Banach-Mazur distance to the cube and the Dvoretzky–Rogers factorization. Isr. J. Math. **62**(2), 169–180 (1988)
8. J. Bourgain, L. Tzafriri, Restricted invertibility of matrices and applications, in *Analysis at Urbana, Volume II, Urbana, 1986–1987*. London Mathematical Society Lecture Note Series, vol. 138 (Cambridge University Press, Cambridge, 1989), pp. 61–107
9. P. J. Davis, in *Circulant Matrices*. Pure and Applied Mathematics (Wiley, New York/Chichester/Brisbane, 1979). ISBN:0-471-05771-1. A Wiley-Interscience Publication
10. J. Diestel, H. Jarchow, A. Tonge, in *Absolutely Summing Operators*. Cambridge Studies in Advanced Mathematics, vol. 43 (Cambridge University Press, Cambridge, 1995). ISBN:0-521-43168-9. doi:10.1017/CBO9780511526138
11. K. Fan, On a theorem of Weyl concerning eigenvalues of linear transformations. I. Proc. Natl. Acad. Sci. U. S. A. **35**, 652–655 (1949)
12. A.A. Giannopoulos, A note on the Banach–Mazur distance to the cube, in *Geometric Aspects of Functional Analysis, Israel, 1992–1994*. Operator Theory: Advances and Applications, vol. 77 (Birkhäuser, Basel, 1995), pp. 67–73
13. A. A. Giannopoulos, A proportional Dvoretzky-Rogers factorization result. Proc. Am. Math. Soc. **124**(1), 233–241 (1996)
14. A. Grothendieck, Résumé de la théorie métrique des produits tensoriels topologiques. Bol. Soc. Mat. São Paulo **8**, 1–79 (1953)
15. I. Kra, S. R. Simanca, On circulant matrices. Not. Am. Math. Soc. **59**(3), 368–377 (2012)
16. A. W. Marcus, D. A. Spielman, N. Srivastava, Ramanujan graphs and the solution of the Kadison–Singer problem, in *Proceedings of the 2014 International Congress of Mathematicians, Volume III* (2014), pp. 363–386. Available at http://www.icm2014.org/en/vod/proceedings

17. A. W. Marcus, D. A. Spielman, N. Srivastava, Interlacing families I: Bipartite Ramanujan graphs of all degrees. Ann. Math. (2) **182**(1), 307–325 (2015)
18. A. W. Marcus, D. A. Spielman, N. Srivastava, Interlacing families II: mixed characteristic polynomials and the Kadison-Singer problem. Ann. Math. (2) **182**(1), 327–350 (2015)
19. A. W. Marcus, D. A. Spielman, N. Srivastava, Interlacing families III: improved restricted invertibility estimates (2016, in preparation)
20. A. Naor, Sparse quadratic forms and their geometric applications [following Batson, Spielman and Srivastava]. Astérisque, (348): Exp. No. 1033, viii, 189–217 (2012). Séminaire Bourbaki: vol. 2010/2011. Exposés 1027–1042
21. A. Pajor, *Sous-espaces l_1^ndes espaces de Banach*. Travaux en Cours [Works in Progress], vol. 16 (Hermann, Paris, 1985). ISBN:2–7056-6021–6. With an introduction by Gilles Pisier
22. A. Pietsch, Absolut p-summierende Abbildungen in normierten Räumen. Stud. Math. **28**, 333–353 (1966/1967)
23. G. Pisier, in *Factorization of Linear Operators and Geometry of Banach Spaces*. CBMS Regional Conference Series in Mathematics. Published for the Conference Board of the Mathematical Sciences, Washington, DC, vol. 60 (American Mathematical Society, Providence, 1986). ISBN:0-8218-0710-2
24. N. Sauer, On the density of families of sets. J. Comb. Theory Ser. A **13**, 145–147 (1972)
25. S. Shelah, A combinatorial problem; stability and order for models and theories in infinitary languages. Pac. J. Math. **41**, 247–261 (1972)
26. D. A. Spielman, N. Srivastava, An elementary proof of the restricted invertibility theorem. Isr. J. Math. **190**, 83–91 (2012)
27. N. Srivastava, R. Vershynin, Covariance estimation for distributions with $2+\varepsilon$ moments. Ann. Probab. **41**(5), 3081–3111 (2013)
28. E. Størmer, in *Positive Linear Maps of Operator Algebras*. Springer Monographs in Mathematics (Springer, Heidelberg, 2013). ISBN:978-3-642-34368-1; 978-3-642-34369-8. doi:10.1007/978-3-642-34369-8
29. S. J. Szarek, On the geometry of the Banach-Mazur compactum, in *Functional Analysis (Austin, TX, 1987/1989)*, Lecture Notes in Mathematics, vol. 1470 (Springer, Berlin, 1991), pp. 48–59. doi:10.1007/BFb0090211
30. S.J. Szarek, M. Talagrand, An “isomorphic” version of the Sauer-Shelah lemma and the Banach-Mazur distance to the cube, in *Geometric Aspects of Functional Analysis (1987–1988)*. Lecture Notes in Mathematics, vol. 1376 (Springer, Berlin, 1989), pp. 105–112. doi:10.1007/BFb0090050
31. N. Tomczak-Jaegermann, *Banach-Mazur Distances and Finite-Dimensional Operator Ideals*. Pitman Monographs and Surveys in Pure and Applied Mathematics, vol. 38 (Longman Scientific & Technical, Harlow; co-published in the United States with John Wiley & Sons, Inc., New York, 1989). ISBN:0-582-01374-7
32. J.A. Tropp, Column subset selection, matrix factorization, and eigenvalue optimization, in *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms* (SIAM, Philadelphia, 2009), pp. 978–986
33. R. Vershynin, John’s decompositions: selecting a large part. Isr. J. Math. **122**, 253–277 (2001)
34. P. Youssef, Restricted invertibility and the Banach-Mazur distance to the cube. Mathematika **60**(1), 201–218 (2014)

Rational Polygons: Odd Compression Ratio and Odd Plane Coverings

Rom Pinchasi and Yuri Rabinovich

Abstract Let P be a polygon with rational vertices in the plane. We show that for any finite odd-sized collection of translates of P, the area of the set of points lying in an odd number of these translates is bounded away from 0 by a constant depending on P alone.

The key ingredient of the proof is a construction of an odd cover of the plane by translates of P. That is, we establish a family $\mathcal{F}$ of translates of P covering (almost) every point in the plane a uniformly bounded odd number of times.

1 Introduction

The starting point of this research is the following isoperimetric-type problem about translates of compact sets in $\mathbb{R}^d$:

Let $X \subset \mathbb{R}^d$ be a compact set, and let $Z \subset \mathbb{R}^d$ be a finite set of odd cardinality. Consider the finite odd-sized collection $\mathcal{F} = \{X + z\}_{z \in Z}$ of translates of X. Let *$U \subset \mathbb{R}$ be the set of all points that belong to an odd number of the members of $\mathcal{F}$. How small can be the Lebesgue measure of U in terms of the Euclidean measure of X?*

Denoting the infimum of this value by $Vol_{odd}(X)$, called the *odd volume* of X, we define the *odd compression ratio* of X as $\alpha^\circ(X) = Vol_{odd}(X) \,/\, Vol(X)$, where $Vol(X)$ is the Euclidean volume of X. Observe that $\alpha^\circ(X) \leq 1$, as $\mathcal{F}$ may consist of a single element X. Clearly, $\alpha^\circ(X)$ is an affine invariant.

It was observed by the second author about a decade ago that α° of a unit d-cube Q^d is 1. Indeed, informally, consider $\mathbb{R}^d$ under the action (i.e., translation) of $\mathbb{Z}^d$.

Supported by ISF grant (grant No. 409/16).

R. Pinchasi
Mathematics Department, Technion–Israel Institute of Technology, 32000 Haifa, Israel
e-mail: room@math.technion.ac.il

Y. Rabinovich (✉)
Department of Computer Science, Haifa University, Haifa, Israel
e-mail: yuri@cs.haifa.ac.il

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_28

The unit cube (with parts of its boundary removed) is a fundamental domain of $\mathbb{R}^d/\mathbb{Z}^d$. The quotient map $\phi : \mathbb{R}^d \to Q^d$ maps any translate of Q^d onto Q^d in a one-to-one manner. Moreover, the quotient map satisfies

$$\phi\left(\bigoplus_{T\in\mathcal{F}} T\right) = \bigoplus_{T\in\mathcal{F}} \phi(T),$$

where $\bigoplus$ denotes the set-theoretic union modulo 2, i.e., the set of all points covered by an odd number of the members of $\mathcal{F}$. Since the quotient map is locally volume preserving, it is globally volume-nondecreasing, and so one concludes that the volume of $\bigoplus_{T\in\mathcal{F}} T$ is at least that of $\phi(\bigoplus_{T\in\mathcal{F}} T) = \bigoplus_{T\in\mathcal{F}} \phi(T) = \bigoplus_{T\in\mathcal{F}} Q^d = Q^d$.

A similar argument shows that α° of a centrally symmetric planar hexagon is 1 as well. But what about other sets, i.e., a triangle? The second author vividly remembers discussing this question with Jirka Matoušek in a pleasant cafe at Malá Strana, laughing that they are too old for Olympiad-type problems... [1]

The value of α^0 of the triangle (recall that any two triangles are affinely equivalent) was determined by the first author in [1]; it is $\frac{1}{2}$.

Next significant progress on the problem was obtained in [2]. It was shown there that for a union of two disjoint intervals of length 1 on a line with a certain irrational distance between them, the odd compression ratio is 0. The proof uses some algebra of polynomials, and Diophantine approximation. The construction easily extends to higher dimensions. In addition, [2] introduced a technique for obtaining lower bounds on $\alpha^\circ(X)$, and used it to show that for X's that are unions of finitely many cells of the 2-dimensional grid, $\alpha^\circ(X) > 0$.

In the present paper we further develop the technique of [2], and use it to prove that for any planar rational polygon P, the odd compression ratio $\alpha^\circ(P)$ is bounded away from 0 by some positive constant explicitly defined in terms of P. In fact, the statement applies to any compact planar figure with piecewise linear boundary, and (finitely many) rational vertices. In view of the above mentioned result from [2], the assumption of rationality cannot in general be dropped.

Perhaps more importantly, the value of $\alpha^\circ(X)$ is related here to the value of some other natural geometric invariant of X. The other invariant is $\theta^\circ(X)$, the smallest possible *average density* in an *odd cover* of $\mathbb{R}^2$ by a family $\mathcal{F}$ of translates of X. By odd cover we mean that every point $p \in \mathbb{R}^2$, with a possible exception of a measure 0 set, is covered by the members of $\mathcal{F}$ an odd and uniformly bounded number of times.

While [2] does not directly consider odd covers of $\mathbb{R}^2$, it still implies that $\alpha^\circ(X) \geq \theta^\circ(X)^{-1}$. We include here two complete proofs of this useful inequality.

The existence of odd covers of $\mathbb{R}^2$ by translates of a rational polygon is by no means obvious. Most of the present paper is dedicated to constructing such covers. We are aware of no related results in the literature.

[1] A discrete version of the problem about the translates of a square in $\mathbb{R}^2$ had indeed found its way into a mathematical olympiad [3].

While many of the results and constructions presented here can be easily extended to higher dimensions, some essential parts resist simple generalization, and more work is required in order to understand the situation there.

To conclude the Introduction, we hope that the present paper will somewhat elucidate the meaning of the odd compression ratio $\alpha^\circ(X)$, and that the odd covers introduced here will prove worthy of further study.

2 Preliminaries

2.1 Two Basic Operators

In what follows, we shall extensively use the following two operators on subsets of $\mathbb{R}^2$: $\oplus$ and $\mathring{+}$. Let us briefly discuss them here.

The first operator $\oplus$ is the set-theoretic union modulo 2. Given a family $\mathcal{F}$ of subsets of $\mathbb{R}^2$ so that any $p \in \mathbb{R}^2$ is covered at most finitely many times by $\mathcal{F}$, $\bigoplus_{X\in\mathcal{F}} X$ is the set of all points of $\mathbb{R}^2$ covered by an odd number of members in $\mathcal{F}$. Observe that $\oplus$ is commutative and associative.

The second operator, $\mathring{+}$, is less standard. It is the Minkowski sum modulo 2:

$$X \mathring{+} Z \;=\; \bigoplus_{x\in X,\, z\in Z} x+z \;=\; \bigoplus_{z\in Z}(X+z)\,,$$

where $X+z$ denotes the translate of X by z. I.e., $a \in X \mathring{+} Z$ if and only if the number of representations of a of the form $a = x + z$ is an odd natural number. Unlike the Minkowski sum, $X \mathring{+} Z$ is well defined only when every $a \in \mathbb{R}^2$ has at most finitely many representations of the form $x + z$ as above. This requirement is met, e.g., when Z is finite, or when Z is a discrete set of points at distance $\geq \epsilon > 0$ from each other, and X is bounded. Since the Minkowski sum extends to any finite number of sets, and it is commutative and associative, the same holds for $\mathring{+}$ (provided, as before, that every a has finitely many representations).

Moreover, the following distributive law holds. Let $\mathcal{G}$ be a family of sets in $\mathbb{R}^2$, and let $S \subset \mathbb{R}^2$. Assume that the family of sets $\{Y + s\}_{Y\in\mathcal{G},\, s\in S}$ covers any point of $\mathbb{R}^2$ at most finitely many times. Then:

$$\left(\bigoplus_{Y\in\mathcal{G}} Y\right) \mathring{+}\, S \;=\; \bigoplus_{Y\in\mathcal{G}}(Y \mathring{+} S)\,. \tag{1}$$

Indeed, the equality is trivial when S consists of a single element. Thus, by definition of $\mathring{+}$,

$$\left(\bigoplus_{Y\in\mathcal{G}} Y\right) \mathring{+} S = \bigoplus_{s\in S}\left(\left(\bigoplus_{Y\in\mathcal{G}} Y\right) + s\right) = \bigoplus_{s\in S}\bigoplus_{Y\in\mathcal{G}}(Y+s) \overset{*}{=} \bigoplus_{Y\in\mathcal{G}}\bigoplus_{s\in S}(Y+s)$$
$$= \bigoplus_{Y\in\mathcal{G}}(Y \mathring{+} S)\,.$$

It remains to validate the change of order of summation in the starred equality. For $a \in \mathbb{R}^2$ consider the set $\{(Y,s) \,|\, a \in Y+s\} \subseteq \mathcal{G}\times S$. By our assumptions, this set is always finite. Therefore, for any a, $\mathbf{1}_a\big(\bigoplus_{s\in S}\bigoplus_{Y\in\mathcal{G}}(Y+s)\big) = \bigoplus_{s\in S}\bigoplus_{Y\in\mathcal{G}}\mathbf{1}_a(Y+s)$ has only finitely many nonzero terms. Hence, the order of summation in the double sum $\bigoplus_{s\in S}\bigoplus_{Y\in\mathcal{G}}(Y+s)$ is interchangeable.

Finally, notice that similarly to Minkowski sum, $X \mathring{+} \emptyset = \emptyset$, while $X \oplus \emptyset = X$.

2.2 Covers and Their Densities

It is important to stress that throughout this paper whenever we speak on covers or odd covers of the plane it always means covering **up to a set of measure** 0, *even if it is not explicitly said so. This convention helps to avoid discussing unnecessary technicalities related to the boundaries of the sets in the cover.*

For every compact measurable set $X \subset \mathbb{R}^2$, we denote by $A(X)$ the Euclidean area of X. Let $Z \subseteq \mathbb{R}^2$ be a discrete set. The family $\mathcal{F} = \{X+z\}_{z\in Z}$ has a *uniformly bounded degree* if there exists a constant $d_{\mathcal{F}}$ such that every $a \in \mathbb{R}^2$ belongs to at most $d_{\mathcal{F}}$ members of $\mathcal{F}$. Further, such $\mathcal{F}$ is called a *cover* of $\mathbb{R}^2$ if $X+Z=\mathbb{R}^2$. I.e., the cover degree of any $a \in \mathbb{R}^2$ by the members of a cover $\mathcal{F}$ is uniformly bounded, and, up to a set of measure 0, it is strictly positive.

The (lower) *density* of $\mathcal{F}$ with a uniformly bounded degree, $\rho(\mathcal{F})$, is defined by

$$\rho(\mathcal{F}) = \liminf_{n\to\infty} \frac{\sum_{z\in Z} A(Q_n \cap (X+z))}{n^2},$$

where Q_n is the $n\times n$ square centered at the origin. Clearly, $\rho(\mathcal{F}) \geq 1$ when $\mathcal{F}$ is a cover or $\mathbb{R}^2$.

Since $\sum_{z\in Z} A(Q_n \cap X+z)/n^2$ is precisely the average of the cover degrees $d_{\mathcal{F}}(a)$ where a ranges over Q_n, the density $\rho(\mathcal{F})$ can be viewed as a kind of an average degree of the cover of $\mathbb{R}^2$ by $\mathcal{F}$.

Claim 2.1 *Fixing Z and varying the (measurable) X, the density of the family $\mathcal{F} = \{X+z\}_{z\in Z}$ is proportional to $A(X)$. I.e., $\rho(\mathcal{F}) = c_Z \cdot A(X)$, where c_Z is a constant depending solely on Z. (In particular, when Z is a lattice, c_Z is the reciprocal of the area of the fundamental domain of Z.)*

Proof (Sketch) Assume w.l.o.g., that X contains the origin, and let $\delta = Diam(X)$. Consider $\Delta_n = \left| \sum_{z \in Z} A(Q_n \cap X + z) - |Z \cap Q_n| \cdot A(X) \right|$. How big can it be? On the one hand,

$$|Z \cap Q_{n-2\delta}| \cdot A(X) \;\leq\; \sum_{z \in Z} A(Q_n \cap X + z) \;\leq\; |Z \cap Q_{n+2\delta}| \cdot A(X)\,,$$

and therefore

$$\Delta_n \;\leq\; |Z \cap Q_{n+2\delta}| \cdot A(X) - |Z \cap Q_{n-2\delta}| \cdot A(X) \;=\; |Z \,\cap\, (Q_{n+2\delta} \setminus Q_{n-2\delta})| \cdot A(X)\,.$$

On the other hand, since $Z \,\cap\, (Q_{n+2\delta} \setminus Q_{n-2\delta}) \,+\, X$ is contained in $Q_{n+4\delta} \setminus Q_{n-4\delta}$, covering no point there more than $d_{\mathcal{F}}$ times, it follows that $|Z \,\cap\, (Q_{n+2\delta} \setminus Q_{n-2\delta})| \cdot A(X)$ is at most $O(n) \cdot \delta \cdot d_{\mathcal{F}}$. Hence, $\Delta_n = O(n) \cdot \delta \cdot d_{\mathcal{F}}$, and so $\Delta_n / n^2 \to 0$. The conclusion follows:

$$\rho(\mathcal{F}) \;=\; \liminf_{n \to \infty} \frac{\sum_{z \in Z} A(Q_n \cap X + z)}{n^2} \;=\; \liminf_{n \to \infty} \frac{|Z \cap Q_n| \cdot A(X) \pm \Delta_n}{n^2} \;=$$

$$\liminf_{n \to \infty} \frac{|Z \cap Q_n|}{n^2} \cdot A(X) \;=\; c_Z \cdot A(X)\,.$$

The fact that for a lattice Z, $\lim_{n \to \infty} |Z \cap Q_n| / n^2$ is the inverse of the the area of the fundamental domain of Z, is well known (see, e.g., [4]). □

The *covering density* of X, $\theta(X) \geq 1$, is defined as the infimum of $\rho(\mathcal{F})$ over all covers of the form $\mathcal{F} = \{X + z\}_{z \in Z}$. It is well known (see, e.g., [4]) that $\theta(X)$ is an affine invariant.

2.3 Odd Covers

Let $X \subset \mathbb{R}^2$ be a compact set of a positive area $A(X) > 0$. The family $\mathcal{F} = \{X + z\}_{z \in Z}$ for $Z \subseteq \mathbb{R}^2$ is called an *odd cover* of $\mathbb{R}^2$ if $X \mathring{+} Z$ is well defined, and equals to $\mathbb{R}^2$ up to a set of measure 0. Notice that if $\mathcal{F}$ is an odd cover of $\mathbb{R}^2$, then in particular it is a cover of $\mathbb{R}^2$. As before, we shall further require that the maximal degree of the cover of $\mathbb{R}^2$ by $\mathcal{F}$ is uniformly bounded.

The *odd covering density* of a compact X, $\theta^\circ(X) \geq 1$ is defined as the infimum of $\rho(\mathcal{F})$ over all odd covers $\mathcal{F}$ as above. If no such $\mathcal{F}$ exists, set $\theta^\circ(X) = \infty$. Notice that $\theta^\circ(X) \geq \theta(X)$. Similarly to the usual covering density $\theta(X)$, the odd covering density $\theta^\circ(X)$ is an affine invariant. This intuitively plausible statement can be proved formally along the same lines as the standard proof of the corresponding statement for the usual covers (see, e.g., [4]).[2]

2.4 Odd Compression Ratio: The Definition

Let $X \subset \mathbb{R}^2$ be a compact set of area $0 < A(X) < \infty$. Define $A_{odd}(X)$, the *odd area* of X, to be the maximum number such that for any finite and odd-sized collection $\mathcal{F}$ of translates of X, the set of all points in $\mathbb{R}^2$ belonging to an odd number of members of $\mathcal{F}$ has area $\geq A_{odd}(X)$. I.e., $A_{odd}(X)$ is the infimum of $A(X \overset{\circ}{+} K)$ over all finite odd-sized sets $K \subset \mathbb{R}^2$ (see [1, 2]).

Define $\alpha^\circ(X)$, the *odd compression ratio* of X, as $A_{odd}(X)/A(X)$. Clearly, $0 \leq \alpha^\circ(X) \leq 1$, and it is an affine invariant.

3 The Odd Cover Lemma

The following lemma, a variant, and in fact a special case, of Lemma 1 from [2], is a useful tool for obtaining lower bounds on the odd compression ratio of X. For completeness, we provide two different proofs for it. The first is shorter and simpler due to the preparation done in Sect. 2.2. It is a streamlined variant of the proof used in [2]. The second proof follows a somewhat different logic, and can be viewed as a generalization of the factor-space argument mentioned in the Introduction.

Lemma 3.1 *For any compact set X of a positive measure in $\mathbb{R}^2$, the odd compression ratio of X is at least the reciprocal of its odd covering density. That is,*

$$\alpha^\circ(X) \;\geq\; \theta^\circ(X)^{-1}\,.$$

Proof (A) Let $\mathcal{F} = \{X+z\}_{z\in Z}$ be an odd cover of $\mathbb{R}^2$ of density $\rho(\mathcal{F})$, and maximal cover degree $d_{\mathcal{F}} < \infty$. (If no such $\mathcal{F}$ exists, the lemma is trivially true.) Let $K \subset \mathbb{R}^2$ be any finite set of odd cardinality. Set $Y = X \overset{\circ}{+} K$.

Consider the set $(X \overset{\circ}{+} Z) \overset{\circ}{+} K$. On the one hand, it is equal to $\mathbb{R}^2$, up to a set of measure 0. This is because $(X \overset{\circ}{+} Z) = \mathbb{R}^2$, again up to a set of measure 0, and the cardinality of K is odd.

[2] The requirement that $\mathcal{F}$ has a uniformly bounded degree does not appear in the standard definition of $\theta(X)$, despite the fact that it is used in the proof of the affine invariance of $\theta(X)$ and elsewhere. The reason is that for any $\epsilon > 0$, a cover $\mathcal{F}$ can be easily modified into a *periodic* cover $\mathcal{F}'$ with $\rho(\mathcal{F}') \leq \rho(\mathcal{F}) + \epsilon$, i.e., the corresponding Z' is of the form $\Lambda + K$, where Λ is a lattice, and K is finite (see, e.g., [4]). Thus, w.l.o.g., one may restrict the discussion of $\theta(X)$ to periodic covers, and those are always uniformly bounded for a compact X. In contrast, the odd covers apparently do not allow such a modification, and so the assumption about the uniformly bounded degree seems to be essential for them. This said, all odd covers occurring in this paper are periodic.

On the other hand, using the commutativity of $\mathring{+}$, one concludes that $(X \mathring{+} Z) \mathring{+} K = (X \mathring{+} K) \mathring{+} Z = Y \mathring{+} Z$. In other words, the family $\mathcal{G} = \{Y + z\}_{z \in Z}$ is an odd cover of $\mathbb{R}^2$ of a maximal covering degree at most $d_{\mathcal{F}} \cdot |K|$.

By Claim 2.1, there is a constant c_Z depending only on Z, such that for every measurable set $W \subset \mathbb{R}^2$ such that $\{W + z\}_{z \in Z}$ is a cover of $\mathbb{R}^2$, it holds that $\rho(\{W + z\}_{z \in Z}) = c_Z \cdot A(W)$. Therefore,

$$1 \le \rho(\mathcal{G}) = c_Z \cdot A(Y) = \rho(\mathcal{F}) \cdot \frac{A(Y)}{A(X)} \qquad \Longrightarrow \qquad \rho(\mathcal{F})^{-1} \le \frac{A(Y)}{A(X)}\,.$$

Taking the infimum over all odd-sized K's to minimize $A(Y)/A(X)$, and the infimum over all legal Z's to minimize $\rho(\mathcal{F})$, one concludes that $\theta^\circ(X)^{-1} \le \alpha^\circ(X)$. □

Proof (B) Let $\mathcal{F} = \{X + z\}_{z \in Z}$ be an odd cover of $\mathbb{R}^2$ as before, and let $S \subset \mathbb{R}^2$ be compact. Consider the following mapping ϕ of the compact sets S to the compact subsets of X:

$$\phi(S) = \bigoplus_{z \in Z} (S - z) \cap X = (S \mathring{+} (-Z)) \cap X\,.$$

Claim 3.1

1. $\phi(-X + a) = X$;

2. $\phi\left(\bigoplus_{i=1}^k S_i\right) = \bigoplus_{i=1}^k \phi(S_i)$;

3. $A(\phi(S)) \le A(S) \cdot \tilde{d}_{\mathcal{F}}(S)$, where $\tilde{d}_{\mathcal{F}}(S)$ is the average degree of a cover of S by $\mathcal{F}$, i.e., the average of the cover degrees $d_{\mathcal{F}}(a)$, where a ranges over S.

Proof Indeed, for (1), keeping in mind that $X \mathring{+} Z = \mathbb{R}^2$, and that $a - \mathbb{R}^2 = \mathbb{R}^2$, one gets

$$\begin{aligned}\phi(-X + a) &= \bigoplus_{z \in Z} (-X + a - z) \cap X = X \cap \bigoplus_{z \in Z} a + (-X - z)\\ &= X \cap \ a - (X \mathring{+} Z) = X \cap \mathbb{R}^2 = X.\end{aligned}$$

For (2),

$$\begin{aligned}\phi\left(\bigoplus_{i=1}^k S_i\right) &= \bigoplus_{z \in Z}\left(\left(\bigoplus_{i=1}^k S_i - z\right) \cap X\right) = \bigoplus_{z \in Z}\bigoplus_{i=1}^k (S_i - z) \cap X\\ &= \bigoplus_{i=1}^k \bigoplus_{z \in Z} (S_i - z) \cap X = \bigoplus_{i=1}^k \phi(S_i)\,.\end{aligned}$$

For (3), observing that $(S - z) \cap X = S \cap (X + z) - z$, one concludes that $A(\phi(S)) = A\left(\bigoplus_{z \in Z}[S \cap (X + z) - z]\right) \le \sum_{z \in Z} A(S \cap (X + z)) = A(S) \cdot \tilde{d}_{\mathcal{F}}(S)$. □

Instead of proving a lower bound on $\alpha^{\circ}(X)$, we shall prove one for $\alpha^{\circ}(-X + a)$, with a suitably chosen a. Since $\alpha^{\circ}(X)$ is invariant under affine transformations of $\mathbb{R}^2$, $\alpha^{\circ}(-X + a) = \alpha^{\circ}(X)$. For typographical reasons, set $X_a = -X + a$.

Consider, as before, any finite set $K \subset \mathbb{R}^2$ of odd cardinality, and let $Y_a = X_a \overset{\circ}{+} K$. On the one hand, by Claim 3.1(3), $A(\phi(Y_a)) \le \tilde{d}_{\mathcal{F}}(Y_a) \cdot A(Y_a)$. On the other hand, by Claim 3.1(2)&(1), $\phi(Y_a) = \phi(\bigoplus_{k \in K}(X_a + k)) = \bigoplus_{k \in K} \phi(X_a + k) = \bigoplus_{k \in K} X = X$. Thus,

$$\tilde{d}_{\mathcal{F}}(Y_a) \cdot A(Y_a) \ \ge\ A(\phi(Y_a)) \ =\ A(X) \qquad \Longrightarrow \qquad \frac{A(Y_a)}{A(X_a)} \ \ge\ \tilde{d}_{\mathcal{F}}(Y_a)^{-1} \,.$$

It remains to choose the translation vector a as to minimize $\tilde{d}_{\mathcal{F}}(Y_a)$. Getting back to the discussion of Sect. 2.2, a simple averaging argument shows that for a random uniform $a \in Q_n$, the expected value of $\tilde{d}_{\mathcal{F}}(Y_a)$ gets arbitrarily close to $\tilde{d}_{\mathcal{F}}(Q_n)$ as n tends to infinity. Keeping in mind the definition of $\rho_{\mathcal{F}}$, this implies in turn that there is a sequence of a's such that $\tilde{d}_{\mathcal{F}}(Y_a)$ approaches $\rho_{\mathcal{F}}$. Minimizing over all legal odd covers $\mathcal{F}$, one concludes that the infimum of $\tilde{d}_{\mathcal{F}}(Y_a)$ over $a \in \mathbb{R}^2$ is at most $\theta^{\circ}(X)$. □

To demonstrate the usefulness of Lemma 3.1, assume that there is a tiling of $\mathbb{R}^2$ by translates of X. Then, $\theta^{\circ}(X) = 1$, implying $\alpha^{\circ}(X) = 1$. This yields the aforementioned result about the non-compressibility of the square and the centrally symmetric hexagon.

Further, assume that X is a triangle (a, b, c). Let Λ be the lattice spanned by $\{\frac{1}{2}(b - a),\ \frac{1}{2}(c - a)\}$. Then, $\mathcal{F} = \{X + z\}_{z \in \Lambda}$ is an odd cover of $\mathbb{R}^2$ covering each point in the plane either 1 or 3 times, with $\rho(\mathcal{F}) = 2$. This implies $\alpha^{\circ}(X) \ge \frac{1}{2}$, matching the optimal bound of [1].

4 Odd Covers by Stripe Patterns

A *stripe pattern* is a (non-singular) affine image of the set $\{(x, y) \in \mathbb{R}^2 \mid \lfloor y \rfloor \text{ is even}\}$. I.e., it is an infinite set of parallel stripes of equal width w, such that the distance between any two adjacent stripes is w as well (see Fig. 1). The *direction* of a stripe pattern is, expectedly, the direction of a boundary line of any stripe in it. The *width* of the stripe pattern is the w as above.

We start with the following simple but useful observation about stripe patterns. The easy verification is left to the reader.

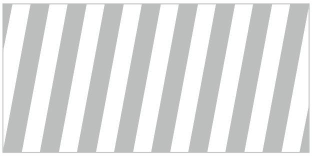


Fig. 1 A stripes pattern

Observation 4.1 *Let S be a stripe pattern, and let ℓ and r be the two lines delimiting one of the stripes in S. Then, for every $a \in \ell$, $b \in r$, and $v = b - a$, it holds that:*

1. $S \overset{\circ}{+} \{0, v\} = S \oplus (S + v) = \mathbb{R}^2$.
2. $S \overset{\circ}{+} \{0, \frac{1}{2}v\} = S \oplus (S + \frac{1}{2}v)$ *is a stripe pattern with the same direction as S, whose width is equal to half of the width of S.*

The main result of this section is:

Lemma 4.1 *Let $S_1, \ldots, S_k$ be stripe patterns with pairwise distinct directions, and let $T = S_1 \oplus \cdots \oplus S_k$. Then, there exists a finite (and efficiently computable) set of vectors $U \subset \mathbb{R}^2$, $|U| \leq 2^{k-1}$, such that $T \overset{\circ}{+} U = \bigoplus_{u_i \in U} (T + u_i) = \mathbb{R}^2$, up to a set of measure* 0.

It will be technically more convenient to prove the following more general statement:

Lemma 4.2 *Let $S_1, \ldots, S_k$ be stripe patterns with pairwise distinct directions, and let $\{Z_i\}_{i=1}^k$ be a family of finite nonempty subsets of $\mathbb{R}^2$, with $Z_1 = \{0\}$. Let $T = \bigoplus_{i=1}^k (S_i \overset{\circ}{+} Z_i)$. Then, as before, there exists a finite (and efficiently computable) set of vectors $U \subset \mathbb{R}^2$, $|U| \leq 2^{k-1}$, such that $T \overset{\circ}{+} U = \bigoplus_{u_i \in U} (T + u_i) = \mathbb{R}^2$, up to a set of measure* 0.

Lemma 4.1 follows from Lemma 4.2 by setting $Z_i = \{0\}$ for all $i \geq 2$.

Notice the special role of S_1 in the statement of Lemma 4.2. In fact, the condition $Z_1 = \{0\}$ is essential even for $k = 1$. It is easy to verify that, using the notation of Observation 4.1, no finite set of translates of the set $T = S_1 \overset{\circ}{+} \{0, \frac{2}{3}v\}$ can oddly cover the plane.[3]

Proof (of Lemma 4.2) For every $i = 1, 2, \ldots, k$, let ℓ_i and r_i denote the two parallel lines delimiting some stripe in S_i. By the assumptions of the Lemma, for different i's these have different directions, and therefore intersect.

The proof proceeds by induction on k.

For $k = 1$, the statement follows from Observation 4.1(1).

For $k = 2$, let a and b be the intersection points of ℓ_1 and r_1 with ℓ_2, respectively. Setting $v_2 = b - a$, we have $S_1 \overset{\circ}{+} \{0, v_2\} = \mathbb{R}^2$, by Observation 4.1(1). Moreover,

[3] Perhaps expectedly, the same T has also the complementary extremal property: $T \overset{\circ}{+} \{0, \frac{1}{3}v, \frac{2}{3}v\} = \emptyset$.

since v_2 has the same direction as of S_2, we have $S_2 \mathring{+} \{0, v_2\} = \emptyset$. Keeping this in mind we have:

$$T \mathring{+} \{0, v_2\} \;=\; \big(S_1 \oplus (S_2 \mathring{+} Z_2)\big) \mathring{+} \{0, v_2\} \;=\; \big(S_1 \mathring{+} \{0, v_2\}\big) \oplus \big(S_2 \mathring{+} Z_2 \mathring{+} \{0, v_2\}\big) \;=$$

$$=\; \mathbb{R}^2 \oplus \big(S_2 \mathring{+} \{0, v_2\} \mathring{+} Z_2\big) \;=\; \mathbb{R}^2 \oplus (\emptyset \mathring{+} Z_2) \;=\; \mathbb{R}^2 \oplus \emptyset \;=\; \mathbb{R}^2 .$$

For $k > 2$, we proceed as follows. Let v_k be the (well-defined) vector such that, on the one hand, $\ell_k + v_k = r_k$, and on the other hand, $\ell_1 + 2v_k = r_1$. Observation 4.1(1) implies that $S_k \mathring{+} \{0, v_k\} = \mathbb{R}^2$, and hence $(S_k \mathring{+} Z_k) \mathring{+} \{0, v_k\}$ equals $\mathbb{R}^2 \mathring{+} Z_k$, which is either $\emptyset$ or $\mathbb{R}^2$, depending on the parity of Z_k. Observation 4.1(2) implies that $S_1 \mathring{+} \{0, v_k\}$ is a stripe pattern with the same direction as S_1, and half its width. Consequently, $\big(S_1 \mathring{+} \{0, v_k\}\big) \oplus \big((S_k \mathring{+} Z_k) \mathring{+} \{0, v_k\}\big)$ is a stripe pattern with the same direction as S_1 and half its width as well.

Consider now the set $T' = T \mathring{+} \{0, v_k\} = T \oplus (T + v_k)$. Using the properties of the operators $\oplus$ and $\mathring{+}$, one gets:

$$T' = T \mathring{+} \{0, v_k\} \;=\; \left(\bigoplus_{i=1}^{k} (S_i \mathring{+} Z_i) \right) \mathring{+} \{0, v_k\} \;=\; \bigoplus_{i=1}^{k} (\, S_i \mathring{+} Z_i \mathring{+} \{0, v_k\} \,) \qquad (2)$$

As we have just seen, the $\oplus$ of the first and the k'th terms of the latter sum is a stripe pattern S_1' with the same direction as S_1. Thus, setting $Z_i' = Z_i \mathring{+} \{0, v_k\}$, one arrives at

$$T' \;=\; S_1' \;\oplus\; \bigoplus_{i=2}^{k-1} (S_i \mathring{+} Z_i') \qquad (3)$$

By the induction hypothesis applied to T', there exists a finite set $U' \subset \mathbb{R}^2$ such that $T' \mathring{+} U' = \mathbb{R}^2$ up to a set of measure 0. However,

$$T' \mathring{+} U' \;=\; T \mathring{+} \{0, v_k\} \mathring{+} U' \;=\; T \mathring{+} (U' \mathring{+} \{0, v_k\}) \qquad (4)$$

Therefore, setting $U = U' \mathring{+} \{0, v_k\}$, one concludes that $T \mathring{+} U = T' \mathring{+} U' = \mathbb{R}^2$. This completes the construction of the desired set U.

It remains to estimate the size of U. The recursive definition $U = U' \mathring{+} \{0, v_k\}$ for $k > 2$, combined with the base cases $|U| = 2^{k-1}$ for $k = 1, 2$, implies the desired bound: $|U| \leq 2^{k-1}$. □

5 Odd Covers by Rational Polygons: A Special Case

In this section we prove our main theorem for the special case of rational polygons with no two parallel edges.

Given a rational polygon P, let P_{INT} be the integer polygon with minimal area affinely equivalent to P, and let $A_{\mathrm{INT}}(P) = A(P_{\mathrm{INT}})$ be its area.

Theorem 5.1 *Let P be a rational polygon with k vertices, and no parallel edges. Then, there exists a bounded degree odd cover $\mathcal{F}$ of $\mathbb{R}^2$ by translates of P with density $\rho(\mathcal{F}) \leq A_{\mathrm{INT}}(P) \cdot 2^{k-1}$. Consequently, $\alpha^\circ(P) \geq A_{\mathrm{INT}}(P)^{-1} \cdot 2^{-(k-1)}$.*

Before starting with the proof, we need one more observation about the structure of $\oplus$-sums of stripe patterns. For $i = 1, \dots r$, let L_i be an affine image of the family of parallel lines $\{(x, y) \in \mathbb{R}^2 \mid y \in \mathbb{Z}\}$. Respectively, let S_i be a stripe pattern whose boundary is L_i. (Notice that there are exactly two such stripe patterns: S_i and its complement $\overline{S}_i = \mathbb{R}^2 \setminus S_i$.) Assume that $S_1, \dots, S_r$ have pairwise distinct directions. The union of all these lines $\bigcup_{i=1}^{r} L_i$ partitions $\mathbb{R}^2$ into pairwise disjoint open cells, each cell being a convex polygon. Call two cells *adjacent* if they share a 1-dimensional edge.

It is a folklore to show that the cells of $\mathbb{R}^2 \setminus \bigcup_{i=1}^{r} L_i$ can be 2-colored in such a way that any two adjacent cells have different colors.

Claim 5.1 *Let T be the union of all cells of $\mathbb{R}^2 \setminus \bigcup_{i=1}^{r} L_i$ in one color class. Then, (up to the 0-measure boundary of T, i.e., $\bigcup_{i=1}^{r} L_i$) either $T = S_1 \oplus \cdots \oplus S_r$, or $T = \mathbb{R}^2 \setminus (S_1 \oplus \cdots \oplus S_r) = \overline{S}_1 \oplus S_2 \oplus \cdots \oplus S_r$.*

The claim is rather obvious, and can be formally verified, e.g., by induction on r. The full details are left to the reader (see Fig. 2 for an illustration).

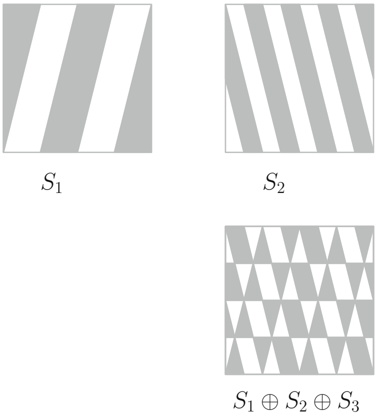


Fig. 2 $\oplus$-sum of three stripes patterns

Proof (of Theorem 5.1) Keeping in mind that both $\theta^\circ(P)$ and $\alpha^\circ(P)$ are affine invariants, one may assume without loss of generality that $P = P_{\mathrm{INT}}$, and that the origin $O = (0,0)$ is a vertex of P. Then, all the vertices of P belong to $\mathbb{Z}^2$. Observe also that some of the edges of P must contain an even number of integer lattice points. Otherwise, the coordinates of the vertices of P would all have the same parity, i.e., they would all be even. Scaling such an all-even P by a factor of $\frac{1}{2}$ would have yielded a smaller integer polygon affinely equivalent to P, contrary to the definition of P_{INT}.

We claim that $P \mathring{+} \mathbb{Z}^2$ is equal to $S_1 \oplus \ldots \oplus S_r$, where $S_1, \ldots, S_1$ are stripe patterns with pairwise distinct directions, and r is at most the number of vertices (=edges) of P. Once this claim is established, the rest easily follows.

Indeed, assuming that the claim holds, by Lemma 4.1 there exists $U \subset \mathbb{R}^2$ with $|U| \leq 2^{r-1}$ such that $(P \mathring{+} \mathbb{Z}^2) \mathring{+} U = \mathbb{R}^2$. Equivalently, the (multi-) family of sets $\mathcal{F} = \{P + z + u\}_{z \in \mathbb{Z}^2,\, u \in U}$ is an odd cover of the plane. To employ the Odd Cover Lemma 3.1, one needs to estimate the density of this cover. Observe that $\{P + z\}_{z \in \mathbb{Z}^2}$ has a bounded maximal degree (being the maximal number of integer lattice points in any translate of P), while its average density is $A(P)$, as mentioned in Claim 2.1. Therefore, the maximal degree of $\mathcal{F}$ is at most $|U|$ times the maximal degree of the cover $\{P + z\}_{z \in \mathbb{Z}^2}$, while $\rho(\mathcal{F})$, the average degree of $\mathcal{F}$, is precisely $A(P) \cdot |U| \leq A(P) \cdot 2^{k-1}$. Hence, $\theta^\circ(P) \leq \rho(\mathcal{F}) \leq A(P) \cdot 2^{k-1}$. Applying the Odd Cover Lemma 3.1 one gets $\alpha^\circ(P) \geq \theta^\circ(P)^{-1} \geq A(P)^{-1} \cdot 2^{-(k-1)}$, as needed.

Thus, it is sufficient to show that $P \mathring{+} \mathbb{Z}^2$ is equal to $S_1 \oplus \ldots \oplus S_r$ as above. In the remainder of this section, we shall focus on proving this claim. The **argument** goes as follows.

Let $E(P)$ denote the set of all edges of P. For $e \in E(P)$, let L_e be the set of all lines parallel to e that contain points of $\mathbb{Z}^2$. Clearly, L_e is a discrete set of lines as in Claim 5.1. Consider a point $x \in \mathbb{R}^2$. It belongs to $P \mathring{+} \mathbb{Z}^2$ exactly when $|P \cap (x - \mathbb{Z}^2)|$ is odd. Unless $x \in \bigcup_{e \in E(P)} L_e$, every point x' in a sufficiently small neighborhood of x will satisfy $|P \cap (x - \mathbb{Z}^2)| = |P \cap (x' - \mathbb{Z}^2)|$. Therefore, $P \mathring{+} \mathbb{Z}^2$ is a union of cells of $\mathbb{R}^2 \setminus \bigcup_{e \in E(P)} L_e$.

Call an edge e of P *active* if it contains an even number of integer lattice points, and *passive* otherwise. Respectively, if e is active, all the lines in L_e are called active, and if it is passive, the lines in L_e are called passive.

Let C_1 and C_2 be two adjacent cells in $\mathbb{R}^2 \setminus \bigcup_{e \in E(P)} L_e$ separated by a line $\ell \in L_e$ for some edge e of P. We claim that if e is active, then exactly one of C_1 and C_2 is contained in $P \mathring{+} \mathbb{Z}^2$, and if e is passive, then either both are contained in $P \mathring{+} \mathbb{Z}^2$, or none of them is.

Indeed, let $I \subset \ell$ denote the common 1-dimensional edge of C_1 and C_2. Observe that the only members in the family $\mathcal{F} = \{P + z\}_{z \in \mathbb{Z}}$ that distinguish between C_1 and C_2, that is, contain exactly one of the two, are those that contain I in their boundary. To get a clearer picture of this subfamily, let $J = [p, q] \subset \ell$ be the smallest interval with integer endpoints containing I. Notice that I has no integer points in its interior. Let us view e as a 1-dimensional interval $[s_e, t_e) \subset \mathbb{R}^2$, parallel to, and having the same orientation as, J. Then, $P + z$ contains I if and only if $p \in e + z$. Or, equivalently, $p - z \in e$.

This means that when e is active (i.e., it contains an even number of points in $\mathbb{Z}^2$), I is covered by an odd number of $(P + z)$'s, and when e is passive, it is covered by an even number of $(P + z)$'s. Consequently, in the former case the degrees of cover of the cells C_1 and C_2 by $\mathcal{F}$ have a different parity, whereas in the latter case the parities are equal. Thus, when e is active, $P \mathring{+} \mathbb{Z}^2$ distinguishes between C_1 and C_2, and when it is passive, it does not. As claimed.

Let $AE(P)$ be the (nonempty!) set of active edges of P. The conclusion is that $P \mathring{+} \mathbb{Z}^2$ is a union of cells of $\mathbb{R}^2 \setminus \bigcup_{e \in AE(P)} L_e$, satisfying the assumptions of Claim 5.1. Hence, $P \mathring{+} \mathbb{Z}^2$ is a $\oplus$-sum of stripe patterns, as desired. This completes the proof of Theorem 5.1. □

The assumption that P has no parallel edges was needed to justify the (tacit) assumption that for every line $\ell \in \bigcup_{e \in E(P)} L_e$, there is a *unique* edge e such that any translate of P may have contained in ℓ. When there are parallel edges, most of the argument still applies, however, it may fail at one fine point. The contributions of parallel edges may cancel out, leaving no active lines, and resulting in $P \mathring{+} \mathbb{Z}^2 = \emptyset$. Unfortunately, this situation indeed does occur for some rational polygons P, for example, for the centrally symmetric ones. To overcome this problem, a more refined family of translates will be constructed.

6 A Theorem About $\mathbb{Z}_2$-Valued Functions on Integer Lattices

We shall need the following result of an independent interest. It will be proven here for any dimension d, but used in Sect. 7 only with $d = 2$.

Let $\mathcal{A}$ be a family of finite subsets of $\mathbb{Z}^d$. A function, or, rather, a weighting, $\mathfrak{F} : \mathbb{Z}^d \to \mathbb{Z}_2$, will be called *stable* with respect to $\mathcal{A}$, if for any $A \in \mathcal{A}$, all integer translates of A have the same $\mathfrak{F}$-weight. That is, the value of $\mathfrak{F}(A + p) = \bigoplus_{x \in A+p} \mathfrak{F}(x)$, does not depend on the choice of $p \in \mathbb{Z}^d$, but solely on A.[4] Further, call $\mathfrak{F}$ 0-*stable* with respect to $\mathcal{A}$, if it is stable, and moreover, for every $A \in \mathcal{A}$, $\mathfrak{F}(A) = 0$. For example, if the function $\mathfrak{F}$ is everywhere 0, then it is 0-stable with respect to any family $\mathcal{A}$. If it is everywhere 1, it is stable with respect to any family $\mathcal{A}$, and 0-stable if $\mathcal{A}$ consists only of sets of even cardinality.

Theorem 6.1 *Let $\mathcal{A}$ be a (possibly infinite) family of non-empty finite subsets of $\mathbb{Z}^d$, and $\mathcal{A} \neq \emptyset$. There exists a function $\mathfrak{F} : \mathbb{Z}^d \to \mathbb{Z}_2$ that is stable, but not* 0*-stable, with respect to $\mathcal{A}$.*

[4] In this section, the operator $\oplus$ that was originally defined on sets, will be sometimes applied to points. For consistency, regard points as single-element sets.

Proof We start with the 1-dimensional case, introducing the key construction to be used in all dimensions.

Case d = 1

We define a family $\{\mathfrak{f}_k\}_{k=0}^{\infty}$ of functions from $\mathbb{Z}$ to $\mathbb{Z}_2$ in the following recursive manner. We will show that one of this functions is the desired function $\mathfrak{F}$:

$\mathfrak{f}_0$ **is identically** 1;
For $k > 0$, $\mathfrak{f}_k(0) = 1$, **and** $\mathfrak{f}_k(t) = \mathfrak{f}_k(t-1) \oplus \mathfrak{f}_{k-1}(t-1)$. [5]

For example, $\mathfrak{f}_1(t)$ is 1 if t is even, and 0 otherwise. The next one, $\mathfrak{f}_2(t)$, is 1 if $t \equiv 0, 3 \pmod 4$, and 0 otherwise. Observe that the repeated application of the recursive formula yields for any $c \in \mathbb{N}$,

$$\mathfrak{f}_k(t+c) = \mathfrak{f}_k(t) \oplus \bigoplus_{i=0}^{c-1} \mathfrak{f}_{k-1}(t+i) . \tag{5}$$

Claim 6.1 *If $\mathfrak{f}_{k-1}$ is 0-stable with respect to a finite $A \subseteq \mathbb{Z}$, then $\mathfrak{f}_k$ is stable with respect to A.*
Indeed, it suffices to show that for any $p \in \mathbb{Z}$, $\mathfrak{f}_k(A+p+1) = \mathfrak{f}_k(A+p)$. By definition of $\mathfrak{f}_k$,

$$\begin{aligned}\mathfrak{f}_k(A+p+1) &= \bigoplus_{t \in A+p+1} \mathfrak{f}_k(t) = \bigoplus_{t \in A+p+1} \mathfrak{f}_k(t-1) \oplus \bigoplus_{t \in A+p+1} \mathfrak{f}_{k-1}(t-1) \\ &= \mathfrak{f}_k(A+p) \oplus \mathfrak{f}_{k-1}(A+p) .\end{aligned}$$

Since $\mathfrak{f}_{k-1}(A+p) = 0$ by assumptions of the claim, one concludes that $\mathfrak{f}_k(A+p+1) = \mathfrak{f}_k(A+p)$.

Claim 6.2 *For any $k \geq 1$, $\mathfrak{f}_k(t) = 0$ for $1 \leq t \leq k$.*
Indeed, apply induction on k. For $k = 1$, $\mathfrak{f}_1(1) = \mathfrak{f}_1(0) \oplus \mathfrak{f}_0(0) = 1 \oplus 1 = 0$. For $k > 1$, using (5), one concludes that for any t in the range,

$$\mathfrak{f}_k(t) = \mathfrak{f}_k(0) \oplus \mathfrak{f}_{k-1}(0) \oplus \mathfrak{f}_{k-1}(1) \oplus \ldots \oplus \mathfrak{f}_{k-1}(t-1) = 1 \oplus 1 \oplus 0 \oplus \ldots \oplus 0 = 0.$$

We proceed to show that one of $\mathfrak{f}_k$'s satisfies the requirements of the theorem. Observe that $\mathfrak{f}_0$ is stable with respect to $\mathcal{A}$. By Claim 6.1, either there exists $k \geq 0$ such that $\mathfrak{f}_k$ is stable, but not 0-stable (precisely as desired), or all $\mathfrak{f}_k$'s are 0-stable. However, the latter situation does not occur. Consider any nonempty $A \in \mathcal{A}$, and let $a = \min(A)$, $b = \max(A)$, $r = b - a$. Then, by Claim 6.2, $\mathfrak{f}_r(A - a) = 1$, and thus $\mathfrak{f}_r$ is not 0-stable. This completes the case $d = 1$.

General Case

Let L be a linear function (without a constant term) from $\mathbb{Z}^d$ to $\mathbb{Z}$ that satisfies two requirements. The first requirement is that the coefficient of x_1 in L is 1. The

[5] Observe that this recursive formula defines $\mathfrak{f}_k(t)$ for both positive and negative values of t. More explicitly, for $t < 0$ it becomes $\mathfrak{f}_k(t) = \mathfrak{f}_k(t+1) \oplus \mathfrak{f}_{k-1}(t)$, reducing either k or $|t|$ just as for $t > 0$.

second requirement is that for some $A \in \mathcal{A}$, L attains a minimum on A at a unique point. Such L's exist. E.g., assuming that A can be translated to a subset of a of the cube $[0, r-1]^d$, the function $\sum_{i=1}^{d} r^{i-1} x_i$ is one-to-one on A by the uniqueness of the base-r representation, and so its minimum on A is attained exactly once.

For $k \geq 0$ and $a \in \mathbb{Z}^d$, define $\mathfrak{F}_k(a) = \mathfrak{f}_k(L(a))$. Respectively, for a finite subset $A \subset \mathbb{Z}^d$, define $\mathfrak{F}_k(A) = \bigoplus_{a \in A} \mathfrak{F}_k(a)$.

The proof proceeds along the same lines as in the 1-dimensional case.

Claim 6.3 *If $\mathfrak{F}_{k-1}$ is 0-stable with respect to a finite $A \subseteq \mathbb{Z}^d$, then $\mathfrak{F}_k$ is stable with respect to it.*

It suffices to show that for any $p \in \mathbb{Z}^d$, and any unit vector $e \in \mathbb{Z}^d$, $\mathfrak{F}_k(A+p+e) = \mathfrak{F}_k(A+p)$. Let $L(e) = c$. If $c = 0$, the statement is trivial. If $c < 0$ the statement reduces to the case $c > 0$ by considering $-e$ instead of e. Thus, without loss of generality, $c > 0$. By (5), the linearity of L, and the first requirement on it,

$$\begin{aligned}\mathfrak{F}_k(A+p+e) &= \bigoplus_{t \in A+p+e} \mathfrak{f}_k(L(t)) = \bigoplus_{t \in A+p} \mathfrak{f}_k(L(t)+c) \\ &= \bigoplus_{t \in A+p} \mathfrak{f}_k(L(t)) \oplus \bigoplus_{i=0}^{c-1} \bigoplus_{t \in A+p} \mathfrak{f}_{k-1}(L(t)+i) = \\ &= \mathfrak{F}_k(A+p) \oplus \bigoplus_{i=0}^{c-1} \bigoplus_{t \in A+p+i \cdot e_1} \mathfrak{f}_{k-1}(L(t)) \\ &= \mathfrak{F}_k(A+p) \oplus \bigoplus_{i=0}^{c-1} \mathfrak{F}_{k-1}(A+p+i \cdot e_1).\end{aligned}$$

Since $\mathfrak{F}_{k-1}$ is 0-stable with respect to A, the second summand is 0, and thus $\mathfrak{F}_k(A+p+e) = \mathfrak{F}_k(A+p)$. This concludes the proof of Claim 6.3.

To conclude the proof of the theorem, observe that $\mathfrak{F}_0$ is stable with respect to $\mathcal{A}$, and thus, by Claim 6.3, either there exists $k \geq 0$ such that $\mathfrak{F}_k$ is stable, but not 0-stable, precisely as desired, or all $\mathfrak{F}_k$'s are 0-stable.

As before, the second possibility does not occur. Indeed, by the second requirement on L, there exists $A \in \mathcal{A}$ on which L attains a unique minimum. Let $p \in A$ be the point on which the minimum is attained, and let $a \in \mathbb{Z}$ denote its value. Then, $L(A - a \cdot e_1) = L(A) - a$ is a subset of $[0, k]$ for some $k \in \mathbb{N}$, and 0 has a unique pre-image $p' = p - a \cdot e_1$. By Claim 6.2, $\mathfrak{F}_k(A - a \cdot e_1) = \mathfrak{f}_k(0) \oplus \bigoplus_{t \in A \setminus p'} \mathfrak{f}_k(L(t)) = 1 \oplus 0 = 1$. □

7 The Main Theorem

We can now prove the main theorem in full generality, making no assumption about parallel edges.

Theorem 7.1 *Let P be a rational polygon with k distinct classes of parallel edges. Then, there exists a bounded degree odd cover $\mathcal{F}$ of $\mathbb{R}^2$ by translates of P with density $\rho(\mathcal{F}) \leq A_{\mathrm{INT}}(P) \cdot 2^{k-1}$. Consequently, $\alpha^{\circ}(P) \geq A_{\mathrm{INT}}(P)^{-1} \cdot 2^{-(k-1)}$.*

Proof While the family of translates will in general be different from the one used in the proof of Theorem 5.1, the logical structure of the proof will be essentially identical. Let us re-examine this structure.

Assuming that $P = P_{INT}$, the first and main goal is to construct a family $\mathcal{F} = \{P + z\}_{z \in Z}$, $Z \subseteq \mathbb{Z}^2$, such that $P \mathring{+} Z$ is a $\oplus$-sum of at most k stripe patterns. (In the former proof, Z was just $\mathbb{Z}^2$.) A close reading of the proof of Theorem 5.1 reveals that in order to prove this fact about $\mathcal{F}$, it is sufficient to show that $\mathcal{F}$ has a certain property. To formulate it we need some definitions.

Let $D(P)$ be the set of all the classes of parallel edges of P, or simply the *directions* of P. For every $d \in D(P)$, let L_d be the set of all lines in $\mathbb{R}^2$ in the direction of d that contain integer lattice points. As before, each L_d is a discrete set of parallel lines with equal distances between any two consecutive ones.

Consider the arrangement of lines $\bigcup_d L_d$. Observe that since $Z \subseteq \mathbb{Z}^2$, for any $(P + z) \in \mathcal{F}$, and any open cell C of $\mathbb{R}^2 \setminus \bigcup_d L_d$, either $C \subseteq (P + z)$, or $C \cap (P + z) = \emptyset$.

Property 7.1 *Let $P, D(P), Z, \mathcal{F}, \bigcup_d L_d$ be as above. Further, let I denote an edge of the arrangement $\bigcup_d L_d$ (i.e., a common 1-dimensional boundary of two adjacent cells) that lies on a line $\ell \in L_d$, $d \in D(P)$. The family $\mathcal{F}$ has the desired property if:*

1. *For* any *edge I of the arrangement $\bigcup_d L_d$ as above, the parity of the number of sets in $\mathcal{F}$ whose boundary contains I depends solely on the corresponding direction d.*
2. *Moreover, there exists $d \in D(P)$ such that this parity is odd. We call such a direction* active, *and denote the set of all active directions by $AD(P)$.*

Once Property 7.1 is established for $\mathcal{F} = \{P + z\}_{z \in Z}$, the **argument** from the proof of Theorem 5.1 implies that $P \mathring{+} Z$ is a (nonempty) union of cells of $\mathbb{R}^2 \setminus \bigcup_{d \in AD(P)} L_d$ that satisfy the assumptions of Claim 5.1. Applying Claim 5.1, one concludes that $P \mathring{+} Z$ is equal to $S_1 \oplus \ldots \oplus S_r$ for some stripe patterns $S_1, \ldots, S_r$, and $r = |AD(P)|$, the number of active directions, is at most $|D(P)| = k$.

Once the main goal is achieved, the rest is easy. Lemma 4.1 is used to conclude that there exists a finite set $U \subset \mathbb{R}^2$ with $|U| \leq 2^{r-1}$, such that $(P \mathring{+} Z) \mathring{+} U = \mathbb{R}^2$. Equivalently, the (multi-) family of sets $\mathcal{F} = \{P + z + u\}_{z \in Z, u \in U}$ is an odd cover of the plane. Since Z is a subset of $\mathbb{Z}^2$, the density of this odd cover is at most $A_{\mathrm{INT}}(P) \cdot |U| \leq A_{\mathrm{INT}}(P) \cdot 2^{k-1}$. (For the appearance of A_{INT}, consult Claim 2.1.) Finally, by the Odd Cover Lemma 3.1, one concludes that $\alpha^{\circ}(P) \geq A_{\mathrm{INT}}(P)^{-1} \cdot 2^{-(k-1)}$, establishing the theorem.

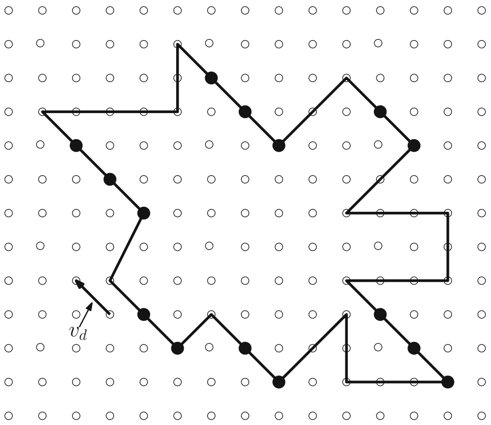


Fig. 3 The points of A_d are the filled discs in the picture

In view of the above, in order to prove Theorem 7.1, it suffices to construct $Z \subseteq \mathbb{Z}^2$ such that $\mathcal{F} = \{P + z\}_{z\in Z}$ has Property 7.1. The remaining part of this section is dedicated to constructing such Z, and proving that $\mathcal{F}$ has the required property.

The set of translates Z is constructed as follows. Assume that $P = P_{INT}$. In particular, the vertices of P are in $\mathbb{Z}^2$. For every direction $d \in D(P)$, define the vector $v_d \in \mathbb{Z}^2$ as the difference between (*any*) pair of two *consecutive* integer lattice points on (*any*) line in L_d, the set of all lines in direction d through an integer lattice point.

Let A_d be the set of all integer lattice points z on the boundary of P such that both z and $z + v_d$ lie on an edge of P in the direction d (see Fig. 3). Let $\mathcal{A} = \{-A_d\}_{d\in D(P)}$. By Theorem 6.1, there exists a $\mathbb{Z}_2$-weighting $\mathfrak{F}$ of $\mathbb{Z}^2$ that is stable, but not 0-stable, with respect to $\mathcal{A}$. Define Z as the support of $\mathfrak{F}$, i.e., $Z = \{z \mid \mathfrak{F}(z) = 1\}$. Finally, define $\mathcal{F} = \{P + z\}_{z\in Z}$. Our goal is to show that the family $\mathcal{F} = \{P + z\}_{z\in Z}$ has Property 7.1.

Call a direction d of an edge of P *active* if $\mathfrak{F}(-A_d) = 1$, and *passive* if $\mathfrak{F}(-A_i) = 0$.

We claim that a point $p \in \mathbb{Z}^2$ belongs to an odd number of sets in $\{A_d + z\}_{z\in Z}$ if d is active, and to an even number of those sets if d is passive. Indeed, the number of solutions of the equation $a + z = p$, where $a \in A_d$, $z \in Z$, is precisely the size of $(-A_d + p) \cap Z$, and hence its parity is $\mathfrak{F}(-A_d + p) = \mathfrak{F}(-A_d)$, as desired.

Let $I \subset \ell \in L_d$, for some d, be an edge in the arrangement of lines $\bigcup_{d\in D(P)}$. Notice that I cannot contain integer lattice points in its (relative) interior. There exists two consecutive integer lattice points p and q on ℓ such that I is contained in the line segment $J = [p, q] \subset \ell$. Observe that $q - p$ is either v_d or $-v_d$; assume w.l.o.g., that $q - p = v_d$.

We claim that the parity of the number of sets from $\mathcal{F} = \{P + z\}_{z\in Z}$ whose boundary contains J is odd if d is active, and even if it is passive. Indeed, J is contained in the boundary of $(P + z)$, $z \in Z$, if and only if $(A_d + z)$ contains p. As we have already seen, the parity of the number of such sets is odd if and only if d is active. In particular, it depends only on d, and not on I, as desired. Moreover, by Theorem 6.1, there exists at least one active direction d.

This concludes the verification of Property 7.1 for the constructed family $\mathcal{F} = \{P + z\}_{z \in Z}$, which in turn concludes the proof of Theorem 7.1. $\square$

Notice that the above proof makes no use of the connectivity of P nor of the connectivity of its boundary. Thus, as it has been already mentioned in the Introduction, Theorem 7.1 applies equally well to any compact figure in $\mathbb{R}^2$ with non-empty interior, piecewise linear boundary, and finite number of vertices, all of which are rational.

References

1. R. Pinchasi, Points covered an odd number of times by translates. Am. Math. Mon. **121**(7), 632–636 (2014)
2. A. Oren, I. Pak, R. Pinchasi, On the odd area of planar sets. Discrete Comput. Geom. **55**(3), 715–724 (2016)
3. The International Mathematics Tournament of the Towns, Fall 2009. Available at http://www.math.toronto.edu/oz/turgor/archives/TT2009F_JAproblems.pdf
4. C.A. Rogers, *Packings and Coverings*. Cambridge Tracts in Mathematics and Mathematical Physics, No. 54 (Cambridge, 1964)

First Order Probabilities for Galton–Watson Trees

Moumanti Podder and Joel Spencer

Abstract In the regime of Galton–Watson trees, first order logic statements are roughly equivalent to examining the presence of specific finite subtrees. We consider the space of all trees with *Poisson* offspring distribution and show that such finite subtrees will be almost surely present when the tree is infinite. Introducing the notion of universal trees, we show that all first order sentences of quantifier depth k depend only on local neighbourhoods of the root of sufficiently large radius depending on k. We compute the probabilities of these neighbourhoods conditioned on the tree being infinite. We give an almost sure theory for infinite trees.

2010 *Mathematics Subject Classification.* Primary05C20; Secondary60F20

1 Introduction and Main Results

For $\lambda > 0$ we let T_λ denote the standard Galton–Watson tree, in which each node independently has *Poisson* offspring with mean λ. We shall set

$$p = p(\lambda) = \Pr[T_\lambda \text{ is infinite}]. \tag{1}$$

As is well known, when $\lambda \leq 1, p(\lambda) = 0$ while when $\lambda > 1, p$ is the unique positive solution to the equation

$$1 - p = e^{-p\lambda}. \tag{2}$$

M. Podder (✉)
Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street,
10012 New York, NY, USA
e-mail: mp3460@nyu.edu

J. Spencer
Computer Science and Mathematics Departments, Courant Institute of Mathematical Sciences,
New York University, 251 Mercer Street, 10012 New York, NY, USA
e-mail: spencer@cims.nyu.edu

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_29

We let T_λ^* denote T_λ conditioned on T_λ being infinite. (When using T_λ^* we tacitly assume $\lambda > 1$.) For any property A of rooted trees we let $\Pr[A], \Pr^*[A]$ denote the probability (as a function of λ) of A in T_λ, T_λ^* respectively.

The *first order logic* for rooted trees consists of equality ($x = y$), parent ($\pi(x, y)$, meaning x is the parent of y), the constant symbol R (the root), the usual Boolean connectives and existential and universal quantification over vertices. A *first order property* is a property that can be written with a sentence A in this language. The quantifier depth of any first order sentence is the number of nested quantifiers involved in expressing the sentence. We illustrate with a few examples what a typical first order sentence looks like.

Example 1.1 Consider the property that there exists a node in the tree that has precisely two children. This can be expressed in first order language as follows:

$$\exists\, u\, [\exists\, v_1\, [\exists\, v_2\, [\pi(u, v_1) \wedge \pi(u, v_2) \wedge [\forall\, v\, \{\pi(u, v) \implies \{(v = v_1) \vee (v = v_2)\}\}]]]].$$

In this particular example, the quantifier depth is 4.

Example 1.2 Consider the property that the root of the tree has precisely one child and precisely one grandchild. Observe that the root of the tree being a designated symbol, this property is written in first order language as follows:

$$\begin{aligned}\exists\, u \big[\exists\, v \big[\pi(R, u) \wedge \pi(u, v) \wedge \big[\forall\, u' \big\{\pi(R, u') \implies \big(u' = u\big)\big\}\big] \\ \wedge \big[\forall\, v' \big\{\pi(u, v') \implies \big(v' = v\big)\big\}\big]\big]\big].\end{aligned}$$

The quantifier depth is 3.

We refer the reader to [2] for further discussion on first order logic. Detailed discussions on random graphs, the general probabilistic methods, and various tools, such as the Azuma's inequalities, can be found in Alon and Spencer [1].

Our main results (Theorem 4.7 and Corollary 4.8) will be a characterization of the possible $\Pr^*[A]$, as functions of λ, where A is a first order property. However, it can be shown that the property of T being infinite is not first order.

Notations 1.3 *Let $v \in T$, T a rooted tree. $T(v)$ denotes the subtree of T that is rooted at v. w is an i-descendant of v if there is a sequence $v = x_0, x_1, \ldots, x_i = w$ so that x_j is the parent of x_{j+1} for $0 \le j < i$. (We say v is a 0-descendant of itself.) (In the Ulam–Harris notation for trees, this can be expressed as $w = (x_0, x_1, \ldots, x_i)$ where $x_0 = v$ and $x_i = w$.) w is a $(\le i)$-descendant of v if it is a j-descendant for some $0 \le j \le i$. (E.g., 3-descendants are great-grandchildren.) We define $d(T)$ to be the depth of the tree, which may be infinite. For $n \ge 1$, $T|_n$ denotes the first n generations of T, along with the root. That is, if $d(T) > n$, then we sever all nodes after the n-th generation (where root is the 0-th generation) and call the truncated tree $T|_n$. If, of course, $d(T) \le n$, then $T|_n = T$. Let T_0 be a finite tree. We say T contains T_0 as a subtree if for some $v \in T$, $T(v) \cong T_0$. We note that this is a first order property. Letting T_0 have s nodes, the first order sentence is that there exist distinct $v_1, \ldots, v_s$ having all the desired parent relations and with $v_1, \ldots, v_s$ having no additional children.*

We use a *fictitious continuation* to analyze T_λ. Let $X_1, X_2, \dots$ be a countable sequence of mutually independent and identically distributed *Poisson*(λ) random variables. Let X_i be the number of children of the i-th node, when the tree is explored using Breadth First Search. (The root is considered the first node so that X_1 is its number of children.) If and when the tree terminates (this occurs when $\sum_{i=1}^{n} X_i = n-1$ for the first time) the remaining (fictitious) X_j are not used. We refer the reader to van der Hofstad [4] for an elaborate discussion on trees and branching processes, especially regarding the survival probability of the process.

Theorem 1.4 *Fix an arbitrary finite tree T_0. Consider the following statement A:*

$$A := \{\textit{either } T \textit{ contains } T_0 \textit{ as a subtree or } T \textit{ is finite}\}. \tag{3}$$

Then $\Pr[A] = 1$.

This is one of the main results of this paper. Note that, in particular, Theorem 1.4 immediately implies that for any arbitrary but fixed finite T_0,

$$\overset{*}{\Pr}[\exists\, v : T(v) \cong T_0] = 1. \tag{4}$$

This gives us a good structural description of the infinite random Galton–Watson tree, in the sense that every local neighbourhood is almost surely present somewhere inside the tree.

1.1 Rapidly Determined Properties

We say (employing a useful notion of Donald Knuth) that an event is *quite surely determined* in a certain parameter s if the probability of the complement of that event is exponentially small in s.

Definition 1.5 Consider the fictitious continuation process T_λ. We say that an event B is *rapidly determined* if *quite surely* B is *tautologically determined* by $X_1, X_2, \dots, X_s$. Here, *tautologically determined* means that for every point ω in the sample space, the realization $(X_1(\omega), X_2(\omega), \dots, X_s(\omega))$ completely determines whether the event B occurs or not. This means that for every sufficiently large $s \in \mathbb{N}$,

$$\Pr[B \text{ is not determined by } X_1, X_2, \dots, X_s] \le e^{-\beta s} \tag{5}$$

where $\beta > 0$ is independent of s.

Theorem 1.6 *The event A described in* (3) *is a* rapidly determined *property.*

We shall now prove Theorem 1.4 subject to Theorem 1.6. Fix an arbitrary finite T_0. Assume Theorem 1.4 is false so that $\Pr[A] < 1$, where A is as in (3). For each $s \in \mathbb{N}$, with probability at least $1 - \Pr[A]$ the values $X_1, \dots, X_s$ do not terminate the tree, nor do they force a copy of T_0. Then A would not be tautologically determined.

So A would not be *rapidly determined* and Theorem 1.6 would be false. Taking the contrapositive, Theorem 1.6 implies Theorem 1.4. We prove Theorem 1.6 in Sect. 2.1.

Remark 1.7 The conclusion of Theorem 1.4 is really that, fixing any finite tree T_0, T_λ^* contains T_0 as a subtree with probability one. We can say a bit more. Let T_0 have root v. For $L \geq 1$ define $T_0[L]$ by adding L new points $v_0, \ldots, v_{L-1}$ and making v_i a child of v_{i-1}, $1 \leq i \leq L-1$ and v a child of v_{L-1}. T_λ^* contains $T_0[L]$ with probability one. But then it contains a T_0 where the root of T_0 is at least distance L from the root of T. We thus deduce that for any finite T_0 and any L there will, with probability one in T_λ^*, be a v at distance at least L from the root such that $T(v) \cong T_0$.

1.2 Ehrenfeucht Games

We use a very standard and well-known tool to analyze first order properties on rooted trees, namely the Ehrenfeucht games. The Ehrenfeucht games are what bridges the gap between mathematical logic and a complete structural description of logical statements on graphs. Fix a positive integer k. The standard k-move Ehrenfeucht game used to analyze first order properties partitions the space of all rooted trees into finitely many equivalence classes. Any two trees belonging to the same equivalence class if and only if they have the same truth value for every first order property of quantifier depth $\leq k$. That is, given a first order sentence A of quantifier depth at most k, if A holds true for one of the trees in an equivalence class, then it holds true for all others in that class as well. This notion is made more precise in the following exposition.

We begin with describing the *standard* game, and later move on to a more specialized variant of the game that is suited to our analysis. Fix $k \geq 1$ and two trees T_1 rooted at R_1 and T_2 rooted at R_2 (these are known to both players). The Ehrenfeucht game $EHR[T_1, T_2; k]$ is a k-round game between two players, the Spoiler and the Duplicator. In each round Spoiler picks a vertex from *either* T_1 or T_2 and then Duplicator picks a vertex from the other tree. Letting $x_1, \ldots, x_k; y_1, \ldots, y_k$ be the vertices selected (in that order) from T_1, T_2 respectively, Duplicator wins if *all* of the following hold:

(i) $x_i = R_1$ iff $y_i = R_2$;
(ii) $\pi(x_i, x_j)$ iff $\pi(y_i, y_j)$, i.e. x_i is the parent of x_j if and only if y_i is the parent of y_j;
(iii) $\pi(R_1, x_i)$ iff $\pi(R_2, y_i)$, i.e., if x_i is a child of the root R_1, then y_i is a child of R_2, and vice versa;
(iv) $x_i = x_j$ iff $y_i = y_j$.

We write $T_1 \equiv_k T_2$ if and only if Duplicator wins $EHR[T_1, T_2; k]$. This equivalence relation partitions all rooted trees into finitely many equivalence classes. It can be shown that two rooted trees T_1, T_2 (with roots R_1, R_2) have the same k-Ehrenfeucht value iff they satisfy precisely the same first order properties of quantifier depth at most k.

We shall now describe the promised modified version of the game. Let T be a rooted tree, $v \in T$, and $r > 0$. Let T^- be the (undirected) tree on the same vertex set with x, y adjacent iff one of them is the parent of the other. Let $B_T(v; r)$ denote the ball of radius r around v. That is,

$$B_T(v; r) = \{u \in T : d(u, v) < r \text{ in } T^-\} \tag{6}$$

Here $d(\cdot, \cdot)$ gives the usual graph distance. (For example, cousins are at distance four.) Let k (the number of rounds) and M (an upper bound on the maximal distance) be fixed. Let T_1, T_2 be trees with *designated nodes* $v_1 \in T_1, v_2 \in T_2$. Set

$$B_i = B_{T_i}(v_i; \lfloor M/2 \rfloor), \quad i = 1, 2.$$

The k-move M-distance preserving Ehrenfeucht game, denoted by $EHR_M[B_1, B_2; k]$, is played on these balls. We add a round zero in which the moves v_1, v_2 must be played. (Essentially these are designated vertices.) As before, each round (1 through k) Spoiler picks a vertex from either T_1 or T_2 and then Duplicator picks a vertex from the other tree. Letting $x_0, \ldots, x_k$; $y_0, \ldots, y_k$ be the vertices selected from T_1, T_2 respectively, Duplicator wins if

- For $0 \leq i, j \leq k$, $d(x_i, x_j) = d(y_i, y_j)$. Equivalently, for all $1 \leq s \leq M$ and all $0 \leq i, j \leq k$, $d(x_i, x_j) = s$ if and only if $d(y_i, y_j) = s$.
- For $0 \leq i, j \leq k$, $\pi(x_i, x_j)$ iff $\pi(y_i, y_j)$.
- For $0 \leq i, j \leq k$, $x_i = x_j$ iff $y_i = y_j$.

Two balls B_1, B_2 (as described above) are said to have *the same* $(M; k)$-*Ehrenfeucht value* if Duplicator wins $EHR_M[B_1, B_2; k]$. We denote this by

$$B_1 \equiv_{M;k} B_2 \tag{7}$$

This being an equivalence relation, the space of all such balls with designated centers, is partitioned into $(M; k)$-*equivalence classes*. We let $\Sigma_{M;k}$ denote the set of all $(M; k)$-equivalence classes.

We create a first order language consisting of $=, \pi(x, y)$ and $d(x, y) = s$ for $1 \leq s \leq M$ (note that s is *not* a variable here). There are only finitely many binary predicates (relations involving two variables). (In general adding the distance function would add an unbounded number of binary predicates. In our case, however, the diameter is bounded by M and so we are only adding the M predicates $d(x, y) = s, 1 \leq s \leq M$.) Hence the number of equivalence classes corresponding to this game will also be finite. That is, $\Sigma_{M;k}$ is a finite set.

1.3 Universal Trees

A *universal tree*, as defined below, shall have the property that once T contains it, all first order statements up to quantifier depth k depend only on the local neighborhood of the root.

Definition 1.8 Fix a positive integer k. Let

$$M_0 = 2 \cdot 3^{k+1}. \tag{8}$$

A finite tree T_0 will be called *universal* if the following phenomenon happens: Take any two trees T_1, T_2 with roots R_1, R_2 such that:

(i) the 3^{k+1} neighbourhoods around the root have the same $(M_0; k)$ value, i.e.

$$B_{T_1}(R_1; 3^{k+1}) \equiv_{M_0;k} B_{T_2}(R_2; 3^{k+1}). \tag{9}$$

(ii) For some $u_1 \in T_1, u_2 \in T_2$ such that

$$d(R_1, u_1) > 3^{k+2}, \quad d(R_2, u_2) > 3^{k+2}, \tag{10}$$

we have

$$T_1(u_1) \cong T_2(u_2) \cong T_0. \tag{11}$$

Then $T_1 \equiv_k T_2$. Equivalently, Duplicator wins the *standard* k-move Ehrenfeucht game played on T_1, T_2.

Remark 1.9 Technically, we should call such a T_0 as described in Definition 1.8 *k-universal*. However, in the sequel, we simply refer to this as *universal* for the convenience of notation, and since the dependence on k will be clear in each context.

We prove in Theorem 3.3 that such a *universal tree* indeed exists, by imposing sufficient structural conditions on it.

Remark 1.10 Fix a certain *universal tree* $UNIV_k$, given $k \in \mathbb{N}$. Using Theorem 1.4, we conclude that T_λ^* will almost surely contain $UNIV_k$. From Remark 1.7, we say further that there will almost surely exist a node v at distance $> 3^{k+2}$ from the root such that

$$T(v) \cong UNIV_k.$$

From the definition of *universal trees*, then the *standard* Ehrenfeucht value of T_λ^* will be determined by the $(M_0; k)$-Ehrenfeucht value of $B_{T_\lambda^*}(R; 3^{k+1})$, the 3^{k+1}-neighbourhood of the root R, where M_0 is as in (8).

1.4 An Almost Sure Theory

Let $\mathcal{B}_i, 1 \leq i \leq N$ for some positive integer N, denote the finitely many $(M_0;k)$-equivalence classes. Note that these are defined on balls of radius 3^{k+1} centered at a designated vertex which is a node in some tree. Then for every realization T of T^*_λ,

$$B_T(R;3^{k+1}) \in \mathcal{B}_i \quad \text{for precisely one } i, 1 \leq i \leq N. \tag{12}$$

Almost surely for two realizations T_1, T_2 of T^*_λ which have the same local neighbourhoods of the roots, i.e.

$$B_{T_1}(R_1;3^{k+1}) \in \mathcal{B}_i, \quad B_{T_2}(R_2;3^{k+1}) \in \mathcal{B}_i \quad \text{for the same } i,$$

we have $T_1 \equiv_k T_2$. As the $\mathcal{B}_i$ are equivalence classes over the space of rooted trees they may be considered properties of rooted trees and so have probabilities $\Pr^*[\mathcal{B}_i]$ in T^*_λ. As they finitely partition the space of all rooted trees

$$\sum_{i=1}^{N} \overset{*}{\Pr}[\mathcal{B}_i] = 1. \tag{13}$$

Let $\mathcal{AS}$ denote the almost sure theory for T^*_λ. That is, $\mathcal{AS}$ consists of all first order sentences B such that $\Pr^*[B] = 1$. We now give an axiomatization of $\mathcal{AS}$. Let $\mathcal{T}$ be defined by the following schema:

$$A\,[T_0] := \{\exists\ v : T(v) \cong T_0\}\,, \text{ for all } T_0 \text{ finite trees.} \tag{14}$$

Theorem 1.11 *Under the probability* $\Pr^*$*,*

$$\mathcal{T} = \mathcal{AS} \tag{15}$$

That is, the first order statements B with $\Pr^*[B] = 1$ *are precisely those statements derivable from the axiom schema $\mathcal{T}$.*

As $\mathcal{T}$ does not depend on λ we also have:

Corollary 1.12 *The almost sure theory $\mathcal{AS}$ is the same for all* $\lambda > 1$.

That $\mathcal{T} \subseteq \mathcal{AS}$ is already clear from Theorem 1.4. To show the reverse inclusion, consider for every $1 \leq i \leq N$, $\mathcal{T} + \mathcal{B}_i$. In this theory every finite T_0 is contained as a subtree and the 3^{k+1}-neighbourhood of the root belongs to the equivalence class $\mathcal{B}_i$. As discussed above in Remark 1.10, this set of information completely determines the *standard* Ehrenfeucht value of the infinite tree. That is, for any first order sentence A of quantifier depth k

$$\text{either } \mathcal{T} + \mathcal{B}_i \models A \quad \text{or } \mathcal{T} + \mathcal{B}_i \models \neg A. \tag{16}$$

The standard notation $T \models A$ for a tree T and a property A means that the property A holds true for tree T. Set

$$J_A = \{1 \leq i \leq N : \mathcal{T} + \mathcal{B}_i \models A\}. \tag{17}$$

Under T_λ^*, A holds *if and only if* $\mathcal{B}_i$ holds for some $i \in J_A$. Thus we can express

$$\overset{*}{\Pr}[A] = \sum_{i \in J_A} \overset{*}{\Pr}[\mathcal{B}_i]. \tag{18}$$

In Sect. 4 we shall use this to express all $\Pr^*[A]$ in reasonably succinct form.

Now suppose, under T_λ^*, that $\Pr^*[A] = 1$. As the $\mathcal{B}_i$ partition the neighbourhoods around the roots of trees, this implies that $J_A = \{1, 2, \ldots N\}$. That is, $\mathcal{T} + \mathcal{B}_i \models A$ for all $1 \leq i \leq N$ and $\bigvee_{i=1}^N \mathcal{B}_i$ is a tautology. Hence A is derivable from $\mathcal{T}$. Thus $\mathcal{AS} \subseteq \mathcal{T}$.

In Sect. 4 below, we turn to the computation of the possible $\Pr^*[A]$. As seen above, in the space of T_λ^*, the neighbourhoods around the root of sufficiently large radius are instrumental in determining the *standard* Ehrenfeucht value of the tree. It only makes sense, therefore, to compute the probabilities of having specific neighbourhoods around the root conditioned on the tree being infinite. We shall do this in a recursive fashion, using induction on the number of generations below the root that we are considering.

2 Containing All Finite Trees

2.1 A Rapidly Determined Property

We prove here Theorem 1.6. We fix an arbitrary finite tree T_0 with depth $d(T_0) = d_0$, following the notation given in Notations 1.3. We alter the *fictitious continuation* process T_λ described previously. If for some finite, first $n \in \mathbb{N}$, we have $\sum_{i=1}^n X_i = n - 1$, then the actual tree has vertices $1, \ldots, n$. If this phenomenon does not happen for any finite n, then we have one infinite tree described by our fictitious continuation. If the tree does abort after n vertices, we begin a new tree with vertex $n + 1$ as the root, and generate it from $X_{n+1}, X_{n+2}, \ldots$. Again, if this tree terminates at some n_1 we begin a new tree with vertex $n_1 + 1$. Continuing, we generate an infinite forest, with vertices the positive integers. We call this the *forest* process and label it T_λ^{for}.

Then we define, for every $s \in \mathbb{N}$, the event (in T_λ)

$$good(s) = \{A \text{ is completely determined by } X_1, \ldots X_s\}, \tag{19}$$

where A is as in (3). Set $bad(s) = good(s)^c$. For every node $i \in \mathbb{N}$, define in T_λ^{for}

$$I_i = \mathbf{1}_{T(i)\cong T_0}. \tag{20}$$

That is, I_i is the indicator function of the event that *in the random forest* $T(i) \cong T_0$. Set

$$Y = \sum_{i=1}^{\lfloor \epsilon^{d_0} s \rfloor} I_i, \tag{21}$$

where, with foresight, we require

$$0 < \epsilon < \frac{1}{\lambda + 1}. \tag{22}$$

(Our ϵ is chosen sufficiently small so that quite surely in s, in T_λ^{for}, all of the ($\leq d_0$)-descendants j of all $i \leq s\epsilon$ have $j \leq s$.) We create a martingale, setting, for $1 \leq i \leq s$,

$$Y_i = E[Y|X_1, X_2, \dots X_i], \quad Y_0 = E[Y]. \tag{23}$$

In T_λ^{for}, for $x \in \mathbb{R}^+, i \in \mathbb{N}$, set

$$\mathcal{S}_i(x) = \{\text{indices of all } i\text{-descendants of nodes } 1, 2, \dots \lfloor x \rfloor\} \tag{24}$$

with $\mathcal{S}_0(x) = \{1, 2, \dots \lfloor x \rfloor\}$, where an i-descendant is as described in Notations 1.3. Define, for $i \in \mathbb{N}$,

$$g_i(x) = \text{highest index recorded in } \bigcup_{j=0}^{i} \mathcal{S}_j(x). \tag{25}$$

Lemma 2.1 *For any* $x \in \mathbb{R}^+, d \in \mathbb{N}$,

$$g_d(x) = g_1^d(x). \tag{26}$$

Here g_1^d *denotes the d-times composition of* g_1.

Proof We prove this using induction on d. For $d = 1$ this is true by definition of g_1. For $d = 2$, the highest possible index of all the children and grandchildren of $1, 2, \dots \lfloor x \rfloor$ is equal to the highest index of the children of the nodes $1, 2, \dots g_1(\lfloor x \rfloor) = g_1(x)$, which is $g_1(g_1(x))$. Now suppose we have proved the claim for some $d \in \mathbb{N}, d \geq 2$. Once again, a similar argument comes into play. The highest index among all the $(d + 1)$-descendants of nodes $1, 2, \dots \lfloor x \rfloor$, is also equal to the highest index among all the d-descendants of the nodes $1, 2, \dots g_1(x)$, which by induction hypothesis will be $g_1^d(g_1(x)) = g_1^{d+1}(x)$. □

When $g_{d_0}(\lfloor \epsilon^{d_0} s \rfloor) \leq s$, the descendents j of $1, \ldots, \lfloor \epsilon^{d_0} s \rfloor$ down to generation d_0 will all have $j \leq s$. Thus Y will be completely determined by $X_1, \ldots, X_s$. That is,

$$g_1^{d_0}(\lfloor \epsilon^{d_0} s \rfloor) \leq s \quad \Rightarrow \quad Y_s = Y. \tag{27}$$

A few observations about the function $g_1(\cdot)$ are important. First,

$$g_1(x) \geq \lfloor x \rfloor \quad \text{for all } x \in \mathbb{R}^+. \tag{28}$$

In T_λ^{for} every time the tree terminates, we start a new tree, and that uses up one extra index for the root of the new tree. But while counting the nodes $1, 2, \ldots, \lfloor x \rfloor$, for any $x \in \mathbb{R}^+$, at most $\lfloor x \rfloor$ many new trees need be started. Therefore

$$g_1(x) \leq \lfloor x \rfloor + \sum_{i=1}^{\lfloor x \rfloor} X_i. \tag{29}$$

Further, by the definition of $g_1(\cdot)$, it is clear that it is monotonically increasing.

We shall use the inequality in (29) to show that, for ϵ as chosen in (22), quite surely in s, we have $Y_s = Y$, i.e. Y is tautologically determined by $X_1, \ldots, X_s$ with exponentially small failure probability in s. This involves showing that for i this small, i.e. $1 \leq i \leq \lfloor \epsilon^{d_0} s \rfloor$, $T(i)$ is quite surely determined by $X_1, \ldots, X_s$.

We employ Chernoff bounds. For $x \in \mathbb{R}^+$ and any $\alpha > 0$,

$$\begin{aligned}
\Pr[g_1(\epsilon x) > x] &= \Pr[e^{\alpha g_1(\epsilon x)} > e^{\alpha x}] \\
&\leq E[e^{\alpha g_1(\epsilon x)}]e^{-\alpha x} \\
&\leq E[e^{\alpha(\epsilon x + \sum_{i=1}^{\lfloor \epsilon x \rfloor} X_i)}]e^{-\alpha x} \\
&= e^{\alpha \epsilon x} \prod_{i=1}^{\lfloor \epsilon x \rfloor} E[e^{\alpha X_i}]e^{-\alpha x} \\
&= e^{-(1-\epsilon)\alpha x} \{\exp[\lambda (e^\alpha - 1)]\}^{\lfloor \epsilon x \rfloor} \\
&\leq e^{-(1-\epsilon)\alpha x} \{\exp[\lambda (e^\alpha - 1)]\}^{\epsilon x} \\
&= \exp\{-[(1-\epsilon)\alpha - \lambda(e^\alpha - 1)\epsilon]x\}.
\end{aligned} \tag{30}$$

It can be checked that for any $\alpha \in \left(0, \log\left(\frac{1-\epsilon}{\lambda\epsilon}\right)\right)$, the exponent in (30) is negative, i.e. $-[(1-\epsilon)\alpha - \lambda(e^\alpha - 1)\epsilon] < 0$. We set

$$\eta = [(1-\epsilon)\alpha - \lambda(e^\alpha - 1)\epsilon] \tag{31}$$

Observe that η is positive. Now we have the upper bound:

$$\Pr[g_1(\epsilon x) > x] \leq e^{-\eta x}. \tag{32}$$

We make the following claim:

Lemma 2.2 *For any $d, s \in \mathbb{N}$,*

$$\Pr[g_1^d(\epsilon^d s) > s] \leq \sum_{i=0}^{d-1} e^{-\epsilon^i \eta s}. \tag{33}$$

Proof We prove this using induction on d. We have already seen that this holds for $d = 1$. This initiates the induction hypothesis. Suppose it holds for some $d \in \mathbb{N}$. Then

$$\begin{aligned}
\Pr[g_1^{d+1}(\epsilon^{d+1} s) > s] &= \Pr[g_1^{d+1}(\epsilon^{d+1} s) > s, g_1(\epsilon^{d+1} s) > \epsilon^d s] \\
&\quad + \Pr[g_1^{d+1}(\epsilon^{d+1} s) > s, g_1(\epsilon^{d+1} s) \leq \epsilon^d s] \\
&\leq \Pr[g_1(\epsilon^{d+1} s) > \epsilon^d s] + \Pr[g_1^d(\epsilon^d s) > s] \\
&\leq e^{-\eta \cdot \epsilon^d s} + \sum_{i=0}^{d-1} e^{-\epsilon^i \eta s}, \quad \text{by induction hypothesis and (32);} \\
&= \sum_{i=0}^{d} e^{-\epsilon^i \eta s}.
\end{aligned}$$

This completes the proof. □

From Lemma 2.2, we conclude that

$$\Pr[g_1^{d_0}(\lfloor \epsilon^{d_0} s \rfloor) > s] \leq \sum_{i=0}^{d_0 - 1} e^{-\epsilon^i \eta s}, \tag{34}$$

From (27), this means

$$\Pr[Y_s = Y] \geq 1 - \sum_{i=0}^{d_0 - 1} e^{-\epsilon^i \eta s}. \tag{35}$$

As promised earlier, we therefore have that, quite surely, $Y_s = Y$. In the following definition, we describe the event $Y_s = Y$ as *globalgood*(s), emphasizing the dependence on the parameter s. What we can conclude from the above computation is that *globalgood*(s) fails to happen with only exponentially small failure probability in s.

Definition 2.3 *globalgood*(s) is the event $Y_s = Y$. *globalbad*(s) is the complement of *globalgood*(s).

We now claim that the martingale $\{Y_i : 0 \leq i \leq s\}$ satisfies a Lipschitz Condition.

Lemma 2.4 *There exists constant* $C > 0$ *such that for* $1 \leq i \leq s$,

$$|Y_i - Y_{i-1}| \leq C. \tag{36}$$

Proof For $1 \leq i \leq \lfloor \epsilon^{d_0} s \rfloor$, fix a sequence $\vec{x} = (x_1, \dots x_{i-1}) \in (\mathbb{N} \cup \{0\})^{i-1}$, and then consider

$$y_i = E[Y | X_1 = x_1, \dots X_{i-1} = x_{i-1}, X_i] = \sum_{1 \leq j \leq \lfloor \epsilon^{d_0} s \rfloor} E[I_j | X_1 = x_1, \dots X_{i-1} = x_{i-1}, X_i]$$

and

$$y_{i-1} = E[Y | X_1 = x_1, \dots X_{i-1} = x_{i-1}] = \sum_{1 \leq j \leq \lfloor \epsilon^{d_0} s \rfloor} E[I_j | X_1 = x_1, \dots X_{i-1} = x_{i-1}].$$

I_j will be affected by the extra information about X_i only if either $j = i$ or node j is an ancestor of node i at distance $\leq d_0$ from i. If $j = i$, then it will of course affect the conditional expectation because X_i gives the number of children of j in that case. When $j > i$, this is immediate, because any subtree rooted at j has no involvement of X_i. When $j < i$, but not an ancestor of i, i is not a part of the subtree $T(j)$ rooted at j. Therefore X_i, the number of children of node i, does not contribute anything to the probability of the presence of T_0 rooted at j. When j is an ancestor of i but at distance $> d_0$ from i, i won't be a part of the subtree $T(j)|_{d_0}$ at all.

When j is an ancestor of i and at distance d_0 from i, then i is a leaf node of $T(j)|_{d_0}$ and therefore X_i, the number of children of i, will actually play a role, because to ensure that $T(j) \cong T_0$, the leaf nodes of $T(j)|_{d_0}$ must have no children of their own in T_λ^{for}.

That is, we need be concerned with the at most d_0 ancestors of node i, plus i itself, and for each of them, the difference in the conditional expectations of I_j can be at most 1. Denoting by $\sum^*$ the sum over $j = i$ and j an ancestor of i at distance $\leq d_0$ from i, this gives us:

$$\begin{aligned} |y_i - y_{i-1}| &= \left| \sum^{*} E[I_j | X_1 = x_1, \dots X_{i-1} = x_{i-1}, X_i] - E[I_j | X_1 = x_1, \dots X_{i-1} = x_{i-1}] \right| \\ &\leq \sum^{*} \left| E[I_j | X_1 = x_1, \dots X_{i-1} = x_{i-1}, X_i] - E[I_j | X_1 = x_1, \dots X_{i-1} = x_{i-1}] \right| \\ &\leq d_0 + 1. \end{aligned}$$

The final inequality follows from the argument above that $\sum^*$ involves summing over at most $d_0 + 1$ many terms, and each summand is at most 1, since each summand is the difference of the expectations of indicator random variables. This proves Lemma 2.4, with $C = d_0 + 1$.

□

Given Lemma 2.4 we apply Azuma's inequality. Consider the martingale

$$Y_i' = \frac{E[Y] - Y_i}{d_0 + 1}, \quad 0 \le i \le s.$$

Set, for a typical node v in a random Galton–Watson tree T with *Poisson*(λ) offspring distribution,

$$\Pr[T(v) \cong T_0] = p_0, \tag{37}$$

so that $E[Y] = \lfloor \epsilon^{d_0} s \rfloor p_0$. Applying Azuma's inequality to $\{Y_i', 0 \le i \le s\}$, for any $\beta > 0$,

$$\Pr[Y_s' > \beta\sqrt{s}] < e^{-\beta^2}.$$

We choose

$$\beta = \frac{\epsilon^{d_0} p_0 \sqrt{s}}{2(d_0 + 1)}.$$

This gives

$$\Pr\left[Y_s < \frac{\epsilon^{d_0} p_0}{2} \cdot s - p_0\right] < \exp\left\{-\frac{\epsilon^{2d_0} p_0^2}{4(d_0+1)^2} \cdot s\right\}. \tag{38}$$

Writing

$$\xi = \frac{\epsilon^{d_0} p_0}{2}, \quad \varphi = \frac{\epsilon^{2(d_0)} p_0^2}{4(d_0+1)^2},$$

we can rewrite the above inequality as

$$\Pr[Y_s < \xi s - p_0] < e^{-\varphi s}. \tag{39}$$

Putting everything together, we get for all s large enough:

$$\begin{aligned}\Pr[Y = 0] &= \Pr[Y = 0, Y_s = Y] + \Pr[Y = 0, Y_s \neq Y] \\ &\le \Pr[Y_s < \xi s - p_0] + \Pr[Y_s \neq Y] \\ &\le e^{-\varphi s} + \sum_{i=0}^{d_0-2} e^{-\epsilon^i \eta s}; \quad \text{from (35) and (39)};\end{aligned}$$

which is an upper bound exponentially small in s. This gives us the proof of Theorem 1.6.

3 Universal Trees Exist!

In this section, we shall establish sufficient conditions that guarantee the existence of *universal trees*. Fixing $k \in \mathbb{N}$, set $M_0 = 2 \cdot 3^{k+1}$ as in (8). *Assume* T_0 is a finite tree with root R_0 with the following properties:

(i) For every $\sigma \in \Sigma_{M_0;k}$, there are distinct nodes $v_{i;\sigma} \in T_0, 1 \leq i \leq k$, with the following conditions satisifed: for every $\sigma \in \Sigma_{M_0;k}$ and every $1 \leq i \leq k$, we have

$$d(R_0, v_{i;\sigma}) > 3^{k+2}; \tag{40}$$

for every $\sigma_1, \sigma_2 \in \Sigma_{M_0;k}$ and $1 \leq i_1, i_2 \leq k$, with $(\sigma_1, i_1) \neq (\sigma_2, i_2)$, we have

$$d(v_{i_1;\sigma_1}, v_{i_2;\sigma_2}) > 3^{k+2}; \tag{41}$$

and for all $1 \leq i \leq k, \sigma \in \Sigma_{M_0;k}$,

$$B(v_{i;\sigma}; 3^{k+1}) \in \sigma. \tag{42}$$

(ii) For every $1 \leq i \leq k$, every choice of $u_1, \ldots u_{i-1} \in T_0$, and every choice of $\sigma \in \Sigma_{M_0;k}$, there exists a vertex $u_i \in T_0$ such that

$$d\left(u_i, u_j\right) > 3^{k+2}, \text{ for all } 1 \leq j \leq i-1, \tag{43}$$

$$d\left(R_0, u_i\right) > 3^{k+2}, \tag{44}$$

and

$$B\left(u_i; 3^{k+1}\right) \in \sigma. \tag{45}$$

Remark 3.1 Observe that Condition (ii) is stronger than Condition (i) and actually implies the latter. However, for pedagogical clarity, and since (i) gives a nice structural description of the *Christmas tree* that is described in Theorem 3.3, we retain (i). Furthermore, we state (i) before (ii) since, we feel, it is an easier condition to visualize.

Lemma 3.2 *T_0 with properties described above will be a universal tree.*

Proof Recall the definition of universal trees. We start with two trees T_1, T_2 with roots R_1, R_2, and which satisfy the following conditions:

(i) The balls $B(R_1; 3^{k+1}), B(R_2; 3^{k+1})$ satisfy

$$B(R_1; 3^{k+1}) \equiv_{M_0;k} B(R_2; 3^{k+1}). \tag{46}$$

(ii) For some $u_1 \in T_1, u_2 \in T_2$ such that

$$d(R_1, u_1) > 3^{k+2}, \quad d(R_2, u_2) > 3^{k+2}, \tag{47}$$

we have each of $T_1(u_1)$ and $T_2(u_2)$ isomorphic to T_0. If $\varphi_1 : T_0 \to T_1(u_1), \varphi_2 : T_0 \to T_2(u_2)$ are these isomorphisms, then

$$\varphi_1(v_{i;\sigma}) = v^{(1)}_{i;\sigma}, \quad \varphi_2(v_{i;\sigma}) = v^{(2)}_{i;\sigma},$$

for all $\sigma \in \Sigma_{M_0;k}$, $1 \le i \le k$.

Now we give a winning strategy for the Duplicator. This is in a somewhat similar favour to the arguments given in Spencer and Thoma [3]. We assume that since R_1, R_2 are designated vertices, $x_0 = R_1, y_0 = R_2$. Let (x_i, y_i) be the pair chosen from $T_1 \times T_2$ in the i-th move, for $1 \le i \le k$. Now, we claim the following:

The Duplicator can play the game such that, for each $0 \le i \le k$,

- he can maintain

$$B(x_i; 3^{k+1-i}) \equiv_{M_0;k} B(y_i; 3^{k+1-i}),$$

(Our proof only needs

$$B(x_i; 3^{k+1-i}) \equiv_{M_0;k-i} B(y_i; 3^{k+1-i}),$$

but the stronger assumption is a bit more convenient);

- for all $0 \le j < i$ such that $x_j \in B(x_i; 3^{k+1-i})$, the corresponding $y_j \in B(y_i; 3^{k+1-i})$, and vice versa, according to the winning strategy of $EHR_{M_0}[B(x_i; 3^{k+1-i}), B(y_i; 3^{k+1-i}); k]$. Again, this is overkill as one need only consider the Ehrenfeucht game of $k - i$ moves at this point.

We prove this using induction on the number of moves played so far. For $i = 0$, we have chosen $x_0 = R_1, y_0 = R_2$, and we already have imposed the condition

$$B(R_1; 3^{k+1}) \equiv_{M_0;k} B(R_2; 3^{k+1})$$

in (46). So suppose the claim holds for $0 \le j \le i - 1$. Without loss of generality suppose Spoiler chooses $x_i \in T_1$. There are two possibilities:

(i) *Inside move:*

$$x_i \in \bigcup_{j=0}^{i-1} B(x_j; 2 \cdot 3^{k+1-i}). \tag{48}$$

So $x_i \in B(x_l; 2 \cdot 3^{k+1-i})$ for some $0 \le l \le i - 1$. By the induction hypothesis,

$$B(x_l; 3^{k+1-l}) \equiv_{M_0;k} B(y_l; 3^{k+1-l}).$$

Duplicator now follows his winning strategy of $EHR_{M_0}[B(x_l; 3^{k+1-l}), B(y_l; 3^{k+1-l}); k]$ and picks $y_i \in B(y_l; 3^{k+1-l})$. This means that,

$$d(x_i, x_l) < 2 \cdot 3^{k+1-i} \quad \Rightarrow \quad B(x_i; 3^{k+1-i}) \subset B(x_l; 3^{k+1-l}),$$

since $l < i$. In the same way

$$B(y_i; 3^{k+1-i}) \subset B(y_l; 3^{k+1-l}),$$

and further,

$$B(x_i; 3^{k+1-i}) \equiv_{M_0;k} B(y_i; 3^{k+1-i}).$$

This last relation follows from the fact that y_i is chosen corresponding to x_i in the winning strategy of the Duplicator for $EHR_{M_0}[B(x_l; 3^{k+1-l}), B(y_l; 3^{k+1-l}); k]$. Since M_0, as chosen in Eq. (8), is greater than $2 \cdot 3^{k+1-i}$, hence for Duplicator to win $EHR_{M_0}[B(x_l; 3^{k+1-l}), B(y_l; 3^{k+1-l}); k]$, he must be able to win the game played within the smaller balls $B(x_i; 3^{k+1-i})$ and $B(y_i; 3^{k+1-i})$.

(ii) *Outside move:*

$$x_i \notin \bigcup_{j=0}^{i-1} B(x_j; 2 \cdot 3^{k+1-i}). \tag{49}$$

Then we consider $B(x_i; 3^{k+1-i})$ and we know, from (43), (44) and (45), that there exists some $v \in T_2$ such that

$$d(v, y_l) > 3^{k+2}, \quad \text{for all } 0 \leq l \leq i-1,$$

and

$$B(v; 3^{k+1}) \equiv_{M_0;k} B(x_i; 3^{k+1}).$$

We choose $y_i = v$. Note that then we automatically have

$$B(y_i; 3^{k+1-i}) \bigcap \left\{ \bigcup_{j=0}^{i-1} B(y_j; 3^{k+1-i}) \right\} = \phi,$$

and

$$B(x_i; 3^{k+1-i}) \equiv_{M_0;k} B(y_i; 3^{k+1-i}).$$

Once again, Duplicator is choosing y_i so that $B(y_i; 3^{k+1}) \equiv_{M_0;k} B(x_i; 3^{k+1})$, i.e. he wins

$$EHR_{M_0}\left[B(x_i; 3^{k+1}), B(y_i; 3^{k+1}); k\right].$$

Then he must be able to win the game within the smaller balls $B(x_i; 3^{k+1-i})$ and $B(y_i; 3^{k+1-i})$, since his winning involves being able to preserve mutual distances of pairs of nodes up to M_0.

This shows that the Duplicator will win $EHR[T_1, T_2; k]$, which finishes the proof. □

Theorem 3.3 *For each $k \in \mathbb{N}$ there is a universal tree T.*

Proof T will be a *Christmas tree* which is constructed as follows. For each $\sigma \in \Sigma_{M_0:k}$ select and fix a specific ball $B(v; 3^{k+1}) \in \sigma$. For each such σ and each $1 \leq i \leq k$ create disjoint copies $T_{i,\sigma} = B(v_{i,\sigma}; 3^{k+1})$ such that $B(v_{i;\sigma}; 3^{k+1}) \cong B(v; 3^{k+1})$, with the isomorphism mapping $v_{i;\sigma}$ to v. These $B(v_{i;\sigma}; 3^{k+1})$ are the *balls* decorating the Christmas tree. Let $w_{i;\sigma}$ be the *top* vertex of $B(v_{i;\sigma}; 3^{k+1})$. That is, it is that unique node in the ball with no ancestor in the ball. It can be seen that this node is actually the ancestor of $v_{i;\sigma}$ which is at distance 3^{k+1} away from $v_{i;\sigma}$, or in other words, $v_{i;\sigma}$ is a 3^{k+1}-descendant of this node. Let R be the root of T. Draw disjoint paths of length 3^{k+4} from R to each $w_{i;\sigma}$. These will be like the *strings* attaching the balls to the Christmas tree.

We now explain why this T satisfies Conditions (i) and (ii). Once again, for pedagogical clarity, we first show a detailed reasoning why T satisfies (i), although technically, it suffices to verify only (ii). First, observe that the $v_{i;\sigma}$ we have defined in the previous paragraph, for $1 \leq i \leq k$ and $\sigma \in \Sigma_{M_0;k}$, immediately satisfy (40) and (41), since

$$d\left(R, v_{i;\sigma}\right) = d\left(R, w_{i;\sigma}\right) + d\left(v_{i;\sigma}, w_{i;\sigma}\right) = 3^{k+4} + 3^{k+1} > 3^{k+2},$$

for every $\sigma_1, \sigma_2 \in \Sigma_{M_0;k}$, $1 \leq i_1, i_2 \leq k$ with $(\sigma_1, i_1) \neq (\sigma_2, i_2)$, we indeed have

$$d\left(v_{i_1;\sigma_1}, v_{i_2;\sigma_2}\right) = d\left(v_{i_1;\sigma_1}, R\right) + d\left(R, v_{i_2;\sigma_2}\right) > 2 \cdot 3^{k+4} > 3^{k+2}.$$

To see that (42) holds, note that by our construction,

$$B(v_{i;\sigma}; 3^{k+1}) \cong B(v; 3^{k+1}) \in \sigma,$$

with $v_{i;\sigma}$ mapped to v, for all $1 \leq i \leq k$, and for all $\sigma \in \Sigma_{M_0:k}$.

Finally, we verify that (ii) holds. Consider any $1 \leq j \leq k$. Suppose we have selected any $j-1$ vertices $u_1, \dots u_{j-1}$ from T. For any $\sigma \in \Sigma_{M_0:k}$ and $1 \leq i \leq k$, we consider the *branch* of the tree consisting of the ball $B\left(v_{i;\sigma}; 3^{k+1}\right)$ and the string joining R to $w_{i;\sigma}$, and we call that branch*free* if no u_l, $1 \leq l \leq j-1$ is picked from that branch. Since there are k copies of balls for each σ, and $j \leq k$, hence we shall always have at least one *free* branch from each $\sigma \in \Sigma_{M_0:k}$. So we simply choose $u_j = v_{i;\sigma}$ for some i such that the corresponding branch is free.

Since no $u_l,\ 1 \leq l \leq j-1$, belongs to that branch, each of them must be at least as far away from u_j as the root is from $v_{i;\sigma}$. That is, we will have

$$d\left(u_j, u_l\right) > 3^{k+4} + 3^{k+1}; \quad d\left(u_j, R\right) = 3^{k+4} + 3^{k+1}.$$

And of course, by our choice, we would have $B\left(u_j; 3^{k+1}\right) \in \sigma$.

□

4 Probabilities Conditioned on Infiniteness of the Tree

As before, with R the root, $B_T(R; i)$ denotes the neighbourhood of R with radius i, i.e.

$$B_T(R; i) = \{u \in T : d(u, R) < i\}.$$

We define

$$\overline{B_T(R; i)} = \{u \in T : d(u, R) \leq i\}.$$

So, $\overline{B_T(R; i)}$ captures up to the i-th generation of the tree, R being the 0-th generation. For each $i \in \mathbb{N}$ we give a set of equivalence classes Γ_i which will be relatively easy to handle and which we show in Theorem 4.2 is a refinement of $\Sigma_{i:k}$. We set

$$C = \{0, 1, \ldots, k-1, \omega\}. \tag{50}$$

Here ω is a special symbol with the meaning "at least k." That is, to say that there are ω copies of something is to say that there are at least k copies. We set

$$\Gamma_1 = C = \{0, 1, \ldots, k-1, \omega\}. \tag{51}$$

A $\overline{B_T(R; 1)}$ is of type $i \in \Gamma_1$ if the root has i children. Since the game has k rounds, if the roots has x, y children in the two trees with both $x, y \geq k$ then Duplicator wins the modified game. Inductively we now set

$$\Gamma_{i+1} = \{g : \Gamma_i \to C\}. \tag{52}$$

Each child v of the root generates a tree to generation i. This tree belongs to an equivalence class $\sigma \in \Gamma_i$. A $\overline{B_T(R; i+1)}$ has state $g \in \Gamma_{i+1}$ if for all $\sigma \in \Gamma_i$ the root has $g(\sigma)$ children v whose subtree $T(v)$ upto generation i belongs to equivalence class σ, i.e. $T(v)|_i \in \sigma$.

Example 4.1 Consider $k = 4, i = 2$. Then a typical example of $\overline{B_T(R; i)}$ will be: the root has two children with no children, at least four children with one child, three children with two children, no children with three children, and one child with at least four children. Thus $g(0) = 2, g(1) = \omega, g(2) = 3, g(3) = 0, g(\omega) = 1$.

Theorem 4.2 *Γ_i is a refinement on $\Sigma_{i:k}$.*

Proof Let $\overline{B_{T_1}(R_1; i)}, \overline{B_{T_2}(R_2, i)}$ lie in the same Γ_i equivalence class. It suffices to show that Duplicator wins the k-move modified Ehrenfeucht game on these balls. We show this using induction on i.

The case $i = 1$ is immediate. Suppose it holds good for all $i' \leq i - 1$. In the Ehrenfeucht game let Spoiler select $w_1 \in T_1$. Let v_1 be the child of the root such that w_1 belongs to the tree generated by v_1 up to depth $i - 1$, i.e. $T_1(v_1)|_{i-1}$. Duplicator allows Spoiler a free move of v_1. Let σ be the Γ_{i-1} class for $T_1(v_1)|_{i-1}$. In T_2 Duplicator finds a child v_2 of the root R_2 in T_2 such that $T_2(v_2)|_{i-1} \in \sigma$. Duplicator now moves v_2 and then, by induction hypothesis, finds the appropriate response $w_2 \in T_2(v_2)|_{i-1}$ corresponding to w_1. For any further moves by the Spoiler with the same v_1 or v_2, Duplicator plays, inductively, on the two subtrees $T_1(v_1)|_{i-1}, T_2(v_2)|_{i-1}$. And if Spoiler chooses some $y_1 \in B_{T_1}(R_1; i) - T_1(v_1)|_{i-1}$, then again we repeat the same procedure as above. There are only k moves, hence Duplicator can continue in this manner and so wins the Ehrenfeucht game. □

When $\sigma \in \Gamma_i$ we write $\Pr[\sigma], \Pr^*[\sigma]$ for the probabilities, in T_λ, T_λ^* respectively, that $\overline{B_T(R, i)}$ is in equivalence class σ. Let $\Gamma = \Gamma_s$ with $s = 3^{k+1}$.

For any first order A with quantifier depth k let J_A be as in (17). Applying Theorem 4.2 for each $i \in J_A$ the class $\mathcal{B}_i$ splits into finitely many classes $\tau \in \Gamma$. Let K_A denote the set of such classes. The Eq. (18) can be rewritten as

$$\overset{*}{\Pr}[A] = \sum_{\tau \in K_A} \overset{*}{\Pr}[\tau]. \tag{53}$$

For $0 \leq i < k$ set

$$P_i(x) = \Pr[Po(x) = i] = e^{-x} \frac{x^i}{i!}, \tag{54}$$

and set

$$P_\omega(x) = \Pr[Po(x) \geq k] = 1 - \sum_{i=0}^{k-1} P_i(x). \tag{55}$$

We now make use of a special property of the *Poisson* distribution. Let $\Omega = \{1, \ldots, n\}$ be some finite state space. Let $p_i \geq 0$ with $\sum_{i=1}^n p_i = 1$ be some distribution over Ω. Suppose v has *Poisson* mean λ children and each child independently is in state i with probability p_i. The distribution of the number of children of each type is the same as if for each $i \in \Omega$ there were *Poisson* mean

$p_i\lambda$ children of type i and these values were mutually independent. For example, assumming boys and girls equally probable, having *Poisson* mean 5 children is the same as having *Poisson* mean 2.5 boys and, independently, having *Poisson* mean 2.5 girls.

The probability, in T_λ, that the root has u children (including $u = \omega$) is then $P_u(\lambda)$. Suppose, by induction, that $P_\tau(x)$ has been defined for all $\tau \in \Gamma_i$ such that $\Pr(\tau) = P_\tau(\lambda)$. Let $\sigma \in \Gamma_{i+1}$ so that σ is a function $g : \Gamma_i \to C$. In T_λ the root has *Poisson* mean λ children and, for each $\tau \in \Gamma_i$, the i-generation tree rooted at a child is in the class τ with probability $P_\tau(\lambda)$. By the special property above we equivalently say that the root has *Poisson* mean $\lambda P_\tau(\lambda)$ children of type τ for each $\tau \in \Gamma_i$ and that these numbers are *mutually independent*. The probability $P_\sigma(\lambda)$ is then the product, over $\tau \in \Gamma_i$, of the probability that a *Poisson* mean $\lambda P_\tau(\lambda)$ has value $g(\tau)$. Setting

$$P_\sigma(x) = \prod_\tau P_{g(\tau)}(xP_\tau(x)), \tag{56}$$

we have

$$\Pr[\sigma] = P_\sigma(\lambda). \tag{57}$$

Example 4.3 Continuing Example 4.1, set $x_i = e^{-\lambda}\lambda^i/i!$ for $0 \le i < 4$ and $x_\omega = 1 - \sum_{i=0}^{3} x_i$. The root has no child with three children with probability $\exp[-x_3\lambda]$. It has one child with at least four children with probability $\exp[-x_\omega\lambda](x_\omega\lambda)$. It has at least four children with one child with probability $1 - \exp[-x_1\lambda](1 + (x_1\lambda) + (x_1\lambda)^2/2 + (x_1\lambda)^3/6]$. It has two children with no children with probability $\exp[-x_0\lambda](x_0\lambda)^2/2$. It has three children with two children with probability $\exp[-x_2\lambda](x_2\lambda)^3/6$. The probability of the event is then the product of these five values.

While Eq. (57) gives a very full description of the possible $\Pr[\sigma]$ the following less precise description may be more comprehensible.

Definition 4.4 Let $\mathcal{F}$ be the minimal family of function $f(\lambda)$ such that

(i) $\mathcal{F}$ contains the identity function $f(\lambda) = \lambda$ and the constant functions $f_q(\lambda) = q, q \in \mathbb{Q}$.
(ii) $\mathcal{F}$ is closed under finite addition, subtraction and multiplication.
(iii) $\mathcal{F}$ is closed under base e exponentiation. That is, if $f(\lambda) \in \mathcal{F}$ then $e^{f(\lambda)} \in \mathcal{F}$.

We call a function $f(\lambda)$ *nice* if it belongs to $\mathcal{F}$.

In Corollary 4.8 we show that the probability of any first order property, conditioned on the tree being infinite, is actually such a nice function.

Theorem 4.5 *Then for all k and all i, if* $\sigma \in \Gamma_i$ *then* $\Pr[\sigma]$ *is a nice function of* λ.

This is an immediate consequence of the recursion (56).

Example 4.6 The statement “the root has no children which have no children which have no children” is the union of classes σ with $k = 1$, $i = 3$. It has probability $\exp[-\lambda \exp[-\lambda \exp[-\lambda]]]$.

Let T_λ^{fin} denote T_λ conditioned on T_λ being finite. For any k, i and any $\sigma \in \Gamma_i$ let $\Pr^{fin}[\sigma]$ be the probability of event σ in T^{fin}. Assume $\lambda > 1$. Let $p = p(\lambda)$, the probability T_λ is infinite, be given by (1). By duality, T_λ^{fin} has the same distribution as $T_{q\lambda}$, where

$$q(\lambda) = 1 - p(\lambda) = \Pr[T_\lambda \text{ is finite}]. \tag{58}$$

Thus

$$\overset{fin}{\Pr}[\sigma] = P_\sigma(q\lambda). \tag{59}$$

For any k, i and $\sigma \in \Gamma_i$

$$\Pr[\sigma] = \overset{fin}{\Pr}[\sigma]q + \overset{*}{\Pr}[\sigma]p \tag{60}$$

and hence

$$\overset{*}{\Pr}[\sigma] = p^{-1}[\Pr[\sigma] - \overset{fin}{\Pr}[\sigma]q]. \tag{61}$$

For any first order sentence A of quantifier depth k, letting K_A be as in (53),

$$\overset{*}{\Pr}[A] = \sum_{\sigma \in K_A} p^{-1}[\Pr[\sigma] - \overset{fin}{\Pr}[\sigma]q]. \tag{62}$$

Combining previous results gives a description of possible $\Pr^*[A]$.

Theorem 4.7 *Let A be a first order sentence of quantifier depth k. Let K_A be as in (53) Let*

$$f(x) = \sum_{\sigma \in K_A} P_\sigma(x). \tag{63}$$

Then

$$\overset{*}{\Pr}[A] = p^{-1}[f(\lambda) - qf(q\lambda)]. \tag{64}$$

As before, it is also convenient to give a slightly weaker form.

Corollary 4.8 *For any first order sentence A we may express*

$$\overset{*}{\Pr}[A] = p^{-1}[f(\lambda) - qf(q\lambda)] \tag{65}$$

where f is a nice function in the sense of Definition 4.4.

5 Further Results

In this paper, we have so far dealt with Galton–Watson trees with *Poisson* offspring distribution. The results of Sects. 2 and 3 extend to some other classes of offspring distributions. In this section, we outline briefly these extensions. We consider a general probability distribution D on $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$, where p_i is the probability that a typical node in the random tree has exactly i children, $i \in \mathbb{N}_0$. We shall denote the probabilities under this regime by $\Pr_D$. We also assume that the moment generating function of D exists on a non-degenerate interval $[0, \gamma]$ on the real line.

Fix an arbitrary finite T_0 of depth d_0. We assume that $\Pr_D[T_0] > 0$. In other words, this means that if T is the random Galton–Watson tree with offspring distribution D, then $\Pr_D[T \cong T_0] > 0$. Consider the statement

$$A = \{\exists\, v : T(v) \cong T_0\} \vee \{T \text{ is finite}\}. \tag{66}$$

We can show, similar to our results in Sect. 2, that $\Pr_D[A] = 1$, provided (71) holds for some $\alpha \in (0, \gamma]$ and $0 < \epsilon < 1$. Of course, the non-trivial case to consider is when D has expectation greater than 1, as only then does it make sense to talk about the infinite Galton–Watson tree.

The proof of this fact follows the exact same steps as shown in Sect. 2. We consider again a fictitious continuation $X_1, X_2, \ldots$ which are i.i.d. D. For every node i, we let I_i be the indicator for the event $\{T(i) \cong T_0\}$. For a suitable $\epsilon > 0$ that we choose later, we let

$$Y = \sum_{i=1}^{\lfloor \epsilon^{d_0} s \rfloor} I_i, \tag{67}$$

and we define the martingale $Y_i = E[Y | X_1, \ldots, X_i]$ for $1 \leq i \leq s$, with $Y_0 = E[Y]$. Defining g_1 as in Eq. (25), we similarly argue that

$$g_1(x) \leq \lfloor x \rfloor + \sum_{i=1}^{\lfloor x \rfloor} X_i. \tag{68}$$

The only difference is in the estimation of the probability that $g_1(\epsilon x)$ exceeds x. We employ Chernoff bounds again, but we no longer have the succinct form of the moment generating function as in the case of *Poisson*. For any $0 < \alpha \leq \gamma$,

$$\begin{aligned}\Pr[g_1(\epsilon x) > x] &= \Pr[e^{\alpha g_1(\epsilon x)} > e^{\alpha x}] \\ &\leq E[e^{\alpha g_1(\epsilon x)}]e^{-\alpha x} \\ &\leq E[e^{\alpha(\epsilon x + \sum_{i=1}^{\lfloor \epsilon x \rfloor} X_i)}]e^{-\alpha x} \\ &= e^{\alpha \epsilon x} \prod_{i=1}^{\lfloor \epsilon x \rfloor} E[e^{\alpha X_i}]e^{-\alpha x} \\ &= \varphi(\alpha)^{\lfloor \epsilon x \rfloor} e^{-\alpha(1-\epsilon)x}, \end{aligned} \tag{69}$$

where $\varphi(\alpha) = E[e^{\alpha X_1}]$. Since X_1 is non-negative valued, $\varphi(\alpha) > 1$ for $\alpha > 0$, hence we can bound the expression in (69) above by

$$\varphi(\alpha)^{\epsilon x} e^{-\alpha(1-\epsilon)x} = \left\{\varphi(\alpha)^{\epsilon} e^{-\alpha(1-\epsilon)}\right\}^x . \tag{70}$$

If we are able to choose $\alpha > 0$ such that for some $0 < \epsilon < 1$, we have

$$\varphi(\alpha)^{\epsilon} e^{-\alpha(1-\epsilon)} < 1, \tag{71}$$

then the exact same argument as in Sect. 2 goes through, and we have the desired result.

In particular, it is easy to see that (71) is indeed satisfied when D is a probability distribution on a finite state space $\subseteq \mathbb{N}_0$.

The sufficient conditions for a tree to be universal nowhere uses the offspring distribution. Once the results of Sect. 2 hold for a given D, it is not too difficult to see that the conclusion of Remark 1.10 should hold in this regime as well. We hope to return to this more general setting in our future work.

A further object of future study is a more detailed analysis of T_λ at the critical value $\lambda = 1$. While $\Pr^*$ is technically not defined at the critical value, there may well be some approaches via the insipient infinite tree.

References

1. N. Alon, J.H. Spencer, *The Probabilistic Method.* Wiley-Interscience Series in Discrete Mathematics and Optimization, 3rd edn. (Wiley, Hoboken, 2008). ISBN:978-0-470-17020-5, doi:10.1002/9780470277331, http://dx.doi.org/10.1002/9780470277331. With an appendix on the life and work of Paul Erdős
2. J. Spencer, *The Strange Logic of Random Graphs*. Volume 22 of Algorithms and Combinatorics (Springer, Berlin, 2001). ISBN:3-540-41654-4, doi:10.1007/978-3-662-04538-1, http://dx.doi.org/10.1007/978-3-662-04538-1

3. J. Spencer, L. Thoma, On the limit values of probabilities for the first order properties of graphs, in *Contemporary Trends in Discrete Mathematics (Štiřín Castle, 1997)*. Volume 49 of DIMACS Series in Discrete Mathematics & Theoretical Computer Science (American Mathematical Society, Providence, 1999), pp. 317–336
4. R. van der Hofstad, *Random Graphs and Complex Networks*, vol. 1. Lecture Notes (2015). http://www.win.tue.nl/~rhofstad/NotesRGCN.pdf

Crossing-Free Perfect Matchings in Wheel Point Sets

Andres J. Ruiz-Vargas and Emo Welzl

Abstract Consider a planar finite point set P, no three points on a line and exactly one point not extreme in P. We call this a *wheel set* and we are interested in $\mathsf{pm}(P)$, the number of crossing-free perfect matchings on P. (If, contrary to our assumption, all points in a set S are extreme, i.e. in convex position, then it is well-known that $\mathsf{pm}(S) = C_m$, the mth Catalan number, $m := \frac{|S|}{2}$.)

We give exact tight upper and lower bounds on $\mathsf{pm}(P)$ depending on the cardinality of the wheel set P. Simplified to its asymptotics in terms of C_m, these yield

$$\frac{9}{8}C_m(1 + o(1)) \leq \mathsf{pm}(P) \leq \frac{3}{2}C_m(1 + o(1)) \text{ , } m := \frac{|P|}{2}.$$

We characterize the wheel sets (order types) which maximize or minimize $\mathsf{pm}(P)$. Moreover, among all sets S of a given size not in convex position, $\mathsf{pm}(S)$ is minimized for some wheel set. Therefore, leaving convex position increases the number of crossing-free perfect matchings by at least a factor of $\frac{9}{8}$ (in the limit as $|S|$ grows). We can also show that $\mathsf{pm}(P)$ can be computed efficiently.

A connection to origin embracing triangles is briefly discussed.

This research was initiated at the 13th Gremo's Workshop on Open Problems (GWOP), Feldis, Switzerland, June 1–5, 2015. Research of the first author was partially supported by Swiss National Science Foundation grants 200021-137574 and 200020-144531. The second author acknowledges financial support from EuroCores/EuroGiga/ComPoSe, Swiss National Science Foundation grant 20GG21 134318/1.

A.J. Ruiz-Vargas
Mathgeom-DCG, EPF Lausanne, Lausanne, Switzerland
e-mail: andres.ruizvargas@epfl.ch

E. Welzl (✉)
Department of Computer Science, Institute of Theoretical Computer Science, ETH Zürich, Zürich, Switzerland
e-mail: emo@inf.ethz.ch

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_30

1 Introduction

Given finite sets S of points in the plane in general position, i.e. no three points on a line, we are interested in *crossing-free perfect matchings* (CFPMs for short) on S, i.e. perfect matchings on S where in the straight-line geometric embedding on S all segments are pairwise noncrossing. Clearly, for such a perfect matching to exist, $|S|$ has to be even, which we assume further on. We denote by $\mathsf{pm}(S)$ the number of such CFPMs.

In this paper we concentrate on sets in a very special position, namely vertex sets of convex polygons together with one point inside the polygon. But let us first briefly summarize the general situation.

Background Only recently an $O(2^n\text{poly}(n))$ algorithm for computing the number $\mathsf{pm}(S)$ for a set of n points was discovered by Wettstein, [30], with a further improvement to subexponential time by Marx and Miltzow [17]. A tight lower bound of $\mathsf{pm}(S) \geq C_m$ is known, [12], where $m := |S|/2$ and $C_m := \frac{1}{m+1}\binom{2m}{m} = \Theta(\frac{1}{m^{3/2}}4^m)$ is the mth Catalan number. The bound is tight, since equality holds whenever S is in *convex position*, i.e. S is the vertex set of a convex polygon; counting of crossing-free perfect matchings in convex position goes back to 1948, at least, (Motzkin, [18]). In fact, Asinowski [7] proved that equality holds iff S is in convex position, with the only exception of the set of six points consisting of the vertices of a regular pentagon together with its center (and clearly everything of the same order type). Good upper bounds seem elusive, with $O(10.05^n)$, $n := |S|$, the currently best known upper bound (see [24]). Sets with $\Omega(3.093^n)$ CFPMs have been analyzed, see Asinowski and Rote, [8], for this very recent improvement from $\Omega(3^n/\text{poly}(n))$ (obtained for the so-called double-chain configuration, [12]).

There are several other works on crossing-free (sometimes also called plane or noncrossing) perfect matchings, see e.g. [1, 3–5, 9, 15, 23].

Results We go a small step beyond convex position and investigate the case when all but one point z are in convex position, and the extra point z lies in the interior of the convex hull. We call such sets *wheel point sets* or simply *wheel sets* and we reserve the letter "P" for those (to avoid confusion with the general case "S"). For wheel sets P we can draw a clear picture.

We show that $\mathsf{pm}(P)$, P a wheel set, can be computed efficiently (Corollary 3.3). This is a by-product of our main results in Sects. 2, 3, and 4, where we give exact tight upper and lower bounds on $\mathsf{pm}(P)$, P a wheel set, in terms of $m := \frac{|P|}{2}$ (summarized in Theorems 7.1 and 7.2). These yield

$$\frac{9}{8}C_m\left(1+\Theta\left(\frac{1}{m^2}\right)\right) \leq \mathsf{pm}(P) \leq \frac{3}{2}C_m\left(1+\Theta\left(\frac{1}{m^{3/2}}\right)\right) .$$

We can give a characterization of the extremal wheel sets.

With some extra effort (Theorem 5.1 in Sect. 5) including a short excursion to well-formed parentheses strings, we show that wheel sets minimize $\mathsf{pm}(S)$ among

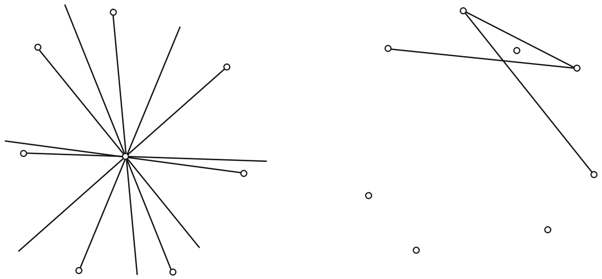


Fig. 1 Wheel sets in symmetric (*left*) and barely-in configuration (*right*)

all sets S not in convex position. In this way we strengthen Asinowski's result in [7], in the sense that not being in convex position does not only enforce an increase in the number of CFPMs (with the one notable exception for 6 points mentioned above), but the number $\mathsf{pm}(P)$ asymptotically goes up by at least a constant factor of $\frac{9}{8}$.

Apart from the concrete results on CFPMs, we view this as a contribution to the understanding of the combinatorics of a set of points "around" a given point, or equivalently, a set of vectors with the origin as a positive linear combination. In Sect. 6 we exemplify this by pointing out a connection to origin embracing triangles which relates to n vertex polytopes in $(n-3)$-space.

Since the results get developed (and are scattered) over several sections of the paper, we conclude with a summary in Sect. 7.

Two special configurations: Symmetric and barely-in In order to give more insight, we would like to be able to address two very specific wheel sets, see Fig. 1. The first one has the property that every line through the extra point and an extreme point is halving, i.e. it is separating the remaining points into two parts of equal size (for example, the vertices of a regular $(2m-1)$-gon together with its center). We call this a wheel set *in symmetric configuration*, and it appears like a good candidate for extremal behavior. In fact, we have mentioned already one instance, namely the symmetric configuration of six points, which was the only nonconvex position with minimal number of CFPMs. Beware, as we will show, this is a misleading hint! For eight points the symmetric configuration maximizes, for ten points it minimizes, for twelve points it maximizes – but this pattern does not continue. Indeed, $2m$ points, m even, in symmetric configuration always maximize, but if $m \geq 7$ is odd, these configurations with $2m$ points play no extremal role anymore.

The second special type of wheel set is the one we obtain by taking $2m$ points in convex position and pushing one of the points, z, inside the convex hull, but barely so. We call this a wheel set in *barely-in configuration*. For such a set P, $\mathsf{pm}(P) = C_m + C_{m-1} = \frac{5}{4}C_m(1 + o(1))$, can easily be established: Note that as we push z inside over the segment e connecting its two neighbors on the convex hull, there are no extra crossings (among points connecting segments) generated. However we lose some crossings with this edge e (in the terminology of [20], the point set is crossing-dominated by convex position). Hence, all C_m matchings from the initial convex position are preserved as crossing-free, and it is easily seen that we gain exactly those which use the edge e (where z was pushed over). The number of those is exactly C_{m-1}, since the barely-in set without the endpoints of e is in convex position.

While this barely-in configuration might be a suspect for minimizing $\mathsf{pm}(P)$ (see experience from counting triangulations below), this is never true. In fact, for six points barely-in *maximizes*. That is, the minimizers are something between symmetric and barely-in, the same for maximizers with $2m$ points, $m \geq 7$ odd.

A relation to counting triangulations It is interesting to compare the situation to the number of triangulations, $\mathsf{tr}(S)$, of a planar point set S. We have $\mathsf{tr}(S) = C_{n-2}$ for S a set of n points in convex position (a classical result that goes back to a question of Euler). Conveniently, Randall et al. [22] have investigated this for wheel configurations.

If P is barely-in, then the number of triangulations is smaller than in convex position, namely $C_{n-2} - C_{n-3}$, and this gives already the minimum possible number for wheel configurations. For n even, the maximum is attained for the symmetric configuration [22, Corollary 6]. Their main tool is a continuous motion argument, by which they can show that as the extra point moves from barely-in deeper into the point set, the number of triangulations increases.

From what we learn below, it becomes clear that such an argument cannot work in our setting, since, in fact, $\mathsf{pm}(P)$ fluctuates under such an into-the-set motion (see Fig. 4 in Sect. 4 for an illustration). Instead we consider the $2m - 1$ lines through the extra point and each of the extreme points, and we note the number of remaining points on the two sides of each of these lines. This information can be summarized in a vector that can be used to compute $\mathsf{pm}(P)$. Some basic simple knowledge from k-set theory turns out to come handy. The counting itself does first some overcounting and subtracts the excess.

Notation $\mathbb{N}$ denotes the set of positive integers, $\mathbb{N}_0$ the set of nonnegative integers, and $\mathbb{R}$ the set of real numbers. For a set $S \subset \mathbb{R}^d$, $\mathsf{conv}(S)$ stands for the convex hull of S.

Catalan numbers Our results and analysis rely heavily on the use of Catalan numbers C_m, $m \in \mathbb{N}_0$ (cf. Stanley's recent book [25]). We list a few properties in "Appendix: Catalan Facts".

Order types We will sometimes refer to the notion of order types in this paper, in order to claim that there is always "basically" a unique wheel set maximizing (minimizing) $\mathsf{pm}(P)$. We decided not to introduce this notion any further here and refer to the literature, cf. [2, 13].

2 A First Identity Via Weighted Diagonals

From now on P is in wheel configuration with $n := 2m := |P|$. We write P as $Q \cup \{z\}$, with Q the $2m-1$ extreme points in P and, hence, z the special nonextreme point. We use *extra point* or *inner point* to address this point z.

An edge connecting two points in Q is either an edge of $\mathsf{conv}(Q)$ or a *diagonal*. Such a diagonal e partitions the remaining $2m-3$ points of Q into two sets (on the sides of the line through e), one of even size, say k, and one of odd size, say ℓ.

We need two notions: The *weight of* e, denoted by $\mu(e)$, is defined to be $C_{k/2}C_{(\ell+1)/2}$. And the diagonal e is called *active* if the extra point z lies on the side with the even number k of points – we call this the *even side of* e; the other side of e is called the *odd side of* e.

Lemma 2.1

$$\mathsf{pm}(P) = (2m-1)C_{m-1} - \sum_{e \text{ active diagonal}} \mu(e)\ .$$

The identity can easily be explained by giving meaning to the terms involved.

In a perfect matching on P there has to be a unique point y that matches with the extra point z. We call this edge yz the *extra edge* in the perfect matching.

Observation 2.2 $(2m-1)C_{m-1}$ *is the number of perfect matchings on P that are crossing-free except for possible crossings involving the extra edge. (We call such perfect matchings* diagonal-crossing-free.*)*

This is obvious, since we have $2m-1$ ways to choose the companion y of z and then we have C_{m-1} ways to complete the matching among the remaining $2m-2$ points in $Q \setminus \{y\}$ with no crossing among those.

Next we consider such a diagonal-crossing-free perfect matching M that is not crossing-free. That is, there are diagonals crossing the extra edge. The first such diagonal encountered when moving along yz from y to z is called *responsible* (for M not being crossing-free). Note that every diagonal-crossing-free perfect matching is either crossing-free or there is a unique responsible diagonal crossing the extra edge. Hence, the previous and the following observation establish Lemma 2.1.

Observation 2.3 *A diagonal e is responsible for crossings in some diagonal-*CFPM*s iff it is active, and if so, it is responsible in exactly* $\mu(e)$ *such perfect matchings.*

Proof If e is responsible for a crossing in a diagonal-CFPM, then all points in Q on the side of e containing the extra point z must be matched to each other. It follows that there has to be an even number of them, i.e. e is active.

Let us assume that z lies on the even side of e and R is the set of points in Q that lie on the other, i.e. odd side of e. Now if e is responsible for the crossings in a diagonal-crossing-free perfect matching M, then by its choice (the first crossing encountered from $y \in R$), the matching M restricted to $R \cup \{z\}$ is a CFPM on $R \cup \{z\}$, a set in convex position. There are $C_{(\ell+1)/2}$ CFPMs on $R \cup \{z\}$, $\ell := |R|$, and all of these can be completed with e and a CFPM on $Q\setminus(R\cup e)$ to a diagonal-crossing-free matching, where e is responsible for crossing the extra edge. The number of such perfect matchings is $C_{k/2}C_{(\ell+1)/2} = \mu(e)$, $k := |Q\setminus(R\cup e)|$. □

3 Diagonals Incident to a Given Extreme Point

For handling the sum $\sum_{e \text{ active diagonal}} \mu(e)$, it appears advantageous to consider the contributions of all the edges incident to a given extreme point $q \in Q$. So let us set

$$\lambda(q) = \lambda(q, P) := \sum_{q\in e \text{ active diagonal}} \mu(e)\ .$$

Then we have

$$\sum_{e \text{ active diagonal}} \mu(e) = \frac{1}{2}\sum_{q\in Q}\lambda(q)\ . \tag{1}$$

We will see that a simple parameter determines $\lambda(q)$. For that consider the line through q and z (the extra point) which splits $P\setminus\{q,z\}$ into sets of size $i>0$ and $j>0$, $i+j=2m-2$. We are interested in the difference between i and j and we call $|i-j|/2$ the *index*, ind(q), *of* q. Note that $0\le \operatorname{ind}(q)\le m-2$. The maximal value occurs for $i=1$ and $j=2m-3$, while $\operatorname{ind}(q)=0$ indicates that the line through qz halves $P\setminus\{q,z\}$. So, for example, if P is the symmetric configuration, then $\operatorname{ind}(q)=0$ for all $q\in Q$.

Lemma 3.1

$$\lambda(q) = \lambda_{\operatorname{ind}(q),m} := C_m - 2C_{m-1} + \lambda'_{\operatorname{ind}(q),m}$$

where, for $0\le i\le m-2$,

$$\lambda'_{i,m} := \begin{cases} 0 & \textit{if } m\equiv i \pmod 2 \textit{ and} \\ C_{(m-1-i)/2}C_{(m-1+i)/2} & \textit{if } m\not\equiv i \pmod 2. \end{cases}$$

Proof See Figs. 2 and 3 for illustration. Let $q_1,\ldots, q_{2m-2}$ be the clockwise order around q of the points in $Q\setminus\{q\}$ such that qq_1 and qq_{2m-2} are edges of the convex hull of P. Let $j=\operatorname{ind}(q)+m-1$. Without loss of generality we may assume that the line passing through the points q and z is not horizontal and has the points $\{q_1,\ldots, q_j\}$ to its left and $\{q_{j+1},\ldots,q_{2m-2}\}$ to its right. Split the diagonals incident to q into the pairs $(qq_2, qq_3), (qq_4, qq_5),\ldots, (qq_{2m-4}, qq_{2m-3})$; in each pair the two diagonals have the same weight. If $m\equiv i$ then exactly one diagonal from each pair

$$\boxed{}\ \boxed{C_1C_3}\ \boxed{C_1C_3}\ \boxed{C_2C_2}\ \boxed{C_2C_2}\ \boxed{C_3C_1}\ \boxed{C_3C_1}\ \boxed{}$$

Diagonals and convex hull edges incident to an extreme point, diagonals with their weight.

$$\boxed{}\ \underline{\boxed{C_1C_3}}\ \boxed{C_1C_3}\ \underline{\boxed{C_2C_2}}\ \boxed{z}\ \underline{\boxed{C_2C_2}}\ \boxed{C_3C_1}\ \underline{\boxed{C_3C_1}}\ \boxed{}$$

The extra point in the middle, hence $\mathrm{ind}(q) = 0$ and
$\lambda(q) = C_1C_3 + C_2C_2 + C_2C_2 + C_1C_3 = X + \underbrace{C_2C_2}_{\lambda_{0,5}}.$

$$\underbrace{\boxed{}\ \underline{\boxed{C_1C_3}}\ \boxed{C_1C_3}}_{3}\ \boxed{z}\ \underbrace{\boxed{C_2C_2}\ \underline{\boxed{C_2C_2}}\ \boxed{C_3C_1}\ \underline{\boxed{C_3C_1}}\ \boxed{}}_{5}$$

$\mathrm{ind}(q) = |5-3|/2 = 1$ and $\lambda(q) = C_1C_3 + C_2C_2 + C_3C_1 = X.$

$$\boxed{}\ \underline{\boxed{C_1C_3}}\ \boxed{z}\ \underline{\boxed{C_1C_3}}\ \boxed{C_2C_2}\ \underline{\boxed{C_2C_2}}\ \boxed{C_3C_1}\ \underline{\boxed{C_3C_1}}\ \boxed{}$$

$\mathrm{ind}(q) = |6-2|/2 = 2$ and $\lambda(q) = C_1C_3 + C_1C_3 + C_2C_2 + C_3C_1 = X + \underbrace{C_1C_3}_{\lambda_{2,5}}.$

$$\boxed{}\ \boxed{z}\ \boxed{C_1C_3}\ \underline{\boxed{C_1C_3}}\ \boxed{C_2C_2}\ \underline{\boxed{C_2C_2}}\ \boxed{C_3C_1}\ \underline{\boxed{C_3C_1}}\ \boxed{}$$

$\mathrm{ind}(q) = |7-1|/2 = 3$ and $\lambda(q) = C_1C_3 + C_2C_2 + C_3C_1 = X.$

Fig. 2 Scenarios around an extreme point for $|P| = 10$, i.e. $m = 5$ odd; the *boxes* indicate the eight edges to the other extreme points, the $\boxed{z}$ the position of the extra point z among them. Active edges are indicated by *underlined boxes*. X stands for $C_1C_3 + C_2C_2 + C_3C_1 = C_5 - 2C_4$

is active and it follows that

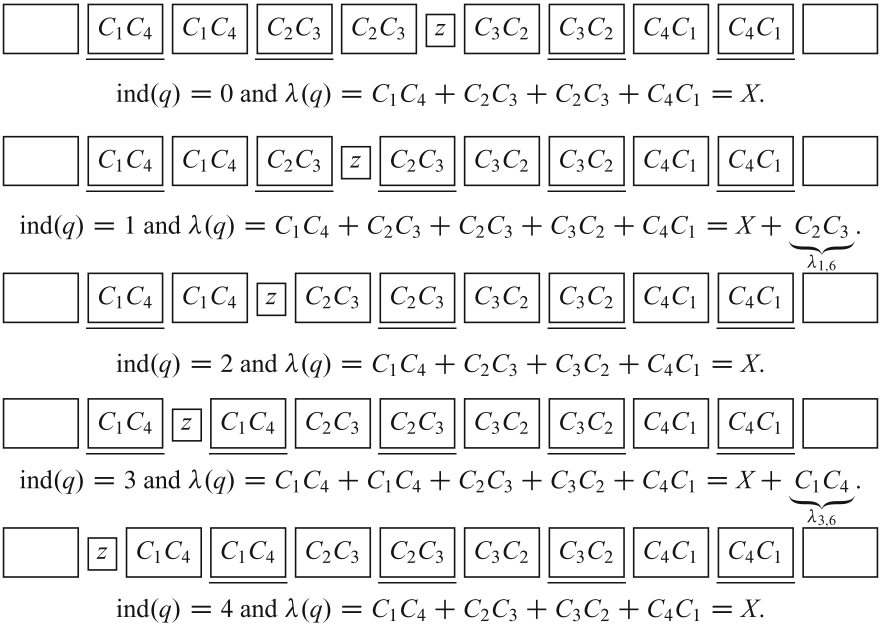
$$\lambda(q) = \sum_{2\le\ell\le 2m-4,\ \ell\text{ even}} \mu(qq_\ell) = \sum_{2\le\ell\le 2m-4,\ \ell\text{ even}} C_{(2m-2-\ell)/2}C_{\ell/2}$$
$$= \sum_{\ell=1}^{m-2} C_{m-1-\ell}C_\ell = C_m - 2C_{m-1},$$

where the last equality follows from Fact A.3 (Segner Recurrence).

On the other hand, if $m \not\equiv i$, exactly one diagonal from each pair is active with the exception of the pair (qq_j, qq_{j+1}) for which both diagonals are active, therefore to the value calculated in the previous case we must add $\mu(qq_j) = C_{(m-1-\mathrm{ind}(q))/2}C_{(m-1+\mathrm{ind}(q))/2}$ and the lemma follows. □

Theorem 3.2 *For* $m := \frac{|P|}{2}$,

$$\mathsf{pm}(P) = \frac{3}{2}C_m - \frac{1}{2}\sum_{q\in Q}\lambda'_{\mathrm{ind}(q),m}\,.$$


C1C4 C1C4 C2C3 C2C3 z C3C2 C3C2 C4C1 C4C1
C1C4 C1C4 C2C3 z C2C3 C3C2 C3C2 C4C1 C4C1
C1C4 C1C4 z C2C3 C2C3 C3C2 C3C2 C4C1 C4C1
C1C4 z C1C4 C2C3 C2C3 C3C2 C3C2 C4C1 C4C1
z C1C4 C1C4 C2C3 C2C3 C3C2 C3C2 C4C1 C4C1

$$\mathrm{ind}(q) = 0 \text{ and } \lambda(q) = C_1C_4 + C_2C_3 + C_2C_3 + C_4C_1 = X.$$

$$\mathrm{ind}(q) = 1 \text{ and } \lambda(q) = C_1C_4 + C_2C_3 + C_2C_3 + C_3C_2 + C_4C_1 = X + \underbrace{C_2C_3}_{\lambda_{1,6}}.$$

$$\mathrm{ind}(q) = 2 \text{ and } \lambda(q) = C_1C_4 + C_2C_3 + C_3C_2 + C_4C_1 = X.$$

$$\mathrm{ind}(q) = 3 \text{ and } \lambda(q) = C_1C_4 + C_1C_4 + C_2C_3 + C_3C_2 + C_4C_1 = X + \underbrace{C_1C_4}_{\lambda_{3,6}}.$$

$$\mathrm{ind}(q) = 4 \text{ and } \lambda(q) = C_1C_4 + C_2C_3 + C_3C_2 + C_4C_1 = X.$$

Fig. 3 Scenarios around an extreme point for $|P| = 12$, i.e. $m = 6$ even. Here $X := C_1C_4 + C_2C_3 + C_3C_2 + C_4C_1 = C_6 - 2C_5$

Proof According to Lemma 2.1, Equation (1), and Lemma 3.1 we have

$$\begin{aligned}
\mathsf{pm}(P) &= (2m-1)C_{m-1} - \frac{1}{2}\left((2m-1)(C_m - 2C_{m-1}) + \sum_{q\in Q}\lambda'_{\mathrm{ind}(q),m}\right)\\
&= (2m-1)(2C_{m-1} - \frac{1}{2}C_m) - \frac{1}{2}\sum_{q\in Q}\lambda'_{\mathrm{ind}(q),m}\\
&= \frac{3}{2}C_m - \frac{1}{2}\sum_{q\in Q}\lambda'_{\mathrm{ind}(q),m} \qquad \text{(see Fact A.5).}
\end{aligned}$$

□

A simple rotational scan around z allows to compute the indices of all points in Q in linear time. However, beyond this we have to compute the Catalan numbers and the sum in Theorem 3.2 involving numbers with linear (in n) number of bits. The recurrence in Fact A.2 is useful for computing the first m Catalan numbers with $O(m)$ operations.

Corollary 3.3 *If the extreme points in P are given in order sorted along the boundary of the convex hull of P, then* $\mathsf{pm}(P)$ *can be computed in linear time plus*

$O(n)$ arithmetic operations (addition, multiplication, integer division) on $O(n)$-bit integers.

Corollary 3.4

(1) If m is even, then the symmetric configuration P gives $\mathsf{pm}(P) = \frac{3}{2}C_m$. This quantity is the maximum attained for wheel configurations and m even.

(2) If m is odd, then the symmetric configuration P gives $\mathsf{pm}(P) = \frac{3}{2}C_m - \frac{2m-1}{2}C^2_{(m-1)/2}$.

Proof From Theorem 3.2 we know $\mathsf{pm}(P) = \frac{3}{2}C_m - \frac{1}{2}\sum_{q\in Q}\lambda'_{\mathrm{ind}(q),m}$. The symmetric configurations have $\mathrm{ind}(q) = 0$ for all $q \in Q$. Now, since

$$\lambda'_{0,m} := \begin{cases} 0 & , m \text{ even, and} \\ (C_{(m-1)/2})^2 & , m \text{ odd,} \end{cases}$$

the claimed values readily follow. Moreover, since $\lambda'_{i,m} \geq 0$ for all i, the maximality for m even holds. □

Frequency vectors While we have resolved already the case of maximizing the number of crossing-free matchings for m even, the remaining cases require a slightly more subtle treatment. If we let $f_i = f_i(P)$ be the number of extreme points with index i, then Lemma 3.1 and Theorem 3.2 tell us that (and how) the *frequency vector* $(f_0, f_1, \ldots f_{m-2})$ determines $\mathsf{pm}(P)$. In preparation of our extremal analysis, we have to understand which frequency vectors are possible.

Lemma 3.5 $(f_0, f_1, \ldots f_{m-2}) \in \mathbb{N}_0^{m-1}$ *is the frequency vector of some wheel set iff*

(i) $\sum_{i=0}^{m-2} f_i = 2m - 1$,
(ii) *f_i is even for $i > 0$, and*
(iii) *If $f_i > 0$ then $f_j > 0$ for all $j < i$.*

An extra property we get as a consequence of the above requirements is that f_0 is always odd (and thus nonzero).

Proof There is a standard argument for this from k-set theory that goes back to the early papers, e.g. [11]. One basically lets a line rotate about the extra point z and observes how this line dissects the points in Q as points in Q are encountered.

First we show that every frequency vector of a wheel set $P = Q \cup \{z\}$ in wheel configuration with $2m$ points satisfies (i–iii) in the statement of the lemma. (i) is obvious.

Note that there is always a point $q \in Q$ such that the line through zq halves $Q \setminus \{q\}$; this follows by observing a directed line rotating about z while observing the points left of it. If such a line starts going through zr, $r \in Q$, with k points to the left, then after rotating by π, we have $n-2-k$ to its left. Since the number of points to the left changes by $+1$ or -1 in each step, there must be a point $q \in Q$ where there are $(k + (n-2-k))/2 = n/2 - 1$ points to the left of the line through zq, a halving line.

Now start with a halving line h (through zq), directed from z towards q, and rotate h (a line, not a ray!) clockwise, while enumerating the points in Q as encountered until q is met again. This gives a sequence $(q = q_1, q_2, \dots q_{2m-1}, q_{2m} = q)$ (we have $\{q_1, q_2, \dots q_{2m-1}\} = Q$). For $i = 1, 2, \dots 2m - 1$, let h_i be the directed line h when it goes through q_i and let $h_{i+1/2}$ be some line through z between h_i and h_{i+1}. Moreover, let $h_{1/2}$ be a line just before the line through q_1.

We define a function $g : \{\frac{i}{2} \mid i = 1, 2, \dots 4m - 1\} \to \mathbb{R}$, by setting $g(x)$ as the number of points left of line h_x, where points on h_x contribute only $\frac{1}{2}$ to the count. We see that for $i \in \{1, 2, \dots 2m - 1\}$,

(a) $g\left(\frac{1}{2}\right) = m - \frac{1}{2}, g\left(2m - \frac{1}{2}\right) = m + \frac{1}{2}$,
(b) $|g\left(i - \frac{1}{2}\right) - g\left(i + \frac{1}{2}\right)| = 1$,
(c) $g(i) = \frac{g\left(i-\frac{1}{2}\right)+g\left(i+\frac{1}{2}\right)}{2} \in \mathbb{N}$,
(d) for $i \neq 1$, $|g(i) - g(i-1)| \leq 1$, and
(e) $\mathrm{ind}(q_i) = |g(i) - m|$.

It follows that m appears an odd number of times as value of $g(i)$, $i \in \{1, 2, \dots, 2m - 1\}$ and any other value appears an even number of times; hence, by (e), $\mathrm{ind}(q_i) = 0$ for an odd number of points and any value other than 0 appears an even number of times. This establishes (ii) in the statement of the lemma. (d) and (e) show (iii).

We still have to show sufficiency of (i–iii). For that let us go back to the proof of necessity and define a sign sequence $\varepsilon_i \in \{-1, +1\}$, $i \in \{1, 2, \dots, 2m - 1\}$, where $\varepsilon_i = +1$ if line h_i is directed from z to q_i and $\varepsilon_i = -1$, otherwise. In our set-up, $\varepsilon_1 = +1$. We have, for i integral,

$$g\left(i + \frac{1}{2}\right) = g\left(i - \frac{1}{2}\right) + \varepsilon_i = g\left(\frac{1}{2}\right) + \varepsilon_1 + \varepsilon_2 + \cdots \varepsilon_i = m - \frac{1}{2} + \sum_{j=1}^{i} \varepsilon_j \,,$$

$$g(i) = m - \frac{1}{2} + \sum_{j=1}^{i-1} \varepsilon_j + \frac{\varepsilon_i}{2} \text{ , and } \mathrm{ind}(q_i) = \left| \sum_{j=1}^{i-1} \varepsilon_j + \frac{\varepsilon_i}{2} - \frac{1}{2} \right| \,.$$

We see that $\mathrm{ind}(q_i) = \mathrm{ind}(q_{i-1})$ iff $\varepsilon_i \neq \varepsilon_{i-1}$. Given $(f_0, f_1, \dots, f_{m-2}) \in \mathbb{N}_0^{m-1}$ with f_k the last nonzero entry, we now build a sequence (where "$f_0 \times \mathrm{ind} = 0$" should be read as "gives f_0 times a point with index 0", etc.)

$$\underbrace{+(-+)^{\frac{f_0-1}{2}}}_{f_0 \times \mathrm{ind}=0} \underbrace{+(-+)^{\frac{f_1-2}{2}}}_{(f_1-1) \times \mathrm{ind}=1} \underbrace{+(-+)^{\frac{f_2-2}{2}}}_{(f_2-1) \times \mathrm{ind}=2} \cdots \underbrace{+(-+)^{\frac{f_k-2}{2}}}_{(f_k-1) \times \mathrm{ind}=k} \quad \underbrace{-^k}_{\mathrm{ind}=k,\,(k-1),\dots 1,\text{ one each}} \qquad (2)$$

Conditions (i–iii) guarantee that this is a well-defined sequence of length $2m-1$ with m times "+" and $(m - 1)$ times "−". It is easy to obtain a set Q of $2m - 1$ points on a circle, together with z its center, so that we get this sequence of ε_is; e.g. choose $2m - 1$ unit vectors $v_1, v_2, \dots, v_{2m-1}$, as sorted around the origin $\mathbf{0}$, all positive x-

coordinate, and then set $q_i := \mathbf{0} + \varepsilon_i v_i$. The balance of $+$s and $-$s guarantees also that indeed the line through the first point is halving, which was essential in our argument.

This completes the proof of the lemma as stated. However, as we are at it, let us see how much choice we had in setting up the sequence (2). Obviously, there is choice. For example, both

$$\overset{0}{+}\overset{1}{+}\overset{1}{-}\overset{1}{+}\overset{1}{-}\overset{1}{+}\overset{2}{+}\overset{2}{-}\overset{1}{-} \quad \text{and} \quad \overset{0}{+}\overset{1}{+}\overset{1}{-}\overset{1}{+}\overset{2}{+}\overset{2}{-}\overset{1}{-}\overset{1}{+}\overset{1}{-}$$

generate the vector $(1, 6, 2, 0)$. The reader familiar with order types will agree that the realizations of these sequences give distinct order types of ten points.

However, there are relevant examples where there is no choice:

$$(2m-1, 0, 0, \dots 0) \quad \longrightarrow \quad +\,(-+)^{m-1} \quad \text{(symmetric configuration)}$$

and

$$(1, \underbrace{2, 2, \dots 2}_{\ell\times}, 2k, 0, 0, \dots 0) \quad \longrightarrow \quad +^{\ell+2}\,(-+)^{k-1}-^{\ell+1}$$

These are exactly the frequency vectors which – as we will see below – occur in minimizing and maximizing the number of CFPMs, and thus we will get unique order types realizing those. □

Again, in different disguises the argument is well-known. We wanted to go through it again, since we are not aware that the sufficiency was ever addressed before.

Let us note that the barely-in configuration has $(1, 2, 2, \dots 2, 4)$ as its frequency vector, while the symmetric configuration has frequency vector $(2m-1, 0, 0, \dots 0)$.

Corollary 3.6 *For $m \geq 3$ odd, any wheel set P of $2m$ points satisfies*

$$\mathsf{pm}(P) \leq \frac{3}{2}C_m - \frac{1}{2}C^2_{(m-1)/2} = \frac{3}{2}C_m\left(1 + \Theta\left(\frac{1}{m^{3/2}}\right)\right)$$

with equality iff P has frequency vector $(1, 2(m-1), 0, 0, \dots 0)$.

Proof Recall $\mathsf{pm}(P) = \frac{3}{2}C_m - \frac{1}{2}\sum_{q\in Q}\lambda'_{\mathrm{ind}(q),m}$ from Theorem 3.2. We have $\sum_{q\in Q}\lambda'_{\mathrm{ind}(q),m} \geq \lambda'_{0,m} = C^2_{(m-1)/2}$ since there is at least one q with $\mathrm{ind}(q) = 0$. This lower bound is attained iff all the other points have index 1; any other point of index 0 or a point of index 2 will increase the sum.

The asymptotic version of the bound follows from Fact A.1. □

In particular, for $m = 3$, barely-in with frequency vector $(1, 4)$ maximizes with seven CFPMs – the unique appearance of barely-in as extremal configuration (except for $m = 2$ where there is only one order type for a wheel set).

4 Extremal Analysis: Lower Bounds

The following analysis is tedious. As a small "justification" we provide Fig. 4: It shows that as the extra point moves inwards (among 9 other points in convex position), the number of CFPMs does not change in a simple way (e.g. monotone).

We have to treat the cases of $|P|/2$ even or odd separately.

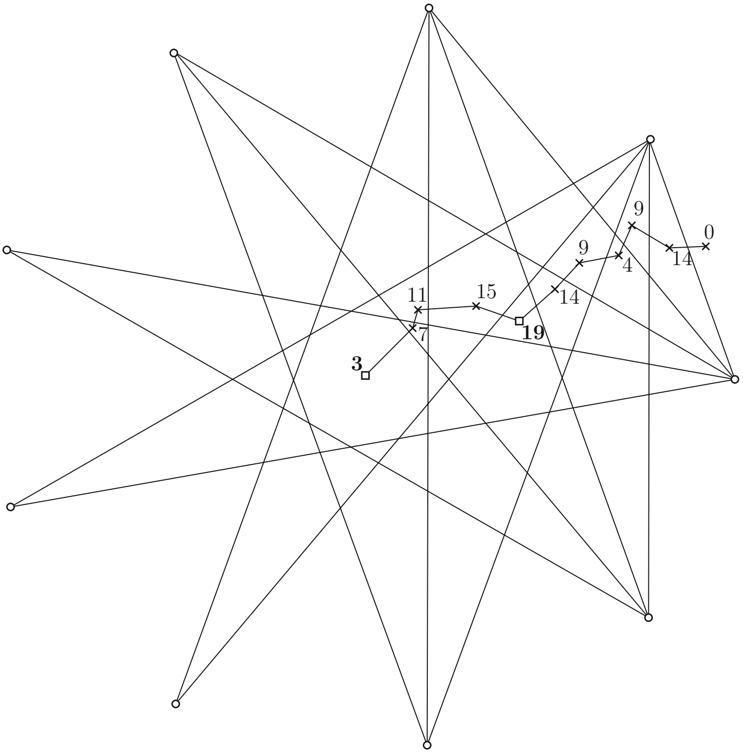


Fig. 4 Nine points plus the extra point moving in from extremal position (ten points in convex position) to central position (symmetric configuration). The numbers indicate the excess of the number of CFPMs beyond $C_5 = 42$

4.1 m Even

If $m = 2$, i.e. we are dealing with four points, then there is only one order type of four points in wheel configuration (which happens to be both of the type symmetric and barely-in). It has three CFPMs, both minimal and maximal.

We first rewrite Theorem 3.2 in terms of the frequency vector.

Lemma 4.1 *Let* $|P| = 2m$*, m even, with frequency vector* $(f_0, f_1, \ldots f_{m-2})$*. Then*

$$\begin{aligned} \mathsf{pm}(P) &= \frac{3}{2}C_m - \frac{1}{2}\sum_{i=0,\text{ odd}}^{m-2} f_i\, C_{(m-1-i)/2}C_{(m-1+i)/2} \\ &= \frac{3}{2}C_m - \underbrace{\frac{1}{2}\sum_{j=0}^{m/2-2} f_{2j+1}\, C_{m/2-j-1}C_{m/2+j}}_{\Delta=\Delta_{\text{even}}:=}\ . \end{aligned}$$

Lemma 4.2 *Let* $|P| = 2m$*,* $m \geq 4$ *even. Then*

$$\mathsf{pm}(P) \geq C_m + C_{m-1} - 2C_{m-2} = \frac{9}{8}C_m\left(1 + \Theta\left(\frac{1}{m^2}\right)\right).$$

This bound is tight and attained iff the frequency vector is $(1, \overbrace{2, 2, \ldots 2}^{m-4}, 6, 0)$. (See Fact A.7 for the asymptotic version of the lower bound.)

Proof $\mathsf{pm}(P)$ is minimized, if Δ – as defined in Lemma 4.1 – is maximized. We observe that the terms $C_{m/2-j-1}C_{m/2+j}$ grow as j grows (see Fact A.4). Therefore, we would like to distribute as much weight as possible to the larger indices of the frequency vector. However, that has a price, since we have to have an entry of 2 at least up to the last nonzero entry. Every other of these entries comes with a coefficient of 0, others with positive coefficients that are however smaller than the last positive. It follows that the only candidates of frequency vectors with extremal property are

$$(\underbrace{1}_{f_0}, \overbrace{2, 2, \ldots 2}^{2j}, \underbrace{2(m-2j-1)}_{f_{2j+1}}, \overbrace{0, 0, \ldots 0}^{m-2j-3}) \text{ for } j = 0, 1, \ldots (m-4)/2$$

where

$$\begin{aligned} \Delta = \Delta_j^m &:= C_{m/2-1}C_{m/2} + C_{m/2-2}C_{m/2+1} + \cdots C_{m/2-j}C_{m/2+j-1} \\ &\quad + (m-2j-1)C_{m/2-j-1}C_{m/2+j}\ . \end{aligned}$$

We want to understand how Δ_j^m behaves as j grows. For that we calculate

$$\begin{aligned}
&\Delta_{j+1}^m - \Delta_j^m \\
&= C_{m/2-j-1}C_{m/2+j} + (m-2j-3)C_{m/2-j-2}C_{m/2+j+1} \\
&\quad -(m-2j-1)C_{m/2-j-1}C_{m/2+j} \\
&= (m-2j-3)C_{m/2-j-2}C_{m/2+j+1} - (m-2j-2)C_{m/2-j-1}C_{m/2+j} \\
&= (2k-1)C_{k-1}C_\ell - 2kC_kC_{\ell-1} \ \ (k := m/2-j-1,\ \ell := m/2+j+1)
\end{aligned}$$

This expression is less than 0 (i.e. $\Delta_{j+1}^m < \Delta_j^m$) iff

$$\begin{aligned}
&2\ell - 1 < 3k \qquad \text{(employ Fact A.6)} \\
\iff \quad & j < (m/2-4)/5\,.
\end{aligned}$$

Hence, the $(\Delta_j^m)_j$-sequence initially falls (at least for m large enough), and then it begins to grow. We are interested in Δ large and therefore, only the first and the last element of the sequence qualify. These are (from vector $(1, 2(m-1), 0, 0, \ldots 0)$)

$$\Delta_0^m = (m-1)C_{m/2-1}C_{m/2}$$

and (from vector $(1, 2, 2, \ldots 2, 6, 0)$)

$$\begin{aligned}
\Delta_{(m-4)/2}^m &= C_{m/2-1}C_{m/2} + C_{m/2-2}C_{m/2+1} + \cdots C_2C_{m-3} + 3C_1C_{m-2} \\
&= \frac{1}{2}C_m - C_{m-1} + 2C_{m-2}
\end{aligned}$$

If we can show $\Delta_{(m-4)/2}^m \geq \Delta_0^m$ then

$$\frac{3}{2}C_m - \left(\frac{1}{2}C_m - C_{m-1} + 2C_{m-2}\right) = C_m + C_{m-1} - 2C_{m-2}$$

establishes the result.

So let us have a look at the sequence $(\Delta_0^m, \Delta_{(m-4)/2}^m)_{m=4,6,\ldots}$

m	4	6	8	10	12
Δ_0^m	6	50	490	5 292	60 984
$\Delta_{(m-4)/2}^m$	6	52	550	6 396	78 812

Note that $\Delta_0^m = \Theta\left(m\left(\frac{2^m}{m^{3/2}}\right)^2\right) = \Theta\left(\frac{4^m}{m^2}\right)$ while $\Delta_{(m-4)/2}^m = \Theta\left(\frac{4^m}{m^{3/2}}\right)$. Therefore, $\Delta_{(m-4)/2}^m \geq \Delta_0^m$ is definitely true for large enough m, but we need to argue that this is true for all $m \geq 4$. This fact is established in Lemma 4.3 below; in fact, $\Delta_{(m-4)/2}^m$

is strictly larger than Δ_0^m for $m \geq 6$ (and for $m = 4$ we have $\Delta_{(m-4)/2}^m = \Delta_0^m$ simply because $(m-4)/2 = 0$).

We still have to establish the uniqueness of the minimizing frequency vector. Note that this is not obvious since $\Delta_{j+1}^m = \Delta_j^m$ may hold (e.g. $\Delta_0^8 = \Delta_1^8 = 490$). We know from the analysis above that this happens iff $j = (m/2-4)/5$, and if – on top – $j = (m-4)/2 - 1$. Then a second minimizing frequency vector would exist. However,

$$(m/2-4)/5 = (m-4)/2 - 1 \Leftrightarrow m = \frac{11}{2}\,.$$

□

Lemma 4.3 *For $m \geq 6$, even,*

$$\frac{1}{2}C_m - C_{m-1} + 2C_{m-2} = \Delta_{(m-4)/2}^m > \Delta_0^m = (m-1)C_{m/2-1}C_{m/2}\,.$$

Proof The claim is true for $m = 6$; we employ induction. Let us write $A(m)$ for $\Delta_{(m-4)/2}^m$ and $B(m)$ for Δ_0^m. Note that we have $A(m) > B(m)$ for all even $m \geq 6$ provided $A(6) \geq B(6)$ (which is true, $52 > 50$) and

$$A(m) > 0, B(m) > 0, \text{ and } \frac{\frac{A(m+2)}{A(m)}}{\frac{B(m+2)}{B(m)}} - 1 \geq 0 \text{ for all even } m \geq 6.$$

We have[1]

$$\frac{\frac{A(m+2)}{A(m)}}{\frac{B(m+2)}{B(m)}} - 1 = \frac{1}{4}\frac{(4m^3 - m^2 - 26m - 66)m}{(m^2 - 2m + 2)(m+3)(m+1)^2}\,.$$

The denominator is positive for all positive m. For positive m the sign of the numerator – and thus the whole expression – is determined by $s(m) = 4m^3 - m^2 - 26m - 66$. Note that $s(3) = -45$ and $s(4) = 70$. That is, $s(4) > 0$, and $s(x)$ (as a real polynomial) has a positive real root between $x = 3$ and $x = 4$. According to Descartes' Sign Rule,[2] this has to be the only positive root and therefore $s(x) > 0$ for all $x \geq 4$. □

[1] We acknowledge the use of Maple for these straightforward but tedious calculations.

[2] Descartes's Sign Rule: The number of positive real roots of a real polynomial is at most the number of sign changes of the coefficients, as read from largest to smallest power (ignoring zero-coefficients). In our polynomial, there is only one sign change, from $4x^3$ to $-x^2$.

4.2 m Odd

Lemma 4.4 *Let $|P| = 2m$, m odd, with frequency vector $(f_0, f_1, \ldots f_{m-2})$. Then*

$$\mathsf{pm}(P) = \frac{3}{2}C_m - \underbrace{\frac{1}{2}\sum_{j=0}^{(m-3)/2} f_{2j}\, C_{(m-1)/2-j}C_{(m-1)/2+j}}_{\Delta=\Delta_{\text{odd}}:=} \ .$$

Lemma 4.5 *Let $|P| = 2m$, $m \geq 7$ odd. Then*

$$\mathsf{pm}(P) \geq C_m + C_{m-1} - 2C_{m-2} = \frac{9}{8}C_m\left(1 + \Theta\left(\frac{1}{m^2}\right)\right).$$

This bound is tight and attained iff the frequency vector is $(1, \overbrace{2, 2, \ldots 2}^{m-4}, 6, 0)$.

If $|P| = 6$ then $\mathsf{pm}(P) \geq 5$ and if $|P| = 10$ then $\mathsf{pm}(P) \geq 45$, both times values attained in the symmetric configuration.

Proof Again we have to analyze for which frequency vectors the value of Δ – as in Lemma 4.4 – is maximized. Plausible frequency vectors for this to happen are

$$(2m-1, 0, 0, \ldots 0) \quad \text{(the symmetric configuration)}, \tag{3}$$

where $\Delta = \Delta_0^m := \frac{2m-1}{2}(C_{(m-1)/2})^2$, and

$$(\underbrace{1}_{f_0}, \overbrace{2, 2, \ldots 2}^{2j-1}, \underbrace{2(m-2j)}_{f_{2j}}, \overbrace{0, 0, \ldots 0}^{m-2j-2}) \ \text{ for } j = 1, 2, \ldots (m-3)/2,$$

where Δ equals

$$\Delta_j^m := \frac{1}{2}(C_{(m-1)/2})^2 + C_{(m-1)/2-1}C_{(m-1)/2+1} + \cdots C_{(m-1)/2-j+1}C_{(m-1)/2+j-1}$$
$$+ (m-2j)C_{(m-1)/2-j}C_{(m-1)/2+j} \ .$$

Beware that plugging 0 for j in Δ_j^m does not give Δ_0^m! For $m = 3$ we have Δ_0^m only, which is therefore the maximum. That is, $\mathsf{pm}(P)$ is minimized in the symmetric configuration – not too surprising, since we have here C_3 CFPMs.

We first show $\Delta_0^m > \Delta_1^m$ for odd $m \geq 5$:

$$\begin{aligned}
\Delta_0^m - \Delta_1^m =& \frac{2m-1}{2}(C_{(m-1)/2})^2 - \left(\frac{1}{2}(C_{(m-1)/2})^2 + (m-2)C_{(m-1)/2-1}C_{(m-1)/2+1}\right)\\
=&(m-1)(C_{(m-1)/2})^2 - (m-2)(C_{(m-1)/2})^2 \frac{\frac{m+1}{2}\,m}{(m-2)\,\frac{m+3}{2}} \text{ (Fact A.2)}\\
=&(C_{(m-1)/2})^2\left(m-1-\frac{m(m+1)}{m+3}\right)\\
=&(C_{(m-1)/2})^2\,\frac{m-3}{m+3} > 0 \quad \text{for } m > 3.
\end{aligned}$$

Now, for $j \geq 1$, we consider

$$\begin{aligned}
\Delta_{j+1}^m - \Delta_j^m &= C_{(m-1)/2-j}C_{(m-1)/2+j} + (m-2j-2)C_{(m-1)/2-j-1}C_{(m-1)/2+j+1}\\
&\quad -(m-2j)C_{(m-1)/2-j}C_{(m-1)/2+j}\\
&= (m-2j-2)C_{(m-1)/2-j-1}C_{(m-1)/2+j+1}\\
&\quad -(m-2j-1)C_{(m-1)/2-j}C_{(m-1)/2+j}\\
&= (2k-1)C_{k-1}C_\ell - 2kC_kC_{\ell-1} \quad \left(k := \frac{m-1}{2} - j,\ \ell := \frac{m+1}{2} + j\right)
\end{aligned}$$

That is, we have the same setting as before in the proof of Lemma 4.2 and it follows that $(\Delta_j^m)_j$ first decreases and then increases; the "switch" happens at $2\ell - 1 = 3k$ iff $j = (m-3)/10$. Hence, we have reduced the possible extremal candidates to the first and last elements $\Delta_0^m = \frac{2m-1}{2}(C_{(m-1)/2})^2$ and, for $j = \frac{m-3}{2}$,

$$\begin{aligned}
\Delta_{(m-3)/2}^m &= \frac{1}{2}(C_{(m-1)/2})^2 + C_{(m-1)/2-1}C_{(m-1)/2+1} + \cdots C_2C_{m-3} + 3C_1C_{m-2}\\
&= \frac{1}{2}C_m - C_{m-1} + 2C_{m-2}
\end{aligned}$$

for frequency vector $(1, 2, 2, \ldots 2, 6, 0)$.

First entries in the sequence $(\Delta_0^m, \Delta_{(m-3)/2}^m)_{m=3,5,\ldots}$ are

m	3	5	7	9	11
Δ_0^m	2.5	18	162.5	1 666	18 522
$\Delta_{(m-3)/2}^m$	2.5	17	166.5	1 859	22 321

We see that for $m = 5$ the symmetric configuration has larger Δ, thus the smaller number of CFPMs. Starting from $m = 7$, the frequency vector $(1, 2, 2, \ldots 2, 6, 0)$

gives the smallest number of CFPMs up to $m = 11$, and also beyond, as to be shown in Lemma 4.6 below.

For uniqueness, we observe that there was only one candidate for $m = 3$ and for $m = 5$ we have shown $\Delta_0^m > \Delta_1^m$. Our last concern is that for $m \geq 7$, $\Delta_{j+1}^m = \Delta_j^m$ (this holds for $j = (m-3)/10$) and $j = (m-3)/2 - 1$. But

$$(m-3)/10 = (m-3)/2 - 1 \Leftrightarrow m = \frac{11}{2}\,.$$

□

Lemma 4.6 *For $m \geq 7$, odd,*

$$\frac{1}{2}C_m - C_{m-1} + 2C_{m-2} = \Delta_{(m-3)/2}^m > \Delta_0^m = \frac{2m-1}{2}(C_{(m-1)/2})^2\,.$$

Proof The claim is true for $m = 7$; again, we use induction. We write $A(m)$ for $\Delta_{(m-3)/2}^m$ and $B(m)$ for Δ_0^m. We have $A(7) = 166.5 > 162.5 = B(7)$ and we obtain

$$\frac{\frac{A(m+2)}{A(m)}}{\frac{B(m+2)}{B(m)}} - 1 = \frac{1}{4}\frac{8m^5 - 6m^4 - 53m^3 - 71m^2 + 60m - 18}{m^2(m^2 - 2m + 2)(2m + 3)(m + 2)}$$

which we will prove nonnegative for $m \geq 7$. The denominator is always positive for positive m, and thus the sign of the numerator $s(m) = 8m^5 - 6m^4 - 53m^3 - 71m^2 + 60m - 18$ determines the sign of the whole expression for positive m. Note that $s(3) = -450$ and $s(4) = 2\,350$. That is, $s(x)$ has a positive real root between $x = 3$ and $x = 4$. However, now Descartes Rule allows for at most three positive roots.

We have a look at the derivative $s'(x) = 40m^4 - 24m^3 - 159m^2 - 142m + 60$ and observe $s'(0) = 60$, $s'(1) = -225$ and $s'(3) = 795$. Hence, $s'(x)$ has a positive root between 0 and 1, and another one between 1 and 3, and there is no further positive root due to Descartes. But that shows that the derivative of $s(x)$ is never 0, thus remains positive, beyond its root between 3 and 4 and therefore, s itself stays positive after that. □

5 Wheel Sets Minimize for Nonconvex Position

We show that wheel sets S minimize $\mathsf{pm}(S)$ among all sets which are not in convex position. (Recall that we assume general position.) In this way we will validate the claim that the number of CFPMs for nonconvex position goes up by a factor of at least $9/8$ – in the limit as the number of points grows – compared with convex position.

Theorem 5.1 *For all* $n := 2m$, $2 \le m \in \mathbb{N}$, *we have*

$$\min_{S \text{ not convex}, \, |S|=n} \mathsf{pm}(S) = \min_{P \text{ wheel set}, \, |P|=n} \mathsf{pm}(P) .$$

Let us consider S, a set not in convex position, which we write as $S = \{z, p_1, p_2, \dots p_{n-1}\}$ with z one of the nonextreme points in S; w.l.o.g. let z be the origin $\mathbf{0}$. We compare $\mathsf{pm}(S)$ with $\mathsf{pm}(S')$, where $S' := \{z, p'_1, p'_2, \dots p'_{n-1}\}$ with $p'_i := \frac{1}{\|p_i\|} p_i$; hence S' is a wheel set (note that S' cannot contain duplicate points since no three points in S are on a common line). Therefore, the following lemma establishes the desired extremal property of wheel sets among sets not in convex position (Theorem 5.1).

Lemma 5.2 *For sets* S *and* S' *as above,* $\mathsf{pm}(S) \ge \mathsf{pm}(S')$.

Proof In fact, we fix the companion, say p_1 and p'_1, resp., of z in the perfect matching and show that S' cannot possibly offer more such CFPMs than S.

Let r_0 be the ray emanating from z containing p_1 (and p'_1), let $\mathcal{M}$ be the CFPMs of S containing zp_1 with no edge crossed by r_0, and, similarly, let $\mathcal{M}'$ be the CFPMs of S' containing zp'_1 not crossed by r_0 (this is implied by non-crossing for S'). We plan to prove

$$|\mathcal{M}| \ge |\mathcal{M}'|$$

which yields the claim from the beginning of the proof, since $\mathcal{M}'$ is exactly the set of CFPMs of S' with edge zp'_1, while $\mathcal{M}$ is a subset of the CFPMs of S with edge zp_1.

In the next step, for the sake of comparison, we partition $\mathcal{M}$ and $\mathcal{M}'$ according to the following rules.

Let $\mathcal{M}_0$ be the set of matchings in $\mathcal{M}$ which are not crossed by the *opposite* ray $-r_0$; so these matchings avoid the line through z and p_1 (other than zp_1 contained in this line). Define $\mathcal{M}'_0$ correspondingly as a subset of $\mathcal{M}'$. Let X and Y, resp., be the set of points of S on the two sides of the line through z and p_1. Then the matchings in $\mathcal{M}_0$ are exactly those composed of a CFPM of X with a CFPM of Y. Since we know that $\mathsf{pm}(R) \ge C_{|R|/2}$ for every set R, we see that

$$|\mathcal{M}_0| \ge C_{|X|/2} C_{|Y|/2} .$$

(Note here that $|X|$ and $|Y|$ may be odd. In order to include that case, we follow the convention that $C_x = 0$ for $x \notin \mathbb{N}_0$.) In an analogous analysis for S', we get sets X' and Y' on the two sides of the line through z and p'_1, where $|X'| = |X|$ and $|Y'| = |Y|$. And since X' and Y' are in convex position, we have

$$|\mathcal{M}'_0| = C_{|X'|/2} C_{|Y'|/2} = C_{|X|/2} C_{|Y|/2}$$

and conclude $|\mathcal{M}_0| \ge |\mathcal{M}'_0|$.

We proceed to the matchings that are crossed by ray $-r_0$. Consider such a matching M in $\mathcal{M}$. Starting with $r = -r_0$, rotate r about z in counterclockwise direction until for the first time no edge in M is crossed, where we consider an intersection in an endpoint of an edge not as crossing. With this in mind, note that in its final position, r goes through some point p of S. Proceed in the same fashion in clockwise direction, yielding a point q; p and q have to lie on opposite sides of the line through z and p_1. We put M into a set $\mathcal{M}_{p,q}$. In this way we have

$$\mathcal{M} = \mathcal{M}_0 \,\dot{\cup}\, \dot{\bigcup_{p,q}} \mathcal{M}_{p,q} \quad \text{(disjoint union)},$$

where p goes through all points on one side of the line though z and p_1, and q goes through the points on the other side. Sets $\mathcal{M}'_{p',q'}$ can be defined accordingly. Our final goal is to show $|\mathcal{M}_{p,q}| \geq |\mathcal{M}'_{p',q'}|$ for all pairs p, q under consideration.

Let us first determine $\mathcal{M}'_{p',q'}$. It is easy to see that these are exactly the matchings, where the first edge encountered when moving from z along $-r_0$ is $p'q'$. In particular, if $p'q'$ crosses r_0 (and not $-r_0$), then $\mathcal{M}'_{p',q'} = \emptyset$. Otherwise, the rays r_0, $r_{p'}$ (from z through p') and $r_{q'}$ (from z through q'), form three convex cones. We partition $S' \backslash \{z, p'_1\}$ into three sets X', Y', and Z', where (i) Z' is the set of points in $S' \backslash \{z, p'_1\}$ in the *closed*[3] cone delimited by $r_{p'}$ and $r_{q'}$, (ii) X' is the set in the *open* cone delimited by r_0 and $r_{p'}$, and (iii) Y' is the set in the *open* cone delimited by r_0 and $r_{q'}$. We see that

$$|\mathcal{M}'_{p',q'}| = \mathsf{pm}(X')\mathsf{pm}(Y')\mathsf{pm}(Z' \setminus \{p', q'\}) = C_{|X'|/2} C_{|Y'|/2} C_{|Z'|/2-1} .$$

As for $\mathcal{M}_{p,q}$, the case of pq crossing r_0 is easy, since this happens iff $p'q'$ crosses r_0; then $\mathcal{M}'_{p',q'} = \emptyset$ and $|\mathcal{M}_{p,q}| \geq |\mathcal{M}'_{p',q'}|$ comes for free.

We are ready for the most interesting situation of r_0, r_p, and r_q forming three convex cones. Let sets X, Y, and Z in these cones be defined in an analogous fashion as above. Then

$$|\mathcal{M}_{p,q}| = \mathsf{pm}(X)\mathsf{pm}(Y)\mathsf{cpm}_z(Z) \geq C_{|X|/2} C_{|Y|/2} \mathsf{cpm}_z(Z) .$$

where $\mathsf{cpm}_z(Z)$, z not in the convex hull of Z, denotes the set of all CFPMs on Z such that every line through z that intersects the interior of $\mathsf{conv}(Z)$ crosses at least one edge in the matching. We call such a matching *z-covering*. Since we will show below that $\mathsf{cpm}_z(Z) \geq C_{|Z|/2-1}$, this gives us the relation $|\mathcal{M}_{p,q}| \geq |\mathcal{M}'_{p',q'}|$ (note $|X| = |X'|$, $|Y| = |Y'|$, and $|Z| = |Z'|$), therefore $|\mathcal{M}| \geq |\mathcal{M}'|$ and the lemma is established. □

We call a perfect matching on a set S *covering* if the vertical projection of all edges in the matching on the x-axis is a connected set. Given a set S and a point z

[3] We choose this cone closed, since we want p' and q' in Z', and later p and q in Z.

outside $\mathsf{conv}(S)$, an appropriate projective transformation (sending a line containing z and avoiding the convex hull of S to infinity) transforms z-covering CFPMs exactly to covering CFPMs.

5.1 Covering Crossing-Free Perfect Matchings

Let $\mathsf{cpm}(S)$ denote the number of covering CFPMs of S. We set out to prove $\mathsf{cpm}(S) \geq C_{|S|/2-1}$.

If S is in convex position and, moreover, the first and last point (w.r.t. x-coordinate) form an edge of $\mathsf{conv}(S)$, then we can easily see that $\mathsf{cpm}(S) = C_{|S|/2-1}$. This is true, since in this configuration a CFPM is covering iff first and last point are matched. For S in convex position without the extra condition, covering CFPMs have a more involved structure and, at first glance, it is not clear what their number should be. As a by-product of our analysis, $\mathsf{cpm}(S) = C_{|S|/2-1}$ will follow also here.

Let us briefly recall that a string w over the alphabet $\{\langle, \rangle\}$ is a *well-formed parentheses string*, WFP for short, if it contains the same number of $\langle$'s and $\rangle$'s, and no prefix of w has more $\rangle$'s than $\langle$'s. (1) and (2) below are basic well-known properties of WFPs, (3) is a trivial consequence.

Proposition 5.3

(1) The number of WFP*s of length* $2m$ *is* C_m.

(2) For every nonempty WFP w *exactly one of the following two holds:*

(i) $w = uv$ *for nonempty* WFP*s* u *and* v *(*w *is called* decomposable*).*
(ii) $w = \langle u \rangle$ *for a* WFP u *(*w *is called* indecomposable*).*

(3) The number of indecomposable WFP*s of length* $2m$ *is* C_{m-1}.

Given a perfect matching M on S, let us label the left endpoints (w.r.t. x-coordinate) of matching edges by $\langle$ and the right endpoints by $\rangle$. Project the points vertically to the x-axis and read their labels from left to right, which gives a string σ_M.

Observation 5.4

(1) σ_M *is a* WFP.
(2) M *is a covering* CFPM *iff it is a* CFPM *and* σ_M *is indecomposable.*

With this in mind, here is the only missing link to the required inequality.

Lemma 5.5 *Let* S *be a set of* $2m$ *points in general position, no two points of same* x*-coordinate. For every* WFP w *of length* $2m$ *there is a* CFPM M *on* S *with* $\sigma_M = w$.

Proof Let M' be a perfect matching – not necessarily crossing-free – on S with $\sigma_{M'} = w$; it is easily seen that such an M' exists. If M' has a crossing between edges ab and cd (where a and c are the left endpoints of the edges) then replace these two edges by ad and cb. These two edges do not cross (since every set of four points has at most one crossing perfect matching), and the projected string of $\langle$'s and $\rangle$'s did not change. While there are crossings, iterate this step. The process will eventually stop, since the Euclidean length of the matching decreases in every step (a simple consequence of the triangle inequality). Hence, we end up with a CFPM M with $\sigma_M = \sigma_{M'} = w$.

(Alternatively, we could have chosen M simply as the perfect matching with $\sigma_M = w$ of smallest Euclidean length and argue that it cannot contain crossings). □

Let us mention, that the proof follows the well-known procedure of untangling crossings in spanning tours or matchings, see e.g. [27], where it is shown that the described untangling process terminates in at most $O(n^3)$ steps.

Lemma 5.6 *Let S, $|S| = 2m$, be a set of points in general position, with no two points of equal x-coordinate. We have*

$$\begin{aligned} \mathsf{cpm}(S) &\geq C_{m-1} \quad \textit{and} \\ \mathsf{cpm}(S) &= C_{m-1} \quad \text{if } S \text{ is in convex position.} \end{aligned}$$

Proof Every WFP w of length $2m$ has at least one CFPM M on S with $\sigma_M = w$. Since there are C_{m-1} indecomposable WFPs, there must be at least C_{m-1} covering CFPMs on S.

For the second fact, recall that S has C_m CFPMs, the same number as the number of WFPs of length $2m$. Therefore, for every WFP w we have exactly one CFPM M_w of S with $\sigma_{M_w} = w$. Precisely the C_{m-1} CFPMs M_w with w indecomposable are covering. □

The reader may have observed already, that the known inequality $\mathsf{pm}(S) \geq C_{|S|/2}$ (from [12]) follows readily as well.

6 Origin Embracing Triangles with Weights

We embed our results in a different context independent of CFPMs. For that, let Q be a planar point set with $\mathbf{0} \in \mathsf{conv}(Q)$ and $Q \cup \{\mathbf{0}\}$ in general position. Let $\mathcal{T}_{\text{embr}}$ denote the set of triples $t \in \binom{Q}{3}$ with $\mathbf{0} \in \mathsf{conv}(t)$, the set of *origin embracing triangles*.

This set $\mathcal{T}_{\text{embr}}$ of origin embracing triangles is of interest, since via the Gale transform there is a bijection between $\mathcal{T}_{\text{embr}}$ and the facets of the convex hull of some appropriate $n := |Q|$ points in $\mathbb{R}^{n-3}$, a simplicial polytope with at most n vertices in $\mathbb{R}^{n-3}$. (In fact, the natural extension to origin embracing simplices in $\mathbb{R}^d$ has an analogous correspondence to simplicial $(n-d-1)$-polytopes, cf. [28, 29, 31].) Other works considering origin embracing triangles or simplices can be found e.g. in [10, 14, 16, 19].

This correspondence to polytopes shows (via the so-called Upper Bound Theorem for convex polytopes, cf. [31]) the following tight bounds.

$$n-2 \le |\mathcal{T}_{\mathrm{embr}}| \le \binom{\lfloor \frac{n}{2}+1 \rfloor}{3} + \binom{\lceil \frac{n}{2}+1 \rceil}{3} \tag{4}$$

Note that $t = \{p, q, r\}$ is in $\mathcal{T}_{\mathrm{embr}}$ iff none of the vectors p, q and r can be expressed as a nonnegative linear combination of the others. Therefore, the three open cones defined by the three rays from $\mathbf{0}$ to p, q, and r, resp., are convex, and they partition $Q \setminus \{p, q, r\}$ into three sets of size a_t, b_t and c_t, respectively.

For functions $f : \mathbb{N}_0 \to \mathbb{R}$ we consider the sum $\sum_{t \in \mathcal{T}_{\mathrm{embr}}} f(a_t) f(b_t) f(c_t)$ and its extremal values. The bounds in (4) above deal with the case of the constant function $f(k) \equiv 1$.

Before we proceed to the choice of f that connects to CFPMs in wheel configurations, we offer the surprising fact that there is a positive function where the sum depends on $|Q|$ only, independent of the configuration.

Observation 6.1

$$\sum_{t \in \mathcal{T}_{\mathrm{embr}}} C_{a_t} C_{b_t} C_{c_t} = C_{|Q|-2}$$

Proof Note that $C_{|Q|-2}$ is the number of triangulations of a convex $|Q|$-gon. Choose Q' as the set of points $q' := \frac{1}{\|q\|} q$, $q \in Q$; this transformation does not change the embracing triangles, but gives a set in convex position. Every triangulation of Q' has a unique triangle (with vertices $\{p', q', r'\} \in \binom{Q'}{3}$) containing $\mathbf{0}$ and thus $t = \{p, q, r\} \in \mathcal{T}_{\mathrm{embr}}$. t appears in this role for exactly $C_{a_t} C_{b_t} C_{c_t}$ triangulations. □

Observation 6.2

$$\sum_{t \in \mathcal{T}_{\mathrm{embr}}} C_{a_t/2} C_{b_t/2} C_{c_t/2} = \mathsf{pm}(Q' \cup \{\mathbf{0}\}) - C_m$$

with $Q' := \{\frac{1}{\|q\|} q \mid q \in Q\}$ and $m := \frac{|Q|+1}{2}$. (We assume $C_x = 0$ for $x \notin \mathbb{N}_0$.)

Proof $P := Q' \cup \{\mathbf{0}\}$ is in wheel configuration with $\mathbf{0}$ the extra point. Recall that the sum to the left is the same for Q and Q'.

In every CFPM M of P, $\mathbf{0}$ has a companion $q' \in Q'$. Let us first consider the case where the line containing $q'\mathbf{0}$ crosses some edge of M and let $p'r'$ be the first edge in M encountered when moving from $\mathbf{0}$ in the direction opposite to q'. The unordered triple $\{q', p', r'\}$ is an embracing triangle in $\mathcal{T}_{\mathrm{embr}}$, and this triple appears in this role for exactly $3C_{a_t/2} C_{b_t/2} C_{c_t/2}$ CFPMs. The factor 3 stems from the fact, that each of the three points can represent the companion of $\mathbf{0}$, while the other two represent the first edge hit.

Therefore, $3 \cdot \sum_{t \in \mathcal{T}_{\mathrm{embr}}} C_{a_t/2} C_{b_t/2} C_{c_t/2}$ is the number of CFPMs of P except for those, where the line through the extra edge $q'\mathbf{0}$ crosses no other edge of the matching. Let k and ℓ be the number of points in Q' on the two sides of the line through $q'\mathbf{0}$ (hence, $k + \ell = |Q'| - 1 = 2(m - 1)$). Then the number of "forgotten" CFPMs with extra edge $q'\mathbf{0}$ is $C_{k/2}C_{\ell/2}$, which we have previously (Lemma 3.1) defined as $\lambda'_{\mathrm{ind}(q'),m}$ (since $\mathrm{ind}(q') = |k - \ell|/2$; assuming $k \geq \ell$, check $(m - 1 - \frac{k-\ell}{2})/2 = \ell/2$ and $(m - 1 + \frac{k-\ell}{2})/2 = k/2$). We have

$$\begin{aligned}
\sum_{t\in\mathcal{T}_{\mathrm{embr}}} C_{a_t/2}C_{b_t/2}C_{c_t/2} &= \frac{1}{3}\left(\mathsf{pm}(Q' \cup \{\mathbf{0}\}) - \sum_{q'\in Q'} \lambda'_{\mathrm{ind}(q'),m}\right)\\
\text{(by Theorem 3.2)} &= \frac{1}{3}\left(\left[\frac{3}{2}C_m - \frac{1}{2}\sum_{q'\in Q'} \lambda'_{\mathrm{ind}(q'),m}\right] - \sum_{q'\in Q'} \lambda'_{\mathrm{ind}(q'),m}\right)\\
&= \frac{1}{2}C_m - \frac{1}{2}\sum_{q'\in Q'} \lambda'_{\mathrm{ind}(q'),m}\\
\text{(by Theorem 3.2)} &= \mathsf{pm}(Q' \cup \{\mathbf{0}\}) - C_m
\end{aligned}$$

□

From the introduction we recall Asinowski's characterization of sets of $2m$ points with C_m CFPMs, [7], and our bounds for wheel sets translate to (1) and (2), resp., in the corollary below.

Corollary 6.3

(1) $\sum_{t\in\mathcal{T}_{\mathrm{embr}}} C_{a_t/2}C_{b_t/2}C_{c_t/2} = 0$ *iff* $|Q|$ *is even or* $|Q| = 5$ *and every line through a point* $q \in Q$ *and* $\mathbf{0}$ *halves* $Q \setminus \{q\}$.

(2) $\frac{1}{8}C_m(1 + o(1)) \leq \sum_{t\in\mathcal{T}_{\mathrm{embr}}} C_{a_t/2}C_{b_t/2}C_{c_t/2} \leq \frac{1}{2}C_m(1 + o(1))$ *for* $m := \frac{|Q|+1}{2}$.

We find this a curious connection, but, admittedly, we cannot supply any motivation for considering such weighted sums over embracing triangles or simplices.

7 Summary

Starting off from the formula $\mathsf{pm}(P) = \frac{3}{2}C_m - \frac{1}{2}\sum_{q\in Q} \lambda'_{\mathrm{ind}(q),m}$ (Theorem 3.2) it was easy to obtain a linear time[4] algorithm for computing $\mathsf{pm}(P)$ (assuming the extreme points sorted). Together with a characterization of possible frequency vectors (Lemma 3.5), tight upper bounds were established (Corollaries 3.4 and 3.6).

[4] Assuming constant time for arithmetic operations!

Theorem 7.1 *Let P be a planar point set of even size in wheel configuration with $m := |P|/2$. Then*

$$\mathsf{pm}(P) \le \begin{cases} \frac{3}{2}C_m & m \text{ even, and} \\ \frac{3}{2}C_m - \frac{1}{2}C^2_{(m-1)/2} = \frac{3}{2}C_m\left(1+\Theta\left(\frac{1}{m^{3/2}}\right)\right) & m \text{ odd.} \end{cases}$$

The symmetric configuration is the unique order type maximizing for m even, and the unique order type with frequency vector $(1, 2(m-1), 0, 0, \ldots 0)$ *is the only one maximizing for m odd.*

Drawing a complete picture for lower bounds (both for wheel and for not convex position) was a bit tedious, primarily because of small-number effects (Lemmas 4.2 and 4.5, and Theorem 5.1; for the claim of unique order types see last paragraphs of proof of Lemma 3.5).

Theorem 7.2 *Let P be a wheel set of even size $2m$. Then*

$$\mathsf{pm}(P) \begin{cases} \ge C_m + C_{m-1} - 2C_{m-2} = \frac{9}{8}C_m\left(1+\Theta\left(\frac{1}{m^2}\right)\right) & m \ne 2, 3, 5, \\ = 3 & m = 2, \\ \ge 5 & m = 3, \text{ and} \\ \ge 45 & m = 5. \end{cases}$$

The symmetric configurations are the unique order types which minimize for $m = 2, 3, 5$. The unique order types realizing frequency vectors $(1, 2, 2, \ldots 2, 6, 0)$ *are the only ones minimizing, otherwise.*

The lower bounds hold for any point set not in convex position.

We have summarized some numbers for small m in Table 1. Note there is also another small-number effect for the relation of symmetric vs. barely-in configurations if m is odd, which does not exhibit its "normal" relation before $n = 2m = 46$.

Relation to counting triangulations, etc An essential step in our proof was the establishment of the formula in Theorem 3.2 and one might be curious whether similar identities are possible for other quantities. Indeed, let $P = Q \cup \{z\}$ be in wheel configuration, z the extra point, $n := |P|$, with $\ell(q)$ the number of points in Q to the left of the directed line from q to z and $\mathrm{r}(q)$ the number of points to the right. Then the number of triangulations of P, $\mathsf{tr}(P)$, can be shown to satisfy

$$\mathsf{tr}(P) = \frac{1}{2}C_{n-1} - \frac{1}{2}\sum_{q \in Q} C_{\ell(q)}C_{\mathrm{r}(q)}\,. \tag{5}$$

With the monotonicity of Catalan products as in Fact A.4 and Lemma 3.5, this immediately implies the result in [22]: The symmetric configuration maximizes and the barely-in configuration minimizes $\mathsf{tr}(P)$, also for n odd. (For n odd, we choose a regular $(n-1)$-gon together with a point close to its center. In this way, $|\ell(q) - \mathrm{r}(q)| = 1$, i.e. $\{\ell(q), \mathrm{r}(q)\} = \{\lfloor \frac{n}{2} - 1 \rfloor, \lceil \frac{n}{2} - 1 \rceil\}$, for all extreme points q.)

Table 1 $\mathsf{pm}_{\text{symm}}(m)$ is $\frac{3}{2}C_m$ for m even and $\frac{3}{2}C_m - \frac{2m-1}{2}C^2_{(m-1)/2}$ for m odd. Values for the number of CFPMs, symmetric and barely-in configuration, and extremal wheel sets; ↑ and ↓ indicate maximizers and minimizers, respectively. Note that for odd m, barely-in (with $\frac{5}{4}C_m(1+o(1))$) majorizes symmetric (with $\frac{3}{2}C_m(1+o(1))$), a small-number effect that is resolved first for 46 points

$\|P\| = n = 2m$			Symmetric		Barely-in	Wheel	
n	m	C_m	$\mathsf{pm}_{\text{symm}}(m)$		$C_m + C_{m-1}$	$\overset{\min}{\mathsf{pm}}(P)$	$\overset{\max}{\mathsf{pm}}(P)$
4	2	2	3 ↕	=	3 ↕	3	3
6	3	5	5 ↓	<	7 ↑	5	7
8	4	14	21 ↑	>	19	15	21
10	5	42	45 ↓	<	56	45	61
12	6	132	198 ↑	>	174	146	198
14	7	429	481	<	561	477	631
16	8	1430	2145 ↑	>	1859	1595	2145
18	9	4862	5627	<	6292	5434	7195
20	10	16, 796	25, 194 ↑	>	21, 658	18, 798	25, 194
22 ⋮	11 ⋮	58, 786 ⋮	69, 657 ⋮	<	75, 582 ⋮	65, 858 ⋮	87, 297 ⋮
44	22	91×10^9	137 223 845 460 ↑	>	115 948 830 660	103×10^9	137×10^9
46	23	343×10^9	436 834 060 065	$\overset{!}{>}$	434 542 177 290	386×10^9	513×10^9

Similarly, there is a formula for origin embracing triangles (for the definition of $\ell(q)$ and $\mathrm{r}(q)$ as above set $z := \mathbf{0}$).

$$|\mathcal{T}_{\text{embr}}| = \binom{n}{3} - \frac{1}{2}\sum_{q\in Q}\left(\binom{\ell(q)}{2} + \binom{\mathrm{r}(q)}{2}\right) \quad \text{(here } n := |Q|\text{).} \tag{6}$$

Again, this allows to read off extremal properties relatively easily.

Formulas as in (5) and (6) and many more can be shown by continuous motion arguments as used e.g. in [6, 22, 26]. This is a topic for ongoing investigations [21].

Acknowledgements We thank Vera Rosta, Patrick Schnider, Shakhar Smorodinsky, Antonis Thomas, and Manuel Wettstein for discussions on and suggestions for the material covered in this paper. We also thank the three referees for carefully reading through the paper with numerous suggestions for improvements. (And thanks to Maple, for sparing us some tedious calculations.)

Appendix: Catalan Facts

Fact A.1 (Definition & Asymptotics) *For all $m \in \mathbb{N}_0$,*

$$C_m := \frac{1}{m+1}\binom{2m}{m} = \frac{(2m)!}{(m+1)!m!} = \Theta\left(\frac{1}{m^{3/2}}4^m\right) .$$

Fact A.2 (Simple Recurrence) *For all $m \in \mathbb{N}_0$,*

$$2 \cdot \frac{2m-1}{m+1} \cdot C_{m-1} = C_m = \frac{1}{2} \cdot \frac{m+2}{2m+1} \cdot C_{m+1} .$$

Fact A.3 (Segner Recurrence) *For all $m \in \mathbb{N}$,*

$$C_m = C_0C_{m-1} + C_1C_{m-2} + \cdots C_{m-1}C_0 = \sum_{i=0}^{m-1} C_iC_{m-1-i}.$$

Fact A.4 *For $k, \ell \in \mathbb{N}$,*

$$C_kC_\ell < C_{k-1}C_{\ell+1} \quad \textit{iff} \quad k \le \ell$$

Proof

$$C_kC_\ell = 2\frac{2k-1}{k+1}C_{k-1} \cdot \frac{1}{2}\frac{\ell+2}{2\ell+1}C_{\ell+1} = \frac{(2k-1)(\ell+2)}{(k+1)(2\ell+1)}C_{k-1}C_{\ell+1}$$

and

$$\begin{aligned} &\frac{(2k-1)(\ell+2)}{(k+1)(2\ell+1)} < 1 \\ &\quad\Leftrightarrow \quad 3k-3 < 3\ell \\ &\qquad\Leftrightarrow \quad k \le \ell \ \text{ since } k \text{ and } \ell \text{ are integers.} \end{aligned}$$

□

Fact A.5 *For $m \in \mathbb{N}$,*

$$(2m-1)\left(2C_{m-1} - \frac{1}{2}C_m\right) = \frac{3}{2}C_m$$

Proof Employ Fact A.2 for expressing C_{m-1} in terms of C_m.

$$\begin{aligned}(2m-1)\left(2C_{m-1}-\frac{1}{2}C_m\right) &= (2m-1)\left(2\left(\frac{1}{2}\frac{m+1}{2m-1}C_m\right)-\frac{1}{2}C_m\right)\\ &= (2m-1)\left(\frac{2(m+1)-(2m-1)}{2(2m-1)}\right)C_m\\ &= \frac{3}{2}C_m\end{aligned}$$

□

Fact A.6 *For $k, \ell \in \mathbb{N}$,*

$$\frac{2k-1}{2k}\frac{C_{k-1}}{C_k} < \frac{C_{\ell-1}}{C_\ell} \quad \textit{iff} \quad 2\ell-1 < 3k$$

Proof

$$\begin{aligned}&\frac{2k-1}{2k}\frac{C_{k-1}}{C_k} < \frac{C_{\ell-1}}{C_\ell}\\ \Leftrightarrow \quad &\frac{2k-1}{2k}\frac{k+1}{2(2k-1)} < \frac{\ell+1}{2(2\ell-1)} \quad \text{due to Fact A.2}\\ \Leftrightarrow \quad &2\ell-1 < 3k\end{aligned}$$

□

Fact A.7 *For all $m \in \mathbb{N}$, $m \geq 2$,*

$$C_m + C_{m-1} - 2C_{m-2} = \frac{9}{8}C_m\left(1+\Theta(1/m^2)\right) .$$

Proof

$$\begin{aligned}C_m + C_{m-1} - 2C_{m-2} &= C_m + \frac{1}{2}\frac{m+1}{2m-1}C_m - 2(\frac{1}{2}\frac{m+1}{2m-1}\cdot\frac{1}{2}\frac{m}{2m-3})C_m\\ &= C_m\left(\frac{9m^2-18m+3}{8m^2-16m+6}\right)\\ &= \frac{9}{8}C_m\left(1-\frac{5}{3(4m^2-8m+3)}\right)\\ &= \frac{9}{8}C_m\left(1+\Theta(1/m^2)\right)\end{aligned}$$

□

References

1. A.K. Abu-Affash, A. Biniaz, P. Carmi, A. Maheshwari, M.H.M. Smid, Approximating the bottleneck plane perfect matching of a point set. Comput. Geom. **48**(9), 718–731 (2015)
2. O. Aichholzer, F. Aurenhammer, H. Krasser, Enumerating order types for small point sets with applications. Order **19**(3), 265–281 (2002)
3. O. Aichholzer, S. Bereg, A. Dumitrescu, A. García Olaverri, C. Huemer, F. Hurtado, M. Kano, A. Márquez, D. Rappaport, S. Smorodinsky, D.L. Souvaine, J. Urrutia, D.R. Wood, Compatible geometric matchings. Comput. Geom. **42**(6–7), 617–626 (2009)
4. G. Aloupis, L. Barba, S. Langerman, D.L. Souvaine, Bichromatic compatible matchings. Comput. Geom. **48**(8), 622–633 (2015)
5. G. Aloupis, J. Cardinal, S. Collette, E.D. Demaine, M.L. Demaine, M. Dulieu, R.F. Monroy, V. Hart, F. Hurtado, S. Langerman, M. Saumell, C. Seara, P. Taslakian, Non-crossing matchings of points with geometric objects. Comput. Geom. **46**(1), 78–92 (2013)
6. A. Andrzejak, B. Aronov, S. Har-Peled, R. Seidel, E. Welzl, Results on k-sets and j-facets via continuous motion, in *Proceedings of the 14th Annual Symposium on Computational Geometry (SoCG'98)*, ed. by R. Janardan (ACM, 1998), pp. 192–199
7. A. Asinowski, The number of non-crossing perfect plane matchings is minimized (almost) only by point sets in convex position. CoRR abs/1502.05332 (2015)
8. A. Asinowski, G. Rote, Point sets with many non-crossing matchings. CoRR abs/1502.04925 (2015)
9. A. Biniaz, P. Bose, A. Maheshwari, M.H.M. Smid, Packing plane perfect matchings into a point set. CoRR abs/1501.03686 (2015)
10. A. Elmasry, K.M. Elbassioni, Output-sensitive algorithms for enumerating and counting simplices containing a given point in the plane, in *Proceedings of the 17th Canadian Conference on Computational Geometry (CCCG'05)*, (2005), pp. 248–251
11. P. Erdős, L. Lovász, A. Simmons, E.G. Straus, Dissection graphs of planar point sets, in *A Survey of Combinatorial Theory (Proceedings of the International Symposium, Colorado State University, Fort Collins, 1971)*, (1973), pp. 139–149
12. A. García Olaverri, M. Noy, J. Tejel, Lower bounds on the number of crossing-free subgraphs of K_N. Comput. Geom. **16**(4), 211–221 (2000)
13. J.E. Goodman, R. Pollack, Multidimensional sorting. SIAM J. Comput. **12**(3), 484–507 (1983)
14. A. Holmsen, J. Pach, H. Tverberg, Points surrounding the origin. Combinatorica **28**(6), 633–644 (2008)
15. M.E. Houle, F. Hurtado, M. Noy, E. Rivera-Campo, Graphs of triangulations and perfect matchings. Graphs and Combinatorics **21**(3), 325–331 (2005)
16. S. Khuller, J.S.B. Mitchell, On a triangle counting problem. Inf. Process. Lett. **33**(6), 319–321 (1990)
17. D. Marx, T. Miltzow, Peeling and nibbling the cactus: Subexponential-time algorithms for counting triangulations and related problems. CoRR abs/1603.07340 (2016)
18. T.S. Motzkin, Relations between hypersurface cross ratios, and a combinatorial formula for partitions of a polygon, for permanent preponderance, and for non-associative products. Bull. Am. Math. Soc. **54**(4), 352–360 (1948)
19. J. Pach, M. Szegedy, The number of simplices embracing the origin, in *Discrete Geometry: In Honor of W. Kuperberg's 60th Birthday*. Pure and Applied Mathematics, ed. by A. Bezdek (Marcel Dekker Inc., New York, 2003), pp. 381–386
20. A. Pilz, E. Welzl, Order on order types, in *Proceedings of the 31st Annual Symposium on Computational Geometry (SoCG'15)*, ed. by L. Arge, J. Pach. Volume 34 of LIPIcs (Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, 2015), pp. 285–299
21. A. Pilz, E. Welzl, M. Wettstein, From crossing-free graphs on wheel sets to embracing simplices and polytopes with few vertices, in *Proceedings of 33rd International Symposium on Computational Geometry (SoCG)* (2017, to appear)

22. D. Randall, G. Rote, F. Santos, J. Snoeyink, Counting triangulations and pseudo-triangulations of wheels, in *Proceedings of the 13th Canadian Conference on Computational Geometry (CCCG'01)* (2001), pp. 149–152
23. M. Sharir, A. Sheffer, E. Welzl, Counting plane graphs: perfect matchings, spanning cycles, and Kasteleyn's technique. J. Comb. Theory Ser. A **120**(4), 777–794 (2013)
24. M. Sharir, E. Welzl, On the number of crossing-free matchings, cycles, and partitions. SIAM J. Comput. **36**(3), 695–720 (2006)
25. R.P. Stanley, *Catalan Numbers* (Cambridge University Press, New York, 2015)
26. H.A. Tverberg, A generalization of Radon's theorem. J. Lond. Math. Soc. **41**(1), 123–128 (1966)
27. J. van Leeuwen, A.A. Schoone, Untangling a travelling salesman tour in the plane, in *Proceedings of the Graph Theoretic Concepts in Computer Science (WG'81)*, ed. by J.R. Mühlbacher (Hanser Verlag, München, 1981), pp. 87–98. Also Technical report RUU-CS-80-11, Departments of Computer Science, Utrecht University, 1980
28. U. Wagner, E. Welzl, A continuous analogue of the upper bound theorem. Discret. Comput. Geom. **26**(2), 205–219 (2001)
29. E. Welzl, Entering and leaving *j*-facets. Discret. Comput. Geom. **25**(3), 351–364 (2001)
30. M. Wettstein, Counting and enumerating crossing-free geometric graphs, in *Proceedings of the 30th Annual Symposium on Computational Geometry (SOCG'14)*, ed. by S.-W. Cheng, O. Devillers (ACM, 2014), pp. 1–10
31. G.M. Ziegler, *Lectures on Polytopes* (Springer, Berlin, 1995)

Network Essence: PageRank Completion and Centrality-Conforming Markov Chains

Shang-Hua Teng

Abstract Jiří Matoušek (1963–2015) had many breakthrough contributions in mathematics and algorithm design. His milestone results are not only profound but also elegant. By going beyond the original objects—such as Euclidean spaces or linear programs—Jirka found the *essence* of the challenging mathematical/algorithmic problems as well as beautiful solutions that were natural to him, but were surprising discoveries to the field.

In this short exploration article, I will first share with readers my initial encounter with Jirka and discuss one of his fundamental geometric results from the early 1990s. In the age of social and information networks, I will then turn the discussion from geometric structures to network structures, attempting to take a humble step towards the holy grail of network science, that is to understand the *network essence* that underlies the observed sparse-and-multifaceted network data. I will discuss a simple result which summarizes some basic algebraic properties of *personalized PageRank matrices*. Unlike the traditional transitive closure of binary relations, the personalized PageRank matrices take "accumulated Markovian closure" of network data. Some of these algebraic properties are known in various contexts. But I hope featuring them together in a broader context will help to illustrate the desirable properties of this Markovian completion of networks, and motivate systematic developments of a network theory for understanding vast and ubiquitous multifaceted network data.

Supported in part by a Simons Investigator Award from the Simons Foundation and by NSF grant CCF-1111270. I would like to thank my Ph.D. student Alana Shine and the three anonymous referees for valuable feedback.

S.-H. Teng (✉)
Computer Science and Mathematics, USC, Los Angeles, CA, USA
e-mail: shanghua.teng@gmail.com

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_31

1 Geometric Essence: To the Memory of Jiří Matoušek

Like many in theoretical computer science and discrete mathematics, my own research has benefited from Jirka's deep insights, especially into computational geometry [64] and linear programming [65]. In fact, our paths accidentally crossed in the final year of my Ph.D. program. As a part of my 1991 CMU thesis [88], I obtained a result on the deterministic computation of a geometric concept, called centerpoints, which led me to learn about one of Jirka's groundbreaking results during this time.

1.1 Centerpoints

The median is a widely-used concept for analyzing one-dimensional data, due to its statistical robustness and its natural algorithmic applications to divide-and-conquer. In general, suppose $P = \{p_1, \ldots, p_n\}$ is a set of n real numbers. For $\delta \in (0, 1/2]$, we call $c \in \mathbb{R}$ a *δ-median* of P if $\max\left(|\{i : p_i < c\}|, |\{j : p_j > c\}|\right) \leq (1-\delta)\, n$. A $\frac{1}{2}$-median of P is known simply as a *median*. Centerpoints are high-dimensional generalization of medians:

Definition 1.1 (Centerpoints) Suppose $P = \{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$ is a point set in $\mathbb{R}^d$. For $\delta \in (0, 1/2]$, a point $\mathbf{c} \in \mathbb{R}^d$ is a *δ-centerpoint* of P if for all unit vectors $\mathbf{z} \in \mathbb{R}^d$, the projection $\mathbf{z}^T\mathbf{c}$ is a *δ-median* of the projections, $\mathbf{z}^T \cdot P = \{\mathbf{z}^T\mathbf{p}_1, \ldots, \mathbf{z}^T\mathbf{p}_n\}$.

Geometrically, every hyperplane $\mathbf{h}$ in $\mathbb{R}^d$ divides the space into two open halfspaces, $\mathbf{h}^+$ and $\mathbf{h}^-$. Let the *splitting ratio* of $\mathbf{h}$ over P, denoted by $\delta_{\mathbf{h}}(P)$, be:

$$\delta_{\mathbf{h}}(P) := \frac{\max\left(|\mathbf{h}^+ \cap P|, |\mathbf{h}^- \cap P|\right)}{|P|} \tag{1}$$

Definition 1.1 can be restated as: $\mathbf{c} \in \mathbb{R}^d$ is a δ-centerpoint of P if the splitting ratio of every hyperplane $\mathbf{h}$ passing through $\mathbf{c}$ is at most $(1-\delta)$. Centerpoints are fundamental to geometric divide-and-conquer [34]. They are also strongly connected to the concept of regression depth introduced by Rousseeuw and Hubert in robust statistics [7, 50].

We all know that every set of real numbers has a median. Likewise—and remarkably—every point set in d-dimensional Euclidean space has a $\frac{1}{d+1}$-centerpoint [30]. This mathematical result can be established by Helly's classical theorem from convex geometry.[1] Algorithmically, Vapnik–Chervonenkis' celebrated sampling theorem [92] (more below) implies an efficient randomized algorithm—at least in theory—for computing a $(\frac{1}{d+1}-\epsilon)$-centerpoint. This "simple"

[1] Helly's Theorem states: Suppose $\mathcal{K}$ is a family of at least $d+1$ convex sets in $\mathbb{R}^d$, and $\mathcal{K}$ is finite or each member of $\mathcal{K}$ is compact. Then, if each $d+1$ members of $\mathcal{K}$ have a common point, there must be a point common to all members of $\mathcal{K}$.

algorithm first takes a "small" random sample, and then obtains its $\frac{1}{d+1}$-centerpoint via linear programming. The complexity of this LP-based sampling algorithm is:

$$2^{O(d)}\left(\frac{d}{\epsilon^2}\cdot\log\frac{d}{\epsilon}\right)^d.$$

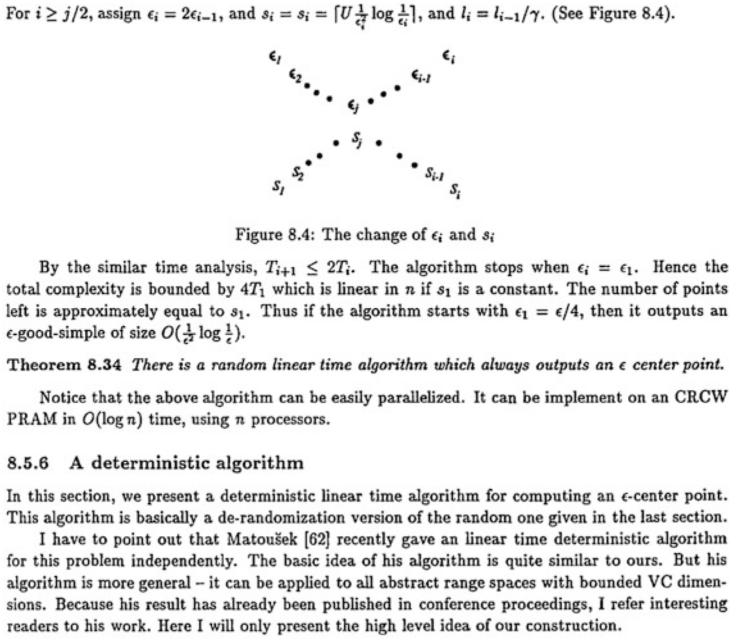
1.2 Derandomization

For my thesis, I needed to compute centerpoints in order to construct geometric separators [67] for supporting finite-element simulation and parallel scientific computing [68]. Because linear programming was too slow, I needed a practical centerpoint algorithm to run large-scale experiments [45]. Because I was a theory student, I was also aiming for a theoretical algorithm to enrich my thesis. For the latter, I focused on derandomization, which was then an active research area in theoretical computer science. For centerpoint approximation without linear programming, my advisor Gary Miller and I quickly obtained a simple and practical algorithm[2] based on Radon's classical theorem[3] [30]. But for derandomization, it took me more than a year to finally design a deterministic linear-time algorithm for computing $(\frac{1}{d+1}-\epsilon)$-centerpoints in any fixed dimensions. It happened in the Spring of 1991, my last semester at CMU. Gary then invited me to accompany him for a month-long visit, starting at the spring break of 1991, at the International Computer Science Institute (ICSI), located near the U.C. Berkeley campus. During the California visit, I ran into Leo Guibas, one of the pioneers of computational geometry.

After I told Leo about my progress on Radon-Tverberg decomposition [90] and centerpoint computation, he mentioned to me a paper by Jirka [64], which was just accepted to the *ACM Symposium on Theory of Computing* (STOC 1991)—before my solution—that beautifully solved the sampling problem for a broad class of computational geometry and statistical learning problems. Jirka's result—see Theorem 1.3 below—includes the approximation of centerpoints as a simple special case. Although our approaches had some ideas in common, I instantly knew that this mathematician—who I later learned was just a year older than me—was masterful and brilliant. I shortened that section of my thesis by referring readers to Jirka's paper [64], and only included the scheme I had that was in common with his bigger result (Fig. 1).

[2] The construction started as a heuristics, but it took a few more brilliant collaborators and years (after my graduation) to rigorously analyzing its performance [29].

[3] Radon's Theorem states: Every point set $Q \subset \mathbb{R}^d$ with $|Q| \geq d+2$ can be partitioned into two subsets (Q_1, Q_2) such that the convex hulls of Q_1 and Q_2 have a common point.



For $i \geq j/2$, assign $\epsilon_i = 2\epsilon_{i-1}$, and $s_i = s_i = \lceil U\frac{1}{\epsilon_i^2}\log\frac{1}{\epsilon_i}\rceil$, and $l_i = l_{i-1}/\gamma$. (See Figure 8.4).

ϵ_1
ϵ_i
ϵ_2
ϵ_{i-1}
ϵ_j
S_j
S_2
S_{i-1}
S_1
S_i

Figure 8.4: The change of ϵ_i and s_i

By the similar time analysis, $T_{i+1} \leq 2T_i$. The algorithm stops when $\epsilon_i = \epsilon_1$. Hence the total complexity is bounded by $4T_1$ which is linear in n if s_1 is a constant. The number of points left is approximately equal to s_1. Thus if the algorithm starts with $\epsilon_1 = \epsilon/4$, then it outputs an ϵ-good-simple of size $O(\frac{1}{\epsilon^2}\log\frac{1}{\epsilon})$.

Theorem 8.34 *There is a random linear time algorithm which always outputs an ϵ center point.*

Notice that the above algorithm can be easily parallelized. It can be implement on an CRCW PRAM in $O(\log n)$ time, using n processors.

8.5.6 A deterministic algorithm

In this section, we present a deterministic linear time algorithm for computing an ϵ-center point. This algorithm is basically a de-randomization version of the random one given in the last section.

I have to point out that Matoušek [62] recently gave an linear time deterministic algorithm for this problem independently. The basic idea of his algorithm is quite similar to ours. But his algorithm is more general – it can be applied to all abstract range spaces with bounded VC dimensions. Because his result has already been published in conference proceedings, I refer interesting readers to his work. Here I will only present the high level idea of our construction.

Fig. 1 Page 66 (Chapter 8) of my thesis

1.3 Matoušek's Theorem: The Essence of Dimensionality

Mathematically, a *range space* Σ is a pair $(X, \mathcal{R})$, where X is a finite or infinite set, and $\mathcal{R}$ is a finite or infinite family of subsets of X. Each $H \in \mathcal{R}$ can be viewed as a classifier of X, with elements in $X \cap H$ as its positive instances. For example, $\mathbb{R}^d$ and its halfspaces form a range space, so do $\mathbb{R}^d$ and its L_p-balls, for any $p > 0$, as well as V and the set of all cliques in a graph $G = (V, E)$. Range spaces greatly extend the concept of *linear separators*.

An important technique in statistical machine learning and computational geometry is sampling. For range spaces, we can measure the quality of a sample as the following:

Definition 1.2 **(ϵ-samples)** Let $\Sigma = (X, \mathcal{R})$ be an n-point range space. A subset $S \subseteq X$ is an *ϵ-sample* or *ϵ-approximation* for Σ if for all $H \in \mathcal{R}$:

$$\left| \frac{|H \cap S|}{|S|} - \frac{|H \cap X|}{|X|} \right| \leq \epsilon \tag{2}$$

For each $S \subseteq X$, the set of distinct classifiers that $\mathcal{R}$ can define is $\mathcal{R}(S) = \{H \cap S : H \in \mathcal{R}\}$. For any $m \leq |X|$, let the *shatter function* for Σ be:

$$\pi_{\mathcal{R}}(m) = \sup_{S \subseteq X, |S|=m} |\mathcal{R}(S)| \tag{3}$$

Theorem 1.3 (Deterministic Sampling—Matoušek) *Let $\Sigma = (X, \mathcal{R})$ be an n-point range space with the shatter function satisfying $\pi_{\mathcal{R}}(m) = O(m^d)$ ($d \geq 1$ a constant). Having a subspace oracle for Σ, and given a parameter r, we can deterministically compute a $(1/r)$-approximation of size $O(dr^2 \log r)$ for Σ, in time $O(n(r^2 \log r)^d)$.*

Matoušek's sampling theorem goes beyond traditional geometry and completely derandomizes the theory of Vapnik–Chervonenkis [92].

Theorem 1.4 (Vapnik and Chervonenkis) *There exists a constant c such that for any finite range space $\Sigma = (X, \mathcal{R})$ and $\epsilon, \delta \in (0, 1)$, if S is a set of $c \cdot \frac{d}{\epsilon^2} \left(\log \frac{d}{\epsilon\delta}\right)$ uniform and independent samples from X, where $d = \mathrm{VC}(\Sigma)$, (see below for definition) then:*

$$\Pr[S \textit{ is an } \epsilon\textit{-sample for } \Sigma] \geq 1 - \delta$$

Matoušek's deterministic algorithm can be applied to geometric classifiers as well as any classifier—known as a *concept space*—that arises in statistical learning theory [91]. The concept of range space has also provided a powerful tool for capturing geometric structures, and played a profound role—both in theory and in practice—for data clustering [38] and geometric approximation [3]. The beauty of Vapnik–Chervonenkis' theory and Matoušek's sampling theorem lies in the *essence of dimensionality*, which is generalized from geometric spaces to abstract range spaces. In Euclidean geometry, the dimensionality comes naturally to many of us. For abstract range spaces, the growth of the *shatter functions* is more intrinsic! If $\pi_{\mathcal{R}}(m) = 2^m$, then there exists a set $S \subseteq X$ of m elements that is *shattered*, i.e., for any subset T of $S \subseteq X$, there exists $H \in \mathcal{R}$ such that $T = H \cap S$. In other words, we can use $\mathcal{R}$ to build classifiers for all subsets of S. There is a beautiful *dichotomy* of polynomial and exponential complexity within the concept of shattering:

- either X has a subset $S \subseteq X$ of size m that can be shattered by $\mathcal{R}$,
- or for any $U \subseteq X$, $|U| \geq m$, $|\{H \cap U : H \in \mathcal{R}\}|$ is polynomial in $|U|$.

The latter case implies that $\mathcal{R}$ can only be used to build a polynomial number of classifiers for U. The celebrated *VC-dimension* of range space $\Sigma = (X, \mathcal{R})$, denoted by $\mathrm{VC}(\Sigma)$, is defined as:

$$\mathrm{VC}(\Sigma) := \arg\max\{m : \pi_{\mathcal{R}}(m) = 2^m\}.$$

This polynomial-exponential dichotomy is established by the following Sauer's lemma.[4]

[4] This lemma is also known as Perles–Sauer–Shelah's lemma.

Lemma 1.5 (Sauer) *For any range space* $\Sigma = (X, \mathcal{R})$ *and* $\forall m > \mathrm{VC}(\Sigma)$, $\pi_{\mathcal{R}}(m) \leq \sum_{k=0}^{\mathrm{VC}(\Sigma)} \binom{m}{k}$.

Sauer's lemma extends the following well-known fact of Euclidean geometry: any set of m hyperplanes in $\mathbb{R}^d$ divides the space into at most $O(m^d)$ convex cells. By the point-hyperplane duality, any set of m points can be divided into at $O(m^d)$ subsets by halfspaces.

Although my construction of ϵ-samples in $\mathbb{R}^d$ was good enough for designing linear-time centerpoint approximation algorithm in fixed dimensions, it did not immediately generalize to arbitrary range spaces, because it was tailored to the geometric properties of Euclidean spaces.

By addressing abstract range spaces, Jirka resolved the intrinsic algorithmic problem at the heart of Vapnik–Chervonenkis' sampling theory. Like Theorem 1.3, many of Jirka's other landmark and breakthrough results are elegant, insightful, and fundamental. By going beyond the original objects—such as Euclidean spaces or linear programs [65]—Jirka usually went directly to the *essence* of the challenging problems to come up with beautiful solutions that were natural to him but remarkable to the field.

2 Backgrounds: Understanding Multifaceted Network Data

To analyze the structures of social and information networks in the age of Big Data, we need to overcome various conceptual and algorithmic challenges both in understanding network data and in formulating solution concepts. For both, we need to capture the *network essence*.

2.1 The Graph Model—A Basic Network Facet

At the most basic level, a network can be modeled as a graph $G = (V, E)$, which characterizes the structure of the network in terms of:

- **nodes**: for example, Webpages, Internet routers, scholarly articles, people, random variables, or counties
- **edges**: for example, links, connections, citations, friends, conditional dependencies, or voting similarities

In general, nodes in many real-world networks may not be "homogeneous" [5], as they may have some additional features, specifying the *types* or *states* of the node elements. Similarly, edges may have additional features, specifying the *levels* and/or *types* of pairwise interactions, associations, or affinities.

Networks with "homogeneous" types of nodes and edges are closest to the combinatorial structures studied under traditional graph theory, which considers both weighted or unweighted graphs. Three basic classes of weighted graphs often appear in applications. The first class consists of *distance networks*, where each edge $e \in E$ is assigned a number $l_e \geq 0$, representing the *length* of edge e. The second class consists of *affinity networks*, where each edge $(u, v) \in E$ is assigned a weight $w_{u,v} \geq 0$, specifying u's *affinity weight* towards v. The third class consists of *probabilistic networks*, where each (directed) edge $(u, v) \in E$ is assigned a probability $p_{u,v} \geq 0$, modeling how a random process connects u to v. It is usually more natural to view maps or the Internet as distance networks, social networks as affinity networks, and Markov processes as probabilistic networks. Depending on applications, a graph may be directed or undirected. Examples of directed networks include: the Web, Twitter, the citation graphs of scholarly publications, and Markov processes. Meanwhile, Facebook "friends" or collaboration networks are examples of undirected graphs.

In this article, we will first focus on affinity networks. An affinity network with n nodes can be mathematically represented as a weighted graph $G = (V, E, \mathbf{W})$. Unless otherwise stated, we assume $V = [n]$ and $\mathbf{W}$ is an $n \times n$ non-negative matrix (for example from $[0, 1]^{n \times n}$). We will follow the convention that for $i \neq j$, $w_{i,j} = 0$, if and only if, $(i, j) \notin E$. If $\mathbf{W}$ is a symmetric matrix, then we say G is *undirected*. If $w_{i,j} \in \{0, 1\}$, $\forall i, j \in V$, then we say G is *unweighted*.

Although they do not always fit, three popular data models for defining pairwise affinity weights are the metric model, feature model, and statistical model. The first assumes that an underlying metric space, $\mathcal{M} = (V, \mathrm{dist})$, impacts the interactions among nodes in a network. The affinities between nodes may then be determined by their *distances* from the underlying metric space: The closer two elements are, the higher their affinity becomes, and the more interactions they have. A standard way to define affinity weights for $u \neq v$ is: $w_{u,v} = \mathrm{dist}(u, v)^{-\alpha}$, for some $\alpha > 0$. The second assumes that there exists an underlying "feature" space, $\mathcal{F} = (V, \mathbf{F})$, that impacts the interactions among nodes in a network. This is a widely-used alternative data model for information networks. In a d-dimensional feature space, $\mathbf{F}$ is an $n \times d$ matrix, where $f_{u,i} \in \mathbb{R}^+ \cup \{0\}$ denotes u's *quality score* with respect the ith feature. Let $\mathbf{f}_u$ denote the uth row of $\mathbf{F}$, i.e., the *feature vector* of node u. The affinity weights $w_{u,v}$ between two nodes u and v may then be determined by the *correlation* between their features: $w_{u,v} \sim \left(\mathbf{f}_u^T \cdot \mathbf{f}_v\right) = \sum_{i=1}^{d} f_{u,i} \cdot f_{v,i}$. The third assumes that there exists an underlying statistical space (such as a stochastic block model, Markov process, or (Gaussian) random field) that impacts the pairwise interactions. The higher the dependency between two elements is, the higher their strength of tie is.

If one thinks that the meaning of weighted networks is complex, the real-world network data is far more complex and diverse. We will have more discussions in Sects. 2.3 and 4.

2.2 Sparsity and Underlying Models

A basic challenge in network analysis is that real network data that we observe is only a reflection of underlying network models. Thus, like machine learning tasks which have to work with samples from an unknown underlying distribution, network analysis tasks typically work with observed network data, which is usually different from the underlying network model. As argued in [11, 48, 56, 89], a real-world social and information network may be viewed as an observed network, induced by a "complete-information" underlying preference/affinity/statistical/geometric/feature/economical model. However, these observed networks are typically sparse with many missing links.

> For studying network phenomena, it is crucial to mathematically understand underlying network models, while algorithmically work efficiently with sparse observed data. Thus, developing systematic approaches to uncover or capture the underlying network model — or the network essence — is a central and challenging mathematical task in network analysis.

Implicitly or explicitly, underlying network models are the ultimate guide for understanding network phenomena, and for inferring missing network data, and distinguishing missing links from absent links. To study basic network concepts, we also need to simultaneously understand the observed and underlying networks. Some network concepts, such as *centrality*, capture various aspects of "dimension reduction" of network data. Others characterizations, such as *clusterability* and *community classification*, are more naturally expressed in a space with dimension higher than that of the observed networks.

Schematically, centrality assigns a numerical score or ranking to each node, which measures the *importance* or *significance* of each node in a network [1, 13–17, 22, 33, 36, 37, 41, 42, 51, 66, 71, 76, 80]. Mathematically, a numerical centrality measure is a mapping from a network $G = (V, E, \mathbf{W})$ to a $|V|$-dimensional real vector:

$$[\,\text{centrality}_{\mathbf{W}}(v)\,]_{v\in V} \in \mathbb{R}^{|V|} \tag{4}$$

For example, a widely used centrality measure is the PageRank centrality. Suppose $G = (V, E, \mathbf{W})$ is a weighted directed graph. The *PageRank centrality* uses an additional parameter $\alpha \in (0, 1)$—known as the *restart constant*—to define a finite Markov process whose transition rule—for any node $v \in V$—is the following:

- with probability α, restart at a random node in V, and
- with probability $(1 - \alpha)$, move to a neighbor of v, chosen randomly with probability proportional to edge weights out of v.

Then, the *PageRank centrality* (with restart constant α) of any $v \in V$ is proportional to v's stationary probability in this Markov chain.

In contrast, clusterability assigns a numerical score or ranking to each subset of nodes, which measures the *coherence* of each group in a network [62, 71, 89].

Mathematically, a numerical clusterability measure is a mapping from a network $G = (V, E, \mathbf{W})$ to a $2^{|V|}$-dimensional real vector:

$$[\,\text{clusterability}_{\mathbf{W}}(S)\,]_{S\subseteq V} \in [0,1]^{2^{|V|}} \tag{5}$$

An example of clusterability measure is *conductance* [62].[5] Similarly, a community-characterization rule [19] is a mapping from a network $G = (V, E, \mathbf{W})$ to a $2^{|V|}$-dimensional Boolean vector:

$$[\,\mathcal{C}_{\mathbf{W}}(S)\,]_{S\subseteq V} \in \{0,1\}^{2^{|V|}} \tag{6}$$

indicating whether or not each group $S \subseteq V$ is a community in G. Clusterability and community-identification rules have much higher dimensionality than centrality. To a certain degree, they can be viewed as a "complete-information" model of the observed network. Thus again:

> Explicitly or implicitly, the formulations of these network concepts are mathematical processes of uncovering or capturing underlying network models.

2.3 *Multifaceted Network Data: Beyond Graph-Based Network Models*

Another basic challenge in network analysis is that real-world network data is much richer than the graph-theoretical representations. For example, social networks are more than weighted graphs. Likewise, the Web and Twitter are not just directed graphs. In general, network interactions and phenomena—such as social influence [55] or electoral behavior [35]—are more complex than what can be captured by nodes and edges. The network interactions are often the result of the interplay between dynamic mathematical processes and static underlying graph structures [25, 44].

2.3.1 Diverse Network Models

The richness of network data and diversity of network concepts encourage us to consider network models beyond graphs [89]. For example, each clusterability measure $[\text{clusterability}_{\mathbf{W}}(S)]_{S\subseteq V}$ of a weighted graph $G = (V, E, \mathbf{W})$ explicitly defines a complete-information, weighted hyper-network:

[5] The *conductance* of a group $S \subset V$ is the ratio of its *external connection* to its *total connection* in G.

Definition 2.1 (Cooperative Model: Weighted Hypergraphs) A weighted hypergraph over V is given by $H = (V, E, \boldsymbol{\tau})$ where $E \subseteq 2^V$ is a set of hyper-edges and $\boldsymbol{\tau} : E \to \mathbb{R}$ is a function that assigns weights to hyper-edges. H is a complete-information cooperative networks if $E = 2^V$.

We refer to weighted hypergraphs as *cooperative networks* because they are the central subjects in classical *cooperative game theory*, but under a different name [81]. An n-person *cooperative game* over $V = [n]$ is specified by a *characteristic function* $\boldsymbol{\tau} : 2^V \to \mathbb{R}$, where for any coalition $S \subseteq V$, $\boldsymbol{\tau}(S)$ denotes the *cooperative utility* of S.

Cooperative networks are generalization of undirected weighted graphs. One can also generalize directed networks, which specify directed node-node interactions. The first one below explicitly captures node-group interactions, while the second one captures group-group interactions.

Definition 2.2 (Incentive Model) An *incentive network* over V is a pair $U = (V, \boldsymbol{u})$. For each $s \in V$, $u_s : 2^{V \setminus \{s\}} \to \mathbb{R}$ specifies s's *incentive utility* over subsets of $V \setminus \{s\}$. In other words, there are $|S|$ utility values, $\{u_s(S \setminus \{s\}\}_{s \in S}$, associated with each group $S \subseteq V$ in the incentive network. For each $s \in S$, the *value* of its interaction with the rest of the group $S \setminus \{s\}$ is explicitly defined as $u_s(S \setminus \{s\})$.

Definition 2.3 (Powerset Model) A *powerset network* over V is a weighted directed network on the powersets of V. In other words, a powerset network $P = (V, \boldsymbol{\theta})$ is specified by a function $\boldsymbol{\theta} : 2^V \times 2^V \to \mathbb{R}$.

For example—as pointed in [25, 55]—a social-influence instance fundamentally defines a powerset network. Recall that a social-influence instance $\mathcal{I}$ is specified by a directed graph $G = (V, E)$ and an influence model $\mathcal{D}$ [32, 55, 78], where G defines the graph structure of the social network and $\mathcal{D}$ defines a stochastic process that characterizes how nodes in each *seed set* $S \subseteq V$ *collectively influence* other nodes using the edge structures of G [55]. A popular influence model is *independent cascade* (IC)[6] [55].

Mathematically, the influence process $\mathcal{D}$ and the network structure G together define a probability distribution $\boldsymbol{P}_{G,\mathcal{D}} : 2^V \times 2^V \to [0, 1]$: For each $T \in 2^V$, $\boldsymbol{P}_{G,\mathcal{D}}[S, T]$ specifies the probability that T is the final activated set when S cascades its influence through the network G. Thus, $P_{\mathcal{I}} = (V, \boldsymbol{P}_{G,\mathcal{D}})$ defines a natural powerset network, which can be viewed as the *underlying network* induced by the interplay between the *static* network structure G and *dynamic* influence process $\mathcal{D}$.

[6] In the classical IC model, each directed edge $(u, v) \in E$ has an influence probability $p_{u,v} \in [0, 1]$. The probability profile defines a discrete-time influence process when given a seed set S: At time 0, nodes in S are activated while other nodes are inactive. At time $t \geq 1$, for any node u activated at time $t - 1$, it has one chance to activate each of its inactive out-neighbor v with an independent probability of $p_{u,v}$. When there is no more activation, this stochastic process ends with a random set of nodes activated during the process.

An important quality measure of S in this process is S's *influence spread* [55]. It can be defined from the powerset model $P_{\mathcal{I}} = (V, \boldsymbol{P}_{G,\mathcal{D}})$ as following:

$$\boldsymbol{\sigma}_{G,\mathcal{D}}(S) = \sum_{T \subseteq V} |T| \cdot \boldsymbol{P}_{G,\mathcal{D}}[S,T].$$

Thus, $(V, \boldsymbol{\sigma}_{G,\mathcal{D}})$ also defines a natural cooperative network [25].

In many applications and studies, *ordinal network models* rather than *cardinal network models* are used to capture the preferences among nodes. Two classical applications of preference frameworks are voting [10] and stable marriage/coalition formation [21, 43, 46, 79]. A modern use of preference models is the *Border Gateway Protocol* (BGP) for network routing between autonomous Internet systems [23, 77].

In a recent axiomatic study of community identification in social networks, Borgs et al. [11, 19] considered the following *abstract* social/information network framework. Below, for a non-empty finite set V, let $L(V)$ denote the set of all *linear orders* on V.

Definition 2.4 (Preference Model) A *preference network* over V is a pair $A = (V, \Pi)$, where $\Pi = \{\boldsymbol{\pi}_u\}_{u \in V} \in L(V)^{|V|}$ is a *preference profile* in which $\boldsymbol{\pi}_u$ specifies u's individual preference.

2.3.2 Understanding Network Facets and Network Concepts

Each network model enables us to focus on different facets of network data. For example, the powerset model offers the most natural framework for capturing the underlying interplay between influence processes and network structures. The cooperative model matches the explicit representation of clusterability, group utilities, and influence spreads. While traditional graph-based network data often consists solely of pairwise interactions, affinities, or associations, a community is formed by a group of individuals. Thus, the basic question for community identification is to understand "how do individual preferences (affinities/associations) result in group preferences or community coherence?" [19] The preference model highlights the fundamental aspect of community characterization. The preference model is also natural for addressing the question of summarizing individual preferences into one collective preference, which is fundamental in the formulation of network centrality [89]. Thus, studying network models beyond graphs helps to broaden our understanding of social/information networks.

Several these network models, as defined above, are highly theoretical models. Their complete-information profiles have exponential dimensionality in $|V|$. To use them as underlying models in network analysis, succinct representations should be constructed to efficiently capture observed network data. For example, both the conductance clusterability measure and the social-influence powerset network are succinctly defined. Characterizing network concepts in these models and effectively applying them to understanding real network data are promising and fundamentally challenging research directions in network science.

3 PageRank Completion

Network analysis is a task to capture the *essence* of the observed networks. For example, graph embedding [61, 89] can be viewed as a process to identify the geometric essence of networks. Similarly, network completion [48, 56, 63], graphon estimation [4, 20], and community recovering in hidden stochastic block models [2] can be viewed as processes to distill the statistical essence of networks. All these approaches build *constructive maps* from observed sparse graphs to underlying complete-information models. In this section, we study the following basic question:

> Given an observed sparse affinity network $G = (V, E, \mathbf{W})$, can we construct a complete-information affinity network that is consistent with G?

This question is simpler than but relevant to matrix and network completion [48, 56], which aims to infer the missing data from sparse, observed network data. Like matrix/network completion, this problem is mathematically an inverse problem. Conceptually, we need to formulate the meaning of "a complete-information affinity network consistent with G."

Our study is also partially motivated by the following question asked in [6, 11], aiming to deriving personalized ranking information from graph-based network data:

> Given a sparse affinity network $G = (V, E, \mathbf{W})$, how should we construct a complete-information preference model that best captures the underlying individual preferences from network data given by G?

We will prove the following basic structural result[7]: Every connected, undirected, weighted graph $G = (V, E, \mathbf{W})$ has an undirected and weighted graph $\overline{G} = (V, \overline{E}, \overline{\mathbf{W}})$, such that:

- **Complete Information**: $\overline{E}$ forms a complete graph with $|V|$ self-loops.
- **Degree and Stationary Preserving**: $\mathbf{W} \cdot \mathbf{1} = \overline{\mathbf{W}} \cdot \mathbf{1}$. Thus, the random-walk Markov chains on G and on $\overline{G}$ have the same stationary distribution.
- **PageRank Conforming**: The transition matrix $\mathbf{M}_{\overline{\mathbf{W}}}$ of the random-walk Markov chain on $\overline{G}$ is *conformal* to the PageRank of G, that is, $\mathbf{M}_{\overline{\mathbf{W}}}^T \cdot \mathbf{1}$ is proportional to the PageRank centrality of G
- **Spectral Approximation**: G and $\overline{G}$ are spectrally similar.

[7] See Theorem 3.5 for the precise statement.

In the last condition, the similarity between G and $\overline{G}$ is measured by the following notion of *spectral similarity* [85]:

Definition 3.1 (Spectral Similarity of Networks) Suppose $G = (V, E, \mathbf{W})$ and $\overline{G} = (V, \overline{E}, \overline{\mathbf{W}})$ are two weighted undirected graphs over the same set V of n nodes. Let $\mathbf{L}_{\mathbf{W}} = \mathbf{D}_{\mathbf{W}} - \mathbf{W}$ and $\mathbf{L}_{\overline{\mathbf{W}}} = \mathbf{D}_{\overline{\mathbf{W}}} - \overline{\mathbf{W}}$ be the *Laplacian matrices*, respectively, of these two graphs. Then, for $\sigma \geq 1$, we say G and $\overline{G}$ are σ*-spectrally similar* if:

$$\forall \mathbf{x} \in \mathbb{R}^n, \quad \frac{1}{\sigma} \cdot \mathbf{x}^T \mathbf{L}_{\overline{\mathbf{W}}} \mathbf{x} \leq \mathbf{x}^T \mathbf{L}_{\mathbf{W}} \mathbf{x} \leq \sigma \cdot \mathbf{x}^T \mathbf{L}_{\overline{\mathbf{W}}} \mathbf{x} \tag{7}$$

Many graph-theoretical measures, such as flows, cuts, conductances, effective resistances, are approximately preserved by spectral similarity [12, 85]. We refer to $\overline{G} = (V, \overline{E}, \overline{\mathbf{W}})$ as the *PageRank essence* or *PageRank completion* of $G = (V, E, \mathbf{W})$.

3.1 The Personalized PageRank Matrix

$\overline{G} = (V, \overline{E}, \overline{\mathbf{W}})$ stated above is derived from a well-known structure in network analysis, the personalized PageRank matrix of a network [8, 89].

3.1.1 Personalized PageRanks

Generalizing the Markov process of PageRank, Haveliwala [49] introduced personalized PageRanks. Suppose $G = (V, E, \mathbf{W})$ is a weighted directed graph and $\alpha > 0$ is a restart parameter. For any distribution $\mathbf{s}$ over V, consider the following Markov process, whose transition rule—for any $v \in V$—is the following:

- with probability α, restart at a random node in V according to distribution $\mathbf{s}$, and
- with probability $(1 - \alpha)$, move to a neighbor of v, chosen randomly with probability proportional to edge weights out of v.

Then, the *PageRank with respect to the starting vector* $\mathbf{s}$, denoted by $\mathbf{p}_{\mathbf{s}}$, is the stationary distribution of this Markov chain.

Let $d_u^{out} = \sum_{v \in V} w_{u,v}$ denotes the *out-degree* of $u \in V$ in G. Then, $\mathbf{p}_{\mathbf{s}}$ is the solution to the following equation:

$$\mathbf{p}_{\mathbf{s}} = \alpha \cdot \mathbf{s} + (1 - \alpha) \cdot \mathbf{W}^T \cdot \left(\mathbf{D}_{\mathbf{W}}^{out}\right)^{-1} \cdot \mathbf{p}_{\mathbf{s}} \tag{8}$$

where $\mathbf{D}_{\mathbf{W}}^{out} = \mathrm{diag}([d_1^{out}, \ldots, d_n^{out}])$ is the diagonal matrix of out degrees. Let $\mathbf{1}_u$ denote the n-dimensional vector whose uth location is 1 and all other entries in $\mathbf{1}_u$ are zeros. Haveliwala [49] referred to $\mathbf{p}_u := \mathbf{p}_{\mathbf{1}_u}$ as the *personalized PageRank* of $u \in V$ in G. Personalized PageRank is asymmetric, and hence to emphasize this fact, we express $\mathbf{p}_u$ as:

$$\mathbf{p}_u = (p_{u\to 1}, \ldots, p_{u\to n})^T .$$

Then $\{\mathbf{p}_u\}_{u\in V}$—the personalized PageRank profile—defines the following matrix:

Definition 3.2 (Personalized PageRank Matrix) The *personalized PageRank matrix* of an n-node weighted graph $G = (V, E, \mathbf{W})$ and restart constant $\alpha > 0$ is:

$$\mathbf{PPR}_{\mathbf{W},\alpha} = [\mathbf{p}_1, \ldots, \mathbf{p}_n]^T = \begin{bmatrix} p_{1\to 1} & \cdots & p_{1\to n} \\ \vdots & \ldots & \vdots \\ p_{n\to 1} & \cdots & p_{n\to n} \end{bmatrix} \tag{9}$$

In this article, we normalize the PageRank centrality so that the sum of the centrality values over all nodes is equal to n. Let $\mathbf{1}$ denote the n-dimensional vector of all 1s. Then, the *PageRank centrality* of G is the solution to the following *Markov random-walk* equation [49, 72]:

$$\mathbf{PageRank}_{\mathbf{W},\alpha} = \alpha \cdot \mathbf{1} + (1-\alpha) \cdot \mathbf{W}^T \left(\mathbf{D}_{\mathbf{W}}^{out}\right)^{-1} \mathbf{PageRank}_{\mathbf{W},\alpha} \tag{10}$$

Because $\mathbf{1} = \sum_u \mathbf{1}_u$, we have:

Proposition 3.3 (PageRank Conforming) *For any $G = (V, E, \mathbf{W})$ and $\alpha > 0$:*

$$\mathbf{PageRank}_{\mathbf{W},\alpha} = \sum_{u\in V} \mathbf{p}_u = \mathbf{PPR}_{\mathbf{W},\alpha}^T \cdot \mathbf{1} \tag{11}$$

Because Markov processes preserve the probability mass of the starting vector, we also have:

Proposition 3.4 (Markovian Conforming) *For any $G = (V, E, \mathbf{W})$ and $\alpha > 0$, $\mathbf{PPR}_{\mathbf{W},\alpha}$ is non-negative and:*

$$\mathbf{PPR}_{\mathbf{W},\alpha} \cdot \mathbf{1} = \mathbf{1} \tag{12}$$

In summary, the PageRank matrix $\mathbf{PPR}_{\mathbf{W},\alpha}$ is a special matrix associated with network G—its row sum is the vector of all 1s and its column sum is the PageRank centrality of G.

3.2 PageRank Completion of Symmetric Networks

PageRank centrality and personalized PageRank matrix apply to both directed and undirected weighted graphs. Both Propositions 3.3 and 3.4 also hold generally. In this subsection, we will focus mainly on undirected weighted networks. In such a case, let $\mathbf{D_W}$ be the diagonal matrix associated with weighted degrees $\mathbf{d_W} = \mathbf{W} \cdot \mathbf{1}$ and let $\mathbf{M_W} = \mathbf{D_W^{-1}W}$ be the standard random-walk transition matrix on G.

To state the theorem below, let's first review a basic concept of Markov chain. Recall that a *Markov chain* over V is defined by an $n \times n$ transition matrix $\mathbf{M}$ satisfying the *stochastic condition*: $\mathbf{M}$ is non-negative and $\mathbf{M} \cdot \mathbf{1} = \mathbf{1}$. A probability vector $\boldsymbol{\pi}$ is the *stationary distribution* of this Markov process if:

$$\mathbf{M}^T\boldsymbol{\pi} = \boldsymbol{\pi} \tag{13}$$

It is well known that every irreducible and ergodic Markov chain has a stationary distribution. Markov chain $\mathbf{M}$ is *detailed-balanced* if:

$$\boldsymbol{\pi}[u]\mathbf{M}[u,v] = \boldsymbol{\pi}[v]\mathbf{M}[v,u], \quad \forall\, u, v \in V \tag{14}$$

We will now prove the following structural result:

Theorem 3.5 (PageRank Completion) *For any weighted directed graph $G = (V, E, \mathbf{W})$ and restart constant $\alpha > 0$:*

A: $\mathbf{PPR}_{\mathbf{W},\alpha}$ *and* $\left(\mathbf{D}_{\mathbf{W}}^{out}\right)^{-1} \cdot \mathbf{W}$ *have the same eigenvectors. Thus, both Markov chains have the same stationary distribution.*
B: $\mathbf{PPR}_{\mathbf{W},\alpha}$ *is detailed-balanced if and only if* $\mathbf{W}$ *is symmetric.*

Furthermore, when $\mathbf{W}$ *is symmetric, let $\overline{G}_\alpha = (V, \overline{E}_\alpha, \overline{\mathbf{W}}_\alpha)$ be the affinity network such that:*

$$\overline{\mathbf{W}}_\alpha = \mathbf{D_W} \cdot \mathbf{PPR}_{\mathbf{W},\alpha} \quad \text{and} \quad \overline{E} = \{(u,v) : \overline{\mathbf{W}}_\alpha[u,v] > 0\} \tag{15}$$

Then, $\overline{G}_\alpha$ satisfies the following conditions:

1. **Symmetry Preserving:** $\overline{\mathbf{W}}^T = \overline{\mathbf{W}}$, *i.e., $\overline{G}_\alpha$ is an undirected affinity network.*
2. **Complete Information***: If G is connected, then $\overline{E}_\alpha$ is a complete graph with $|V|$ self-loops.*
3. **Degree and Stationary Preserving***:* $\mathbf{W} \cdot \mathbf{1} = \overline{\mathbf{W}} \cdot \mathbf{1}$. *Thus,* $\mathbf{D_W} = \mathbf{D}_{\overline{\mathbf{W}}}$ *and the random-walk Markov chains* $\mathbf{M_W}$ *and* $\mathbf{M}_{\overline{\mathbf{W}}}$ *have the same stationary distribution.*
4. **Markovian and PageRank Conforming***:*

$$\mathbf{M}_{\overline{\mathbf{W}}} \cdot \mathbf{1} = \mathbf{1} \quad \text{and} \quad \mathbf{M}_{\overline{\mathbf{W}}}^T \cdot \mathbf{1} = \mathbf{PageRank}_{\mathbf{W},\alpha} \tag{16}$$

5. **Simultaneously Diagonalizable***: For any symmetric* $\mathbf{W}$, *recall* $\mathbf{L_W} = \mathbf{D_W} - \mathbf{W}$ *denotes the Laplacian matrix associated with* $\mathbf{W}$. *Let* $\mathcal{L}_{\mathbf{W}} = \mathbf{D}_{\mathbf{W}}^{-\frac{1}{2}}\mathbf{L_W}\mathbf{D}_{\mathbf{W}}^{\frac{1}{2}} = \mathbf{I} - \mathbf{D}_{\mathbf{W}}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}_{\mathbf{W}}^{-\frac{1}{2}}$ *be the normalized Laplacian matrix associated with* $\mathbf{W}$. *Then,* $\mathcal{L}_{\mathbf{W}}$ *and* $\mathcal{L}_{\overline{\mathbf{W}}}$ *are simultaneously diagonalizable.*

6. **Spectral Densification and Approximation***: For all* $\mathbf{x} \in \mathbb{R}^n$*:*

$$\alpha \cdot \mathbf{L}_{\mathbf{W}} \leq \mathbf{x}^T \left(\frac{1}{1-\alpha} \cdot \mathbf{L}_{\overline{\mathbf{W}}} \right) \mathbf{x} \leq \frac{1}{\alpha} \mathbf{L}_{\mathbf{W}} \tag{17}$$

$$\alpha \cdot \mathcal{L}_{\mathbf{W}} \leq \mathbf{x}^T \left(\frac{1}{1-\alpha} \cdot \mathcal{L}_{\overline{\mathbf{W}}} \right) \mathbf{x} \leq \frac{1}{\alpha} \mathcal{L}_{\mathbf{W}} \tag{18}$$

In other words, G and $\frac{1}{1-\alpha} \cdot \overline{G}_\alpha$ *are* $\frac{1}{\alpha}$*-spectrally similar.*

Remarks We rescale $\mathbf{L}_{\overline{\mathbf{W}}}$ and $\mathcal{L}_{\overline{\mathbf{W}}}$ by $\frac{1}{1-\alpha}$ because $\overline{G}_\alpha$ has self-loops of magnitude $\alpha \mathbf{D}_{\mathbf{W}}$. In other words, $\overline{G}_\alpha$ only uses $(1 - \alpha)$ fraction of its weighted degrees for connecting different nodes in V.

Proof Let $n = |V|$. For any initial distribution $\mathbf{s}$ over V, we can explicitly express $\mathbf{p}_{\mathbf{s}}$ as:

$$\mathbf{p}_{\mathbf{s}} = \alpha \sum_{k=0}^{\infty} (1-\alpha)^k \cdot \left(\mathbf{W}^T \cdot (\mathbf{D}_{\mathbf{W}}^{out})^{-1} \right)^k \cdot \mathbf{s} \tag{19}$$

Consequently: we can express $\mathbf{PPR}_{\mathbf{W},\alpha}$ as:

$$\mathbf{PPR}_{\mathbf{W},\alpha} = \alpha \sum_{k=0}^{\infty} (1-\alpha)^k \cdot \left((\mathbf{D}_{\mathbf{W}}^{out})^{-1} \cdot \mathbf{W} \right)^k \tag{20}$$

Note that $\alpha \sum_{k=0}^{\infty} (1-\alpha)^k = 1$. Thus, $\mathbf{PPR}_{\mathbf{W},\alpha}$ is a convex combination of (multi-step) random-walk matrices defined by $(\mathbf{D}_{\mathbf{W}}^{out})^{-1} \cdot \mathbf{W}$. Statement A follows directly from the fact that $\left((\mathbf{D}_{\mathbf{W}}^{out})^{-1} \cdot \mathbf{W} \right)^k$ is a stochastic matrix for any integer $k \geq 0$.

The following fact is well known (Aldous and Fill, recompiled 2014, Reversible Markov chains and random walks on graphs, Unfinished monograph. Available at http://www.stat.berkeley.edu~aldous/RWG/book.html):

> Suppose $\mathbf{M}$ is a Markov chain with stationary distribution $\boldsymbol{\pi}$. Let $\boldsymbol{\Pi}$ be the diagonal matrix defined by $\boldsymbol{\pi}$. Then, $\mathbf{M}^T \boldsymbol{\Pi}$ is symmetric if and only if the Markov process defined by $\mathbf{M}$ is detailed balanced.

We now assume $\mathbf{W} = \mathbf{W}^T$. Then, Eq. (20) becomes:

$$\mathbf{PPR}_{\mathbf{W},\alpha} = \alpha \sum_{k=0}^{\infty} (1-\alpha)^k \cdot \left(\mathbf{D}_{\mathbf{W}}^{-1} \cdot \mathbf{W} \right)^k \tag{21}$$

The stationary distribution of $\mathbf{D}_{\mathbf{W}}^{-1}\mathbf{W}$—and hence of $\mathbf{PPR}_{\mathbf{W},\alpha}$—is proportional to $\mathbf{d} = \mathbf{W} \cdot \mathbf{1}$. $\mathbf{PPR}_{\mathbf{W},\alpha}$ is detailed balanced because $\overline{\mathbf{W}} = \mathbf{D}_{\mathbf{W}} \cdot \mathbf{PPR}_{\mathbf{W},\alpha}$ is a symmetric matrix. Because $\left((\mathbf{D}_{\mathbf{W}}^{out})^{-1} \cdot \mathbf{W} \right)^k$ (for all positive integers) have a common stationary distribution, $\mathbf{PPR}_{\mathbf{W},\alpha}$ is not detailed balanced when $\mathbf{W}$ is not symmetric. It is also well known—by Eq. (19)—that for all $u, v \in V$, $\mathbf{PPR}_{\mathbf{W},\alpha}[u, v]$ is equal to the probability that a run of random walk starting at u passes by v immediately before it restarts. Thus, when G is connected, $\mathbf{PPR}_{\mathbf{W},\alpha}[u, v] > 0$ for all $u, v \in V$. Thus, $\mathrm{nnz}(\overline{\mathbf{W}}_\alpha) = n^2$, and $\overline{E}_\alpha$, the nonzero pattern of $\overline{\mathbf{W}}_\alpha$, is a complete graph with $|V|$ self-loops. We have now established Condition B and Conditions 1–4.

We now prove Conditions 5 and 6.[8] Recall that when $\mathbf{W} = \mathbf{W}^T$, we can express the personalized PageRank matrix as:

$$\mathbf{PPR}_{\mathbf{W},\alpha} = \alpha \sum_{k=0}^{\infty} (1-\alpha)^k \cdot \left(\mathbf{D}_{\mathbf{W}}^{-1} \cdot \mathbf{W}\right)^k .$$

Thus:

$$\overline{\mathbf{W}}_\alpha = \mathbf{D}_{\mathbf{W}} \cdot \mathbf{PPR}_{\mathbf{W},\alpha} = \left(\alpha \sum_{k=0}^{\infty} (1-\alpha)^k \cdot \mathbf{D}_{\mathbf{W}} \cdot \left(\mathbf{D}_{\mathbf{W}}^{-1}\mathbf{W}\right)^k\right).$$

We compare the Laplacian matrices associated with $\mathbf{W}$ and $\overline{\mathbf{W}}$:

$$\mathbf{L}_{\mathbf{W}} = \mathbf{D}_{\mathbf{W}} - \mathbf{W} = \mathbf{D}_{\mathbf{W}}^{1/2}\left(\mathbf{I} - \mathbf{D}_{\mathbf{W}}^{-1/2}\mathbf{W}\mathbf{D}_{\mathbf{W}}^{-1/2}\right)\mathbf{D}_{\mathbf{W}}^{1/2} = \mathbf{D}_{\mathbf{W}}^{1/2}\mathcal{L}_{\mathbf{W}}\mathbf{D}_{\mathbf{W}}^{1/2}.$$
$$\mathbf{L}_{\overline{\mathbf{W}}} = \mathbf{D}_{\mathbf{W}} - \overline{\mathbf{W}} = \mathbf{D}_{\mathbf{W}}^{1/2}\mathcal{L}_{\overline{\mathbf{W}}}\mathbf{D}_{\mathbf{W}}^{1/2}$$

where

$$\mathcal{L}_{\overline{\mathbf{W}}} = \mathbf{I} - \alpha \sum_{k=0}^{\infty} (1-\alpha)^k \cdot (\mathbf{D}_{\mathbf{W}}^{-1/2}\mathbf{W}\mathbf{D}_{\mathbf{W}}^{-1/2})^k.$$

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ be the n eigenvalues of $\mathbf{D}_{\mathbf{W}}^{-1/2}\mathbf{W}\mathbf{D}_{\mathbf{W}}^{-1/2}$. Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ denote the unit-length eigenvectors of $\mathbf{D}_{\mathbf{W}}^{-1/2}\mathbf{W}\mathbf{D}_{\mathbf{W}}^{-1/2}$ associated with eigenvalues $\lambda_1, \cdots, \lambda_n$, respectively. We have $|\lambda_i| \leq 1$. Let $\mathbf{\Lambda}$ be the diagonal matrix associated with $(\lambda_1, \ldots, \lambda_n)$ and $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$. By the spectral theorem—i.e., the eigenvalue decomposition for symmetric matrices—we have:

$$\mathbf{U}^T\mathbf{D}_{\mathbf{W}}^{-1/2}\mathbf{W}\mathbf{D}_{\mathbf{W}}^{-1/2}\mathbf{U} = \mathbf{\Lambda} \tag{22}$$
$$\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I} \tag{23}$$

[8] Thanks to Dehua Cheng of USC for assisting this proof.

Therefore:

$$\begin{aligned}\mathbf{L}_{\mathbf{W}} &= \mathbf{D}_{\mathbf{W}}^{1/2}\mathbf{U}\mathbf{U}^{T}\left(\mathbf{I}-\mathbf{D}_{\mathbf{W}}^{-1/2}\mathbf{W}\mathbf{D}_{\mathbf{W}}^{-1/2}\right)\mathbf{U}\mathbf{U}^{T}\mathbf{D}_{\mathbf{W}}^{1/2}\\ &= \mathbf{D}_{\mathbf{W}}^{1/2}\mathbf{U}\left(\mathbf{I}-\mathbf{U}^{T}\mathbf{D}_{\mathbf{W}}^{-1/2}\mathbf{W}\mathbf{D}_{\mathbf{W}}^{-1/2}\mathbf{U}\right)\mathbf{U}^{T}\mathbf{D}_{\mathbf{W}}^{1/2}\\ &= \mathbf{D}_{\mathbf{W}}^{1/2}\mathbf{U}\left(\mathbf{I}-\mathbf{\Lambda}\right)\mathbf{U}^{T}\mathbf{D}_{\mathbf{W}}^{1/2}.\end{aligned}$$

Similarly:

$$\begin{aligned}\mathbf{L}_{\overline{\mathbf{W}}_{\alpha}} &= \mathbf{D}_{\mathbf{W}}-\overline{\mathbf{W}}_{\alpha} = \mathbf{D}_{\mathbf{W}}^{1/2}\mathcal{L}_{\overline{\mathbf{W}}}\mathbf{D}_{\mathbf{W}}^{1/2}\\ &= \mathbf{D}_{\mathbf{W}}^{1/2}\left(\mathbf{I}-\alpha\sum_{k=0}^{\infty}(1-\alpha)^{k}\cdot(\mathbf{D}_{\mathbf{W}}^{-1/2}\mathbf{W}\mathbf{D}_{\mathbf{W}}^{-1/2})^{k}\right)\mathbf{D}_{\mathbf{W}}^{1/2}\\ &= \mathbf{D}_{\mathbf{W}}^{1/2}\mathbf{U}\left(\mathbf{I}-\alpha\sum_{k=0}^{\infty}(1-\alpha)^{k}\cdot\mathbf{U}^{T}(\mathbf{D}_{\mathbf{W}}^{-1/2}\mathbf{W}\mathbf{D}_{\mathbf{W}}^{-1/2})^{k}\mathbf{U}\right)\mathbf{U}^{T}\mathbf{D}_{\mathbf{W}}^{1/2}\\ &= \mathbf{D}_{\mathbf{W}}^{1/2}\mathbf{U}\left(\mathbf{I}-\alpha\sum_{k=0}^{\infty}(1-\alpha)^{k}\cdot\mathbf{\Lambda}^{k}\right)\mathbf{U}^{T}\mathbf{D}_{\mathbf{W}}^{1/2}\\ &= \mathbf{D}_{\mathbf{W}}^{1/2}\mathbf{U}\left(\mathbf{I}-\frac{\alpha}{\mathbf{I}-(1-\alpha)\mathbf{\Lambda}}\right)\mathbf{U}^{T}\mathbf{D}_{\mathbf{W}}^{1/2}.\end{aligned}$$

The derivation above has proved Condition (5). To prove Condition (6), consider an arbitrary $\mathbf{x}\in\mathbb{R}^{n}\setminus\{\mathbf{0}\}$. With $\mathbf{y}=\mathbf{U}^{T}\mathbf{D}_{\mathbf{W}}^{1/2}\mathbf{x}$, we have:

$$\begin{aligned}\frac{\mathbf{x}^{T}\frac{1}{1-\alpha}\mathbf{L}_{\overline{\mathbf{W}}}\mathbf{x}}{\mathbf{x}^{T}\mathbf{L}_{\mathbf{W}}\mathbf{x}} &= \frac{1}{1-\alpha}\cdot\frac{\mathbf{x}^{T}\mathbf{D}_{\mathbf{W}}^{1/2}\mathbf{U}\left(\mathbf{I}-\frac{\alpha}{\mathbf{I}-(1-\alpha)\mathbf{\Lambda}}\right)\mathbf{U}^{T}\mathbf{D}_{\mathbf{W}}^{1/2}\mathbf{x}}{\mathbf{x}^{T}\mathbf{D}_{\mathbf{W}}^{1/2}\mathbf{U}\left(\mathbf{I}-\mathbf{\Lambda}\right)\mathbf{U}^{T}\mathbf{D}_{\mathbf{W}}^{1/2}\mathbf{x}}\\ &= \frac{1}{1-\alpha}\cdot\frac{\mathbf{y}^{T}\left(\mathbf{I}-\frac{\alpha}{\mathbf{I}-(1-\alpha)\mathbf{\Lambda}}\right)\mathbf{y}}{\mathbf{y}^{T}\left(\mathbf{I}-\mathbf{\Lambda}\right)\mathbf{y}}\end{aligned}$$

This ratio is in the interval of:

$$\left[\inf_{\lambda:|\lambda|\leq 1}\frac{1}{1-(1-\alpha)\lambda},\ \sup_{\lambda:|\lambda|\leq 1}\frac{1}{1-(1-\alpha)\lambda}\right]=\left[\frac{1}{2-\alpha},\frac{1}{\alpha}\right].$$

□

3.3 PageRank Completion, Community Identification, and Clustering

PageRank completion has an immediate application to the community-identification approaches developed in [11, 19]. This family of methods first constructs a preference network from an input weighted graph $G = (V, E, \mathbf{W})$. It then applies various social-choice aggregation functions [10] to define network communities [11, 19]. In fact, Balcan et al. [11] show that the PageRank completion of G provides a wonderful scheme (see also in Definition 4.10) for constructing preference networks from affinity networks.

In addition to its classical connection with PageRank centrality, PageRank completion also has a direct connection with network clustering. To illustrate this connection, let's recall a well-known approach in spectral graph theory for clustering [9, 24, 62, 84, 86]:

Algorithm: Sweep$(G, \mathbf{v})$

```
Require: $G = (V, E, \mathbf{W})$ and $\mathbf{v} \in \mathbb{R}^{|V|}$
Let $\boldsymbol{\iota}$ be an ordering of $V$ according to $\mathbf{v}$, i.e., $\forall k \in [n-1]$, $\mathbf{v}[\pi(k)] \geq \mathbf{v}[\pi(k+1)]$
Let $S_k = \{\pi(1), \ldots, \pi(k)\}$
Let $k^* = \operatorname{argmin}_k \text{conductance}_{\mathbf{W}}(S_k)$.
Return $S_{k^*}$
```

Both in theory and in practice, the most popular vectors used in Sweep are:

- **Fiedler vector:** the eigenvector associated with the second smallest eigenvalue of the Laplacian matrix $\mathbf{L}_{\mathbf{W}}$ [39, 40, 84].
- **Cheeger vector:** $\mathbf{D}_{\mathbf{W}}^{-1/2}\boldsymbol{v}_2$, where $\boldsymbol{v}_2$ is the eigenvector associated with the second smallest eigenvalue of the normalized Laplacian matrix $\mathcal{L}_{\mathbf{W}}$ [24, 28].

The sweep-based clustering method and Fiedler/Cheeger vectors are the main subject of following beautiful theorem [24] in spectral graph theory:

Theorem 3.6 (Cheeger's Inequality) *For any symmetric weighted graph* $G = (V, E, \mathbf{W})$, *let* λ_2 *be the second smallest eigenvalue of the normalized Laplacian matrix* $\mathcal{L}_{\mathbf{W}}$ *of* G. *Let* $\boldsymbol{v}_2$ *be the eigenvector associated with* λ_2 *and* $S =$ Sweep$(G, \mathbf{D}_{\mathbf{W}}^{-1/2}\mathbf{v}_2)$. *Then:*

$$\frac{\lambda_2}{2} \leq \text{conductance}_{\mathbf{W}}(S) \leq \sqrt{2\lambda_2} \tag{24}$$

By Theorem 3.5, the normalized Laplacian matrices of G and its PageRank completion are simultaneously diagonalizable. Thus, we can also use the eigenvector of the PageRank completion of G to identify a cluster of G whose conductance is guaranteed by the Cheeger's inequality.

> Then, how is the PageRank completion necessarily a better representation of the information contained in the original network?
> For example, with respect to network clustering, what desirable properties does the PageRank completion have that the original graph doesn't?

While we are still looking for a comprehensive answer to these questions, we will now use the elegant result of Andersen, Chung, and Lang [9] to illustrate that the PageRank completion indeed contains more *direct* information about network clustering than the original data $\mathbf{W}$. Andersen et al. proved that if one applies sweep to vectors $\{\mathbf{D}_{\mathbf{W}}^{-1} \cdot \mathbf{p}_v\}_{v \in V}$, then one can obtain a cluster whose conductance is nearly as small as that guaranteed by Cheeger's inequality. Such a statement does not hold for the rows in the original network data $\mathbf{W}$, particularly when $\mathbf{W}$ is sparse.

In fact, the result of Andersen, Chung, and Lang [9] is much stronger. They showed that for any cluster $S \subset V$, if one selects a random node $v \in S$ with probability proportional to the weighted degree d_v of the node, then, with probability at least $1/2$, one can identify a cluster S' of conductance at most $O(\sqrt{\text{conductance}_{\mathbf{W}}(S) \log n})$ by applying sweep to vector $\mathbf{D}_{\mathbf{W}}^{-1} \cdot \mathbf{p}_v$. In other words, the row vectors in the PageRank completion—i.e., the personalized PageRank vectors that represent the individual data associated with nodes—have rich and direct information about network clustering (measured by conductance). This is a property that the original network data simply doesn't have, as one is usually not able to identify good clusters directly from the individual rows of $\mathbf{W}$.

In summary, Cheeger's inequality and its algorithmic proof can be viewed as the mathematical foundation for *global* spectral partitioning, because the Fiedler/Cheeger vectors are formulated from the network data as a whole. From this global perspective, both the original network and its PageRank completion are equally effective. In contrast, from the local perspective of individual-row data, Andersen, Chung, and Lang's result highlights the effectiveness of the PageRank completion to *local clustering* [86]: The row data associated with nodes in the PageRank completion provides effective information for identifying good clusters. Similarly, from the corresponding column in the PageRank completion, one can also directly and "locally" obtains each node's PageRank centrality. In other words, PageRank completion transforms the input network data $\mathbf{W}$ into a "complete-information" network model $\overline{\mathbf{W}}$, and in the process, it distilled the centrality/clusterability information implicitly embedded globally in $\mathbf{W}$ into an ensemble of nodes' "individual" network data that explicitly encodes the centrality information and locally capturing the clustering structures.

4 Connecting Multifaceted Network Data

The formulations highlighted in Sect. 2.3, such as the cooperative, incentive, powerset, and preference models, are just a few examples of network models beyond the traditional graph-based framework. Other extensions include the popular probabilistic graphical model [58] and game-theoretical graphical model [26, 31, 52].

These models use relatively homogeneous node and edge types, but nevertheless represent a great source of expressions for multifaceted and multimodal network data.

While diverse network models enable us to express multifaceted network data, we need mathematical and algorithmic tools to connect them. For some applications such as community identification, one may need to properly use some data facets as metadata to evaluate or cross validate the network solution(s) identified from the main network facets [74].

> But more broadly, for many real-world network analysis tasks, we need a systematic approach to *network composition* whose task is to integrate the multifaceted data into a single effective network worldview. Towards this goal, a basic theoretical step in multifaceted network analysis is to establish a unified worldview for capturing multifaceted network data expressed in various models.

Although fundamental, formulating a unified worldview of network models is still largely an outstanding research problem. In this section, we sketch our preliminary studies in using Markov chains to build a "common platform" for the network models discussed in Sect. 2.3. We hope this study will inspire a general theory for data integration, network composition, and multifaceted network analysis. We also hope that it will help to strengthen the connection between various fields, as diverse as statistical modeling, geometric embedding, social influence, network dynamics, game theory, and social choice theory, as well as various application domains (protein-protein interaction, viral marketing, information propagation, electoral behavior, homeland security, healthcare, etc.), that have provided different but valuable techniques and motivations to network analysis.

4.1 Centrality-Conforming Stochastic Matrices of Various Network Models

Markov chain—a basic statistical model—is also a fundamental network concept. For a weighted network $G = (V, E, \mathbf{W})$, the standard random-walk transition $\left(\mathbf{D}_{\mathbf{W}}^{out}\right)^{-1} \cdot \mathbf{W}$ is the most widely-used stochastic matrix associated with G. Importantly, Sect. 3 illustrates that other Markov chains—such as PageRank Markov chain $\mathbf{PPR}_{\mathbf{W},\alpha}$—are also natural with respect to network data $\mathbf{W}$. Traditionally, a Markov chain is characterized by its stochastic condition, stationary distribution, mixing time, and detailed-balancedness. Theorem 3.5 highlights another important feature of Markov chains in the context of network analysis: *The PageRank Markov chain is conforming with respect to PageRank centrality,* that is, for any network $G = (V, E, \mathbf{W})$ and $\alpha > 0$, we have:

$$\mathbf{PPR}_{\mathbf{W},\alpha}^{T} \cdot \mathbf{1} = \mathbf{PageRank}_{\mathbf{W},\alpha}.$$

> How should we derive stochastic matrices from other network models? Can we construct Markov chains that are centrality-confirming with respect to natural centrality measures of these network models?

In this section, we will examine some centrality-confirming Markov chains that can be derived from network data given by preference/incentive/cooperative/powerset models.

4.1.1 The Preference Model

For the preference model, there is a family of natural Markov chains, based on weighted aggregations in social-choice theory [10]. For a fixed n, let $\mathbf{w} \in (\mathbb{R}^+ \cup \{0\})^n$ be a non-negative and monotonically non-increasing vector. For the discussion below, we will assume that $\mathbf{w}$ is normalized such that $\sum_{i=1}^{n} \mathbf{w}[i] = 1$. For example, while the famous Borda count [93] uses $\mathbf{w} = [n, n-1, \ldots, 1]^T$, the normalized Borda count uses $\mathbf{w} = [n, n-1, \ldots, 1]^T / \binom{n}{2}$.

Proposition 4.1 (Weighted Preference Markov Chain) *Suppose $A = (V, \Pi)$ is a* preference network *over $V = [n]$ and $\mathbf{w}$ is non-negative and monotonically non-increasing weight vector, with $||\mathbf{w}||_1 = 1$. Let $\mathbf{M}_{A,\mathbf{w}}$ be the matrix in which for each $u \in V$, the uth row of $\mathbf{M}_{A,\mathbf{w}}$ is:*

$$\boldsymbol{\pi}_u \circ \mathbf{w} = [\mathbf{w}[\boldsymbol{\pi}_u(1)], \ldots, \mathbf{w}(\boldsymbol{\pi}_u(n))].$$

Then, $\mathbf{M}_{A,\mathbf{w}}$ defines a Markov chain, i.e., $\mathbf{M}_{A,\mathbf{w}}\mathbf{1} = \mathbf{1}$.

Proof $\mathbf{M}_{A,\mathbf{w}}$ is a stochastic matrix because each row of $\mathbf{M}_{A,\mathbf{w}}$ is a permutation of $\mathbf{w}$, and permutations preserve the L1-norm of the vector. □

Social-choice aggregation based on $\mathbf{w}$ also defines the following natural centrality measure, which can be viewed as the collective ranking over V based on the preference profiles of $A = (V, \Pi)$:

$$\text{centrality}_{\Pi,\mathbf{w}}[v] = \sum_{u \in V} \mathbf{w}[\pi_u(v)] \tag{25}$$

Like PageRank Markov chains, weighted preference Markov chains also enjoy the centrality-conforming property:

Proposition 4.2 *For any preference network $A = (V, \Pi)$, in which $\Pi \in L(V)^{|V|}$:*

$$\mathbf{M}_{A,\mathbf{w}}^T \cdot \mathbf{1} = \text{centrality}_{\Pi,\mathbf{w}} \tag{26}$$

4.1.2 The Incentive Model

We now focus on a special family of incentive networks: We assume for $U = (V, \boldsymbol{u})$ and $s \in V$:

1. u_s is monotonically non-decreasing, i.e., for all $T_1 \subset T_2$, $u_s(T_1) \leq u_s(T_2)$.
2. u_s is normalized, i.e., $u_s(V \setminus \{s\}) = 1$.

Each incentive network defines a natural cooperative network, $H_U = (V, \boldsymbol{\tau}_{SocialUtility})$: For any $S \subseteq V$, let the *social utility* of S be:

$$\boldsymbol{\tau}_{SocialUtility}(S) = \sum_{s \in S} u_s(S \setminus \{s\}) \tag{27}$$

The Shapley value [81]—a classical game-theoretical concept—provides a natural centrality measure for cooperative networks.

Definition 4.1 (Shapley Value) Suppose $\boldsymbol{\tau}$ is the characteristic function of a cooperative game over $V = [n]$. Recall that $L(V)$ denotes the set of all permutations of V. Let $S_{\boldsymbol{\pi},v}$ denotes the set of players preceding v in a permutation $\boldsymbol{\pi} \in L(V)$. Then, the *Shapley value* $\boldsymbol{\phi}_{\boldsymbol{\tau}}^{Shapley}[v]$ of a player $v \in V$ is:

$$\boldsymbol{\phi}_{\boldsymbol{\tau}}^{Shapley}[v] = \mathrm{E}_{\boldsymbol{\pi} \sim L(V)} \left[\boldsymbol{\tau}[S_{\boldsymbol{\pi},v} \cup \{v\}] - \boldsymbol{\tau}[S_{\boldsymbol{\pi},v}]\right] \tag{28}$$

The Shapley value $\boldsymbol{\phi}_{\boldsymbol{\tau}}^{Shapley}[v]$ of player $v \in V$ is the *expected marginal contribution* of v over the set preceding v in a random permutation of the players. The Shapley value has many attractive properties, and is widely considered to be the *fairest* measure of a player's power index in a cooperative game.

We can use Shapley values to define both the stochastic matrix and the centrality of incentive networks U. Let centrality_U be the Shapley value of the cooperative game defined by $\boldsymbol{\tau}_{SocialUtility}$. Note that the incentive network U also defines $|V|$ natural individual cooperative networks: For each $s \in V$ and $T \subset V$, let:

$$\boldsymbol{\tau}_s(T) = \begin{cases} u_s(T \setminus \{s\}) & \text{if } s \in T \\ 0 & \text{if } s \notin T \end{cases} \tag{29}$$

Proposition 4.4 (The Markov Chain of Monotonic Incentive Model) *Suppose $U = (V, \boldsymbol{u})$ is an incentive network over $V = [n]$, such that $\forall s \in V$, u_s is monotonically non-decreasing and $u_s(V \setminus \{s\}) = 1$. Let $\mathbf{M}_U$ be the matrix in which for each $s \in V$, the sth row of $\mathbf{M}_U$ is the Shapley value of the cooperative game with characteristic function $\boldsymbol{\tau}_s$. Then, $\mathbf{M}_U$ defines a Markov chain and is centrality-conforming with respect to* centrality_U, *i.e., (1) $\mathbf{M}_U\mathbf{1} = \mathbf{1}$ and (2)* $\mathbf{M}_U^T\mathbf{1} = \text{centrality}_U$.

Proof This proposition is the direct consequence of two basic properties of Shapley's beautiful characterization [81]:

1. The Shapley value is *efficient*: $\sum_{v \in V} \phi_{\boldsymbol{\tau}}[v] = \boldsymbol{\tau}(V)$.
2. The Shapley value is *Linear*: For any two characteristic functions $\boldsymbol{\tau}$ and $\boldsymbol{\omega}$, $\phi_{\boldsymbol{\tau}+\boldsymbol{\omega}} = \phi_{\boldsymbol{\tau}} + \phi_{\boldsymbol{\omega}}$.

By the assumption u_s is monotonically non-decreasing, we can show that every entry of the Shapley value (as given by Eq. (28)) is non-negative. Then, it follows from the efficiency of Shapley values and the assumption that $\forall s \in V, u_s(V \setminus \{s\}) = 1$, that $\mathbf{M}_U$ is a stochastic matrix, and hence it defines a Markov chain. Furthermore, we have:

$$\boldsymbol{\tau}_{SocialUtility} = \sum_{s \in V} \boldsymbol{\tau}_s \tag{30}$$

Because centrality$_U$ is the Shapley value of the cooperative game with characteristic function $\boldsymbol{\tau}_{SocialUtility}$, the *linearity* of the Shapley value then implies $\mathbf{M}_U^T \mathbf{1} =$ centrality$_U$, i.e., $\mathbf{M}_U$ is centrality-conforming with respect to centrality$_U$. □

4.1.3 The Influence Model

Centrality-conforming Markov chain can also be naturally constructed for a family of powerset networks. Recall from Sect. 2.3 that an influence process $\mathcal{D}$ and social network $G = (V, E)$ together define a powerset network, $\boldsymbol{P}_{G,\mathcal{D}} : 2^V \times 2^V \rightarrow [0, 1]$, where for each $T \in 2^V$, $\boldsymbol{P}_{G,\mathcal{D}}[S, T]$ specifies the probability that T is the final activated set when S cascades its influence through G. As observed in [25], the influence model also defines a natural cooperative game, whose characteristic function is the influence spread function:

$$\boldsymbol{\sigma}_{G,\mathcal{D}}(S) = \sum_{T \subseteq V} |T| \cdot \boldsymbol{P}_{G,\mathcal{D}}[S, T], \quad \forall S \subseteq V.$$

Chen and Teng [25] proposed to use the Shapley value of this social-influence game as a centrality measure of the powerset network defined by $\boldsymbol{P}_{G,\mathcal{D}}$. They showed that this social-influence centrality measure, to be denoted by centrality$_{G,\mathcal{D}}$, can be uniquely characterized by a set of five natrual axioms [25]. Motivated by the PageRank Markov chain, they also constructed the following centrality-conforming Markov chain for social-influence models.

Proposition 4.5 (Social-Influence Markov Chain) *Suppose $G = (V, E)$ is a social network and $\mathcal{D}$ is a social-influence process. Let $\mathbf{M}_{G,\mathcal{D}}$ be the matrix in which for each $v \in V$, the vth row of $\mathbf{M}_{G,\mathcal{D}}$ is given by the Shapley value of the cooperative game with the following characteristic function:*

$$\boldsymbol{\sigma}_{G,\mathcal{D},v}(S) = \sum_{T \subseteq V} [v \in T] \cdot \boldsymbol{P}_{G,\mathcal{D}}[S, T] \tag{31}$$

where $[v \in T]$ *is the indicator function for event* $(v \in T)$*. Then,* $\mathbf{M}_{G,\mathcal{D}}$ *defines a Markov chain and is centrality-conforming with respect to* $\text{centrality}_{G,\mathcal{D}}$*, i.e., (*1*)* $\mathbf{M}_{G,\mathcal{D}}\mathbf{1} = \mathbf{1}$ *and (*2*)* $\mathbf{M}_{G,\mathcal{D}}^T\mathbf{1} = \text{centrality}_{G,\mathcal{D}}$.

Proof For all $v \in V$, the characteristic function $\boldsymbol{\sigma}_{G,\mathcal{D},v}$ satisfies the following two conditions:

1. $\boldsymbol{\sigma}_{G,\mathcal{D},v}$ is monotonically non-decreasing.
2. $\boldsymbol{\sigma}_{G,\mathcal{D},v}(V) = 1$.

The rest of the proof is essentially the same as the proof of Proposition 4.4. □

4.2 Networks Associated with Markov Chains

The common feature in the Markovian formulations of Sect. 4.1 suggests the possibility of a general theory that various network models beyond graphs can be succinctly analyzed through the worldview of Markov chains. Such analyses are forms of dimension reduction of network data—the Markov chains derived, such as from social-influence instances, usually have lower dimensionality than the original network models. In dimension reduction of data, inevitably some information is lost. Thus, which Markov chain is formulated from a particular network model may largely depend on through which mathematical lens we are looking at the network data. The Markovian formulations of Sect. 4.1 are largely based on centrality formulations. Developing a more general Markovian formulation theory of various network models remains the subject of future research.

But once we can reduce the network models specifying various aspects of network data to a collection of Markov chains representing the corresponding network facets, we effectively reduce multifaceted network analysis to a potentially simpler task—the analysis of multilayer networks [57, 60]. Thus, we can apply various emerging techniques for multilayer network analysis [47, 73, 94] and network composition [60]. We can further use standard techniques to convert the Markov chains into weighted graphs to examine these network models through the popular graph-theoretical worldview.

4.2.1 Random-Walk Connection

Because of the following characterization, the random-walk is traditionally the most commonly-used connection between Markov chains and weighted networks.

Proposition 4.6 (Markov Chains and Networks: Random-Walk Connection) *For any directed network* $G = (V, E, \mathbf{W})$ *in which every node has at least one out-neighbor, there is a unique transition matrix:*

$$\mathbf{M}_{\mathbf{W}} = \left(\mathbf{D}_{\mathbf{W}}^{out}\right)^{-1} \mathbf{W}$$

that captures the (unbiased) random-walk Markov process on G. Conversely, given a transition matrix $\mathbf{M}$*, there is an infinite family of weighted networks whose random-walk Markov chains are consistent with* $\mathbf{M}$*. This family is given by:*

$$\{\boldsymbol{\Gamma}\mathbf{M} : \boldsymbol{\Gamma} \textit{ is a positive diagonal matrix}\}.$$

The most commonly-used diagonal scaling is $\boldsymbol{\Pi}$, the diagonal matrix of the stationary distribution. This scaling is partially justified by the fact that $\boldsymbol{\Pi}\mathbf{M}$ is an undirected network if and only if $\mathbf{M}$ is a detailed-balanced Markov chain. In fact in such a case, $\boldsymbol{\Gamma}\mathbf{M}$ is symmetric if and only if there exists $c > 0$, $\boldsymbol{\Gamma} = c \cdot \boldsymbol{\Pi}$. Let's call $\boldsymbol{\Pi}\mathbf{M}$ the *canonical Markovian network* of transition matrix $\mathbf{M}$. For a general Markov chain, we have:

$$\mathbf{1}\boldsymbol{\Pi}\mathbf{M} = \boldsymbol{\pi}^T \text{ and } \boldsymbol{\Pi}\mathbf{M}\mathbf{1} = \boldsymbol{\pi} \tag{32}$$

Thus, although canonical Markovian networks are usually directed, their nodes always have the same in-degree and out-degree. Such graphs are also known as the weighted Eulerian graphs.

4.2.2 PageRank Connection

Recall that Theorem 3.5 features the derivation of PageRank-conforming Markov chains from weighted networks. In fact, Theorem 3.5 and its PageRank power series can be naturally extended to any transition matrix $\mathbf{M}$: For any finite irreducible and ergodic Markov chain $\mathbf{M}$ and restart constant $\alpha > 0$, the matrix $\alpha \sum_{k=0}^{\infty}(1-\alpha)^k \cdot \mathbf{M}^k$ is a stochastic matrix that preserves the detailed-balancedness, the stationary distribution, and the spectra of $\mathbf{M}$.

Let's call $\alpha \sum_{k=0}^{\infty}(1-\alpha)^k \cdot \boldsymbol{\Pi}\mathbf{M}^k$ the *canonical PageRank-Markovian network* of transition matrix $\mathbf{M}$.

Proposition 4.7 *For any Markov chain* $\mathbf{M}$*, the random-walk Markov chain of the canonical PageRank-Markovian network* $\alpha \sum_{k=0}^{\infty}(1-\alpha)^k \cdot \boldsymbol{\Pi}\mathbf{M}^k$ *is conforming with respect to the PageRank of the canonical Markovian network* $\boldsymbol{\Pi}\mathbf{M}$.

4.2.3 Symmetrization

Algorithmically, computational/optimization problems on directed graphs are usually harder than they are on undirected graphs. For example, many recent breakthroughs in scalable graph-algorithm design are for limited to undirected graphs [9, 27, 53, 54, 59, 75, 83, 85–87]. To express Markov chains as undirect networks, we can apply the following well-known Markavian symmetrization formulation. Recall a matrix $\mathbf{L}$ is a *Laplacian matrix* if (1) $\mathbf{L}$ is a symmetric matrix with non-positive off-diagonal entries, and (2) $\mathbf{L} \cdot \mathbf{1} = \mathbf{0}$.

Proposition 4.8 (Canonical Markovian Symmetrization) *For any irreducible and ergodic finite Markov chain* $\mathbf{M}$*:*

$$\boldsymbol{\Pi} - \frac{\boldsymbol{\Pi}\mathbf{M} + \mathbf{M}^T\boldsymbol{\Pi}}{2} \tag{33}$$

is a Laplacian matrix, where $\boldsymbol{\Pi}$ *the diagonal matrix associated with* $\mathbf{M}$*'s stationary distribution. Therefore,* $\frac{\boldsymbol{\Pi}\mathbf{M}+\mathbf{M}^T\boldsymbol{\Pi}}{2}$ *is a symmetric network, whose degrees are normalized to stationary distribution* $\boldsymbol{\pi} = \boldsymbol{\Pi} \cdot \mathbf{1}$. *When* $\mathbf{M}$ *is detailed balanced,* $\frac{\boldsymbol{\Pi}\mathbf{M}+\mathbf{M}^T\boldsymbol{\Pi}}{2}$ *is the* canonical Markovian network *of* $\mathbf{M}$.

Proof We include a proof here for completeness. Let $\boldsymbol{\pi}$ be the stationary distribution of $\mathbf{M}$. Then:

$$\mathbf{M}^T\boldsymbol{\pi} = \boldsymbol{\pi}$$

$$\boldsymbol{\Pi} \cdot \mathbf{1} = \boldsymbol{\pi}$$

$$\mathbf{M} \cdot \mathbf{1} = \mathbf{1}$$

Therefore:

$$\left(\boldsymbol{\Pi} - \frac{\boldsymbol{\Pi}\mathbf{M} + \mathbf{M}^T\boldsymbol{\Pi}}{2}\right) \cdot \mathbf{1} = \left(\boldsymbol{\pi} - \frac{\boldsymbol{\Pi}\mathbf{1} + \mathbf{M}^T\boldsymbol{\pi}}{2}\right) = \mathbf{0} \tag{34}$$

The Lemma then follows from the fact that $\frac{1}{2}(\boldsymbol{\Pi}\mathbf{M} + \mathbf{M}^T\boldsymbol{\Pi})$ is symmetric and non-negative. □

Through the PageRank connection, Markov chains also have two extended Markovian symmetrizations:

Proposition 4.9 (PageRank Markovian Symmetrization) *For any irreducible and ergodic finite Markov chain* $\mathbf{M}$ *and restart constant* $\alpha > 0$, *the two matrices below:*

$$\boldsymbol{\Pi} - \alpha \sum_{k=0}^{\infty} (1-\alpha)^k \cdot \frac{\boldsymbol{\Pi}\mathbf{M}^k + (\mathbf{M}^T)^k\boldsymbol{\Pi}}{2} \tag{35}$$

$$\boldsymbol{\Pi} - \alpha \sum_{k=0}^{\infty} (1-\alpha)^k \boldsymbol{\Pi} \cdot \left(\boldsymbol{\Pi}^{-1} \cdot \frac{\boldsymbol{\Pi}\mathbf{M} + \mathbf{M}^T\boldsymbol{\Pi}}{2}\right)^k \tag{36}$$

are both Laplacian matrices. Moreover, the second Laplacian matrix is $\frac{1}{\alpha}$*-spectrally similar to* $(1-\alpha) \cdot \left(\boldsymbol{\Pi} - \frac{\boldsymbol{\Pi}\mathbf{M}+\mathbf{M}^T\boldsymbol{\Pi}}{2}\right)$.

4.2.4 Network Interpretations

We now return to Balcan et al.'s approach [11] for deriving preference networks from affinity networks. Consider the following natural extension of linear orders to express rankings with ties: An *ordered partition* of V is a total order of a partition of V. Let $\overline{L(V)}$ denote the set of all *ordered partitions* of V: For a $\sigma \in \overline{L(V)}$, for $i, j \in V$, we i is *ranked strictly ahead of* j if i and j belong to different partitions, and the partition containing i is ahead of the partition containing j in σ. If i and j are members of the same partition in σ, we say σ is *indifferent of* i and j.

Definition 4.10 (PageRank Preferences) Suppose $G = (V, E, \mathbf{W})$ is a weighted graph and $\alpha > 0$ is a restart constant. For each $u \in V$, let $\boldsymbol{\pi}_u$ be the ordered partition according to the descending ranking of V based on the personalized PageRank vector $\mathbf{p}_u = \mathbf{PPR}_{\mathbf{W},\alpha}[u, :]$. We call $\Pi_{\mathbf{W},\alpha} = \{\boldsymbol{\pi}_u\}_{u \in V}$ the *PageRank preference profile of* V with respect to G, and $A_{\mathbf{W},\alpha} = (V, \Pi_{\mathbf{W},\alpha})$ the *PageRank preference network* of G.

As pointed out in [11], other methods for deriving preference networks from weighted networks exist. For example, one can obtain individual preference rankings by ordering nodes according to shortest path distances, effective resistances, or maximum-flow/minimum-cut values.

> Is the PageRank preference a desirable personalized-preference profile of an affinity network?

This is a basic question in network analysis. In fact, much work has been done. I will refer readers to the beautiful axiomatic approach of Altman and Tennenholtz for characterizing personalized ranking systems [6]. Although they mostly studied unweighted networks, many of their results can be extended to weighted networks. Below, I will use Theorem 3.5 to address the following question that I was asked when first giving a talk about PageRank preferences.

> By taking the ranking information from PageRank matrices — which is usually asymmetric — one may lose valuable network information. For example, when $G = (V, E, \mathbf{W})$ is a undirected network, isn't it desirable to define ranking information according to a symmetric matrix?

At the time, I was not prepared to answer this question and replied that it was an excellent point. Theorem 3.5 now provides an answer. Markov chain theory uses an elegant concept to characterize whether or not a Markov chain $\mathbf{M}$ has an undirected network realization. Although Markov-chain transition matrices are usually asymmetric, if a Markov chain is detailed-balanced, then its transition matrix $\mathbf{M}$ can be diagonally scaled into a symmetric matrix by its stationary distribution. Moreover, $\boldsymbol{\Pi}\mathbf{M}$ is the "unique" underlying undirected network associated with $\mathbf{M}$. By Theorem 3.5, $\mathbf{PPR}_{\mathbf{W},\alpha}$ is a Markov transition matrix with stationary distribution $\mathbf{D}_{\mathbf{W}}$, and thus, $\overline{\mathbf{W}}_\alpha = \mathbf{D}_{\mathbf{W}} \cdot \mathbf{PPR}_{\mathbf{W},\alpha}$ is symmetric if and only if $\mathbf{W}$ is symmetric. Therefore, because the ranking given by $\mathbf{p}_u$ is the same as the ranking given by $\overline{\mathbf{W}}[u, :]$, the PageRank preference profile is indeed derived from a symmetric matrix when $\mathbf{W}$ is symmetric.

We can also define clusterability and other network models based on personalized PageRank matrices. For example:

- **PageRank conductance**:

$$\text{PageRank-conductance}_{\mathbf{W}}(S) := \frac{\sum_{u\in S, v\notin S} \overline{\mathbf{W}}[u,v]}{\min\left(\sum_{u\in S, v\in V} \overline{\mathbf{W}}[u,v], \sum_{u\notin S, v\in V} \overline{\mathbf{W}}[u,v]\right)} \tag{37}$$

- **PageRank utility**:

$$\text{PageRank-utility}_{\mathbf{W}}(S) := \sum_{u\in S, v\in S} \mathbf{PPR}_{\mathbf{W},\alpha}[u,v] \tag{38}$$

- **PageRank clusterability**:

$$\text{PageRank-clusterability}_{\mathbf{W}}(S) := \frac{\text{PageRank-utility}_{\mathbf{W}}(S)}{|S|} \tag{39}$$

Each of these functions defines a cooperative network based on $G = (V, E, \mathbf{W})$. These formulations are connected with the PageRank of G. For example, the Shapley value of the cooperative network given by $\boldsymbol{\tau} = \text{PageRank-utility}_{\mathbf{W}}$ is the PageRank of G.

$\mathbf{PPR}_{\mathbf{W},\alpha}$ can also be used to define incentive and powerset network models. The former can be defined by $u_s(T) = \sum_{v\in T} \mathbf{PPR}_{\mathbf{W},\alpha}[s,v]$, for $s \in V, T \subset V$ and $s \notin T$. The latter can be defined by $\boldsymbol{\theta}_{\mathbf{W}}(S,T) = \frac{\sum_{u\in S, v\in T} \mathbf{PPR}_{\mathbf{W},\alpha}[u,v]}{|S|}$ for $S, T \subseteq V$. $\boldsymbol{\theta}_{\mathbf{W}}(S,T)$ measures the *rate* of PageRank contribution from S to T.

4.3 Multifaceted Approaches to Network Analysis: Some Basic Questions

We will now conclude this section with a few basic questions, aiming to study how structural concepts in one network model can inspire structural concepts in other network models. A broad view of network data will enable us to comprehensively examine different facets of network data, as each network model brings out different aspects of network data. For examples, the metric model is based on geometry, the preference model is inspired by social-choice theory [10], the incentive and cooperative models are based on game-theoretical and economical principles [69, 70, 82], the powerset model is motivated by social influences [32, 55, 78], while the graphon [18] is based on graph limits and statistical modeling. We hope that addressing questions below will help us to gain comprehensive and comparative understanding of these models and the network structures/aspects that these models may reveal. We believe that multifaceted and multimodal approaches to network analysis will become increasingly more essential for studying major subjects in network science.

- How should we formulate *personalized centrality measures* with respect to other commonly-used network centrality measures [1, 13–17, 33, 36, 37, 41, 42, 51, 66, 71, 76, 80]? Can they be used to define meaningful centrality-conforming Markov chains?
- How should we define centrality measures and personalized ranking systems for general incentive or powerset networks? How should we define personalized Shapley value for cooperative games? How should we define weighted networks from cooperative/incentive/powerset models?
- What are natural Markov chains associated with the probabilistic graphical models [58]? How should we define centrality and clusterability for this important class of network models that are central to statistical machine learning?
- What constitutes a community in a probabilistic graphical model? What constitutes a community in a cooperative, incentive, preference, and powerset network? How should we capture network similarity in these models? How should we integrate them if they represents different facets of network data?
- How should we evaluate different clusterability measures and their usefulness to community identification or clustering? For example, PageRank conductance and PageRank clusterability are two different subset functions, but the latter applies to directed networks. How should we define clusterability-conforming centrality or centrality-forming clusterability?
- What are limitations of Markovian worldview of various network models? What are other unified worldview models for multifaceted network data?
- What is the fundamental difference between "directed" and "undirected" networks in various models?
- How should we model networks with non-homogeneous nodes and edge types?

More broadly, the objective is to build a systematic algorithmic framework for understanding multifaceted network data, particular given that many natural network models are highly theoretical in that their complete-information profiles have exponential dimensionality in $|V|$. In practice, they must be succinctly defined. The algorithmic network framework consists of the complex and challenging tasks of integrating sparse and succinctly-represented multifaceted network data $N = (V, F_1, \ldots, F_k)$ into an effective worldview (V, W) based on which, one can effectively build succinctly-represented underlying models for network facets, analyzing the interplay between network facets, and identify network solutions that are consistent with the comprehensive network data/models. What is a general model for specifying multifaceted network data? How should we formulate the problem of *network composition* for multifaceted network data?

5 To Jirka

The sparsity, richness, and ubiquitousness of multifaceted networks data make them wonderful subjects for mathematical and algorithmic studies. Network science has truly become a "universal discipline," with its multidisciplinary roots and interdisciplinary presence. However, it is a fundamental and conceptually challenging task to understand network data, due to the vast network phenomena.

> The holy grail of network science is to understand the network essence that underlies the observed sparse-and-multifaceted network data.

We need an analog of the concept of range space, which provides a united worldview of a family of diverse problems that are fundamental in statistical machine learning, geometric approximation, and data analysis. I wish that I had a chance to discuss with you about the mathematics of networks—beyond just the geometry of graphs—and to learn from your brilliant insights into *the essence of networks*. You and your mathematical depth and clarity will be greatly missed, Jirka.

References

1. K.V. Aadithya, B. Ravindran, T. Michalak, N. Jennings, Efficient computation of the Shapley value for centrality in networks, in *Internet and Network Economics*. Volume 6484 of Lecture Notes in Computer Science (Springer, Berlin/Heidelberg, 2010), pp. 1–13
2. E. Abbe, C. Sandon, Recovering communities in the general stochastic block model without knowing the parameters. CoRR, abs/1506.03729 (2015)
3. P.K. Agarwal, S. Har-Peled, K.R. Varadarajan, Geometric approximation via coresets, in *Combinatorial and Computational Geometry, MSRI* (Cambridge University Press, Cambridge, 2005), pp. 1–30
4. E.M. Airoldi, T.B. Costa, S.H. Chan, Stochastic blockmodel approximation of a graphon: theory and consistent estimation, in *27th Annual Conference on Neural Information Processing Systems 2013* (2013), pp. 692–700
5. N. Alon, V. Asodi, C. Cantor, S. Kasif, J. Rachlin, Multi-node graphs: a framework for multiplexed biological assays. J. Comput. Biol. **13**(10), 1659–1672 (2006)
6. A. Altman, M. Tennenholtz, An axiomatic approach to personalized ranking systems. J. ACM **57**(4), 26:1–26:35 (2010)
7. N. Amenta, M. Bern, D. Eppstein, S.-H. Teng, Regression depth and center points. Discret. Comput. Geom. **23**(3), 305–323 (2000)
8. R. Andersen, C. Borgs, J. Chayes, J. Hopcraft, V.S. Mirrokni, S.-H. Teng, Local computation of PageRank contributions, in *Proceedings of the 5th International Conference on Algorithms and Models for the Web-Graph*, WAW'07 (Springer, 2007), pp. 150–165
9. R. Andersen, F. Chung, K. Lang, Using PageRank to locally partition a graph. Internet Math. **4**(1), 1–128 (2007)
10. K.J. Arrow, *Social Choice and Individual Values*, 2nd edn. (Wiley, New York, 1963)
11. M.F. Balcan, C. Borgs, M. Braverman, J.T. Chayes, S.-H. Teng, Finding endogenously formed communities, in *SODA* (2013), pp. 767–783
12. J. Batson, D.A. Spielman, N. Srivastava, S.-H. Teng, Spectral sparsification of graphs: theory and algorithms. Commun. ACM **56**(8), 87–94 (2013)

13. A. Bavelas, Communication patterns in task oriented groups. J. Acoust. Soc. Am. **22**(6), 725–730 (1950)
14. P. Bonacich, Power and centrality: a family of measures. Am. J. Soc. **92**(5), 1170–1182 (1987)
15. P. Bonacich, Simultaneous group and individual centralities. Soc. Netw. **13**(2), 155–168 (1991)
16. S.P. Borgatti, Centrality and network flow. Soc. Netw. **27**(1), 55–71 (2005)
17. S.P. Borgatti, M.G. Everett, A graph-theoretic perspective on centrality. Soc. Netw. **28**(4), 466–484 (2006)
18. C. Borgs, J. Chayes, L. Lovász, V.T. Sós, B. Szegedy, K. Vesztergombi, Graph limits and parameter testing, in *Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of Computing*, STOC'06 (2006), pp. 261–270
19. C. Borgs, J.T. Chayes, A. Marple, S. Teng, An axiomatic approach to community detection, in *Proceedings of the ACM Conference on Innovations in Theoretical Computer Science*, ITCS'16 (2016), pp. 135–146
20. C. Borgs, J.T. Chayes, A.D. Smith, Private graphon estimation for sparse graphs, in *Annual Conference on Neural Information Processing Systems* (2015), pp. 1369–1377
21. S.J. Brams, M.A. Jones, D.M. Kilgour, Dynamic models of coalition formation: Fallback vs. build-up, in *Proceedings of the 9th Conference on Theoretical Aspects of Rationality and Knowledge*, TARK'03 (2003), pp. 187–200
22. S. Brin, L. Page, The anatomy of a large-scale hypertextual Web search engine. Comput. Netw. **30**(1–7), 107–117 (1998)
23. M. Caesar, J. Rexford, BGP routing policies in ISP networks. Netw. Mag. Global Internetwkg. **19**(6), 5–11 (2005)
24. J. Cheeger, A lower bound for the smallest eigenvalue of the Laplacian, in *Problems in Analysis*, ed by R.C. Gunning (Princeton University Press, Princeton, 1970), pp. 195–199
25. W. Chen, S.-H. Teng, Interplay between social influence and network centrality: a comparative study on Shapley centrality and single-node-influence centrality, in *Proceedings of the 26th International Conference on World Wide Web, WWW*, Perth (ACM, 2017), pp. 967–976
26. X. Chen, X. Deng, S.-H. Teng, Settling the complexity of computing two-player nash equilibria. J. ACM **56**(3), 14:1–14:57 (2009)
27. D. Cheng, Y. Cheng, Y. Liu, R. Peng, S.-H. Teng, Efficient sampling for Gaussian graphical models via spectral sparsification, in *Proceedings of the 28th Conference on Learning Theory*, COLT'05 (2015)
28. F.R.K. Chung, *Spectral Graph Theory (CBMS Regional Conference Series in Mathematics, No. 92)* (American Mathematical Society, 1997)
29. K. Clarkson, D. Eppstein, G.L. Miller, C. Sturtivant, S.-H. Teng, Approximating center points with and without linear programming, in *Proceedings of 9th ACM Symposium on Computational Geometry* (1993), pp. 91–98
30. L. Danzer, J. Fonlupt, V. Klee, Helly's theorem and its relatives. Proc. Symp. Pure Math. Am. Math. Soc. **7**, 101–180 (1963)
31. C. Daskalakis, P.W. Goldberg, C.H. Papadimitriou, The complexity of computing a nash equilibrium. SIAM J. Comput. **39**(1), 195–259 (2009)
32. P. Domingos, M. Richardson, Mining the network value of customers, in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'01 (2001), pp. 57–66, CoRR, abs/cond-mat/0502230
33. L. Donetti, P.I. Hurtado, M.A. Munoz, Entangled networks, synchronization, and optimal network topology (2005)
34. H. Edelsbrunner, *Algorithms in Combinatorial Geometry* (Springer, New York, 1987)
35. H. Eulau, The columbia studies of personal influence: social network analysis. Soc. Sci. Hist. **4**(02), 207–228 (1980)
36. M.G. Everett, S.P. Borgatti, The centrality of groups and classes. J. Math. Soc. **23**(3), 181–201 (1999)
37. K. Faust, Centrality in affiliation networks. Soc. Netw. **19**(2), 157–191 (1997)

38. D. Feldman, M. Langberg, A unified framework for approximating and clustering data, in *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, STOC'11 (2011), pp. 569–578
39. M. Fiedler, Algebraic connectivity of graphs. Czechoslov. Math. J. **23**(2), 298–305 (1973)
40. M. Fiedler, A property of eigenvectors of nonnegative symmetric matrices and its applications to graph theory. Czechoslov. Math. J. **25**(100), 619–633 (1975)
41. L.C. Freeman, A set of measures of centrality based upon betweenness. Sociometry **40**, 35–41 (1977)
42. L.C. Freeman, Centrality in social networks: conceptual clarification. Soc. Netw. **1**(3), 215–239 (1979)
43. D. Gale, L.S. Shapley, College admissions and the stability of marriage. Am. Math. Mon. **69**(1), 9–15 (1962)
44. R. Ghosh, S.-H. Teng, K. Lerman, X. Yan, The interplay between dynamics and networks: centrality, communities, and Cheeger inequality, in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'14 (2014), pp. 1406–1415
45. J.R. Gilbert, G.L. Miller, S.-H. Teng, Geometric mesh partitioning: implementation and experiments. SIAM J. Sci. Comput. **19**(6), 2091–2110 (1998)
46. D. Gusfield, R.W. Irving, *The Stable Marriage Problem: Structure and Algorithms* (MIT Press, Cambridge, 1989)
47. Q. Han, K.S. Xu, E.M. Airoldi, Consistent estimation of dynamic and multi-layer block models. CoRR, abs/1410.8597 (2015)
48. S. Hanneke, E.P. Xing, Network completion and survey sampling, in *AISTATS*, ed. by D.A.V. Dyk, M. Welling. Volume 5 of JMLR Proceedings (2009), pp. 209–215
49. T. Haveliwala, Topic-sensitive Pagerank: a context-sensitive ranking algorithm for web search. Trans. Knowl. Data Eng. **15**(4), 784–796 (2003)
50. M. Hubert, P.J. Rousseeuw, The catline for deep regression. J. Multivar. Anal. **66**, 270–296 (1998)
51. L. Katz, A new status index derived from sociometric analysis. Psychometrika **18**(1), 39–43 (1953)
52. M.J. Kearns, M.L. Littman, S.P. Singh, Graphical models for game theory, in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI'01 (2001), pp. 253–260
53. J.A. Kelner, Y.T. Lee, L. Orecchia, A. Sidford, An almost-linear-time algorithm for approximate max flow in undirected graphs, and its multicommodity generalizations, in *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA'14 (2014), pp. 217–226
54. J.A. Kelner, L. Orecchia, A. Sidford, Z.A. Zhu, A simple, combinatorial algorithm for solving SDD systems in nearly-linear time, in *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, STOC'13 (2013), pp. 911–920
55. D. Kempe, J. Kleinberg, E. Tardos, Maximizing the spread of influence through a social network, in *KDD'03* (ACM, 2003), pp. 137–146
56. M. Kim, J. Leskovec, The network completion problem: inferring missing nodes and edges in networks, in *SDM* (SIAM/Omnipress, 2011), pp. 47–58
57. M. Kivelä, A. Arenas, M. Barthelemy, J.P. Gleeson, Y. Moreno, M.A. Porter, Multilayer networks. CoRR, abs/1309.7233 (2014)
58. D. Koller, N. Friedman, *Probabilistic Graphical Models: Principles and Techniques – Adaptive Computation and Machine Learning* (The MIT Press, Cambridge, 2009)
59. I. Koutis, G. Miller, R. Peng, A nearly-mlogn time solver for SDD linear systems, in *2011 52nd Annual IEEE Symposium on Foundations of Computer Science (FOCS)* (2011), pp. 590–598
60. K. Lerman, S. Teng, X. Yan, Network composition from multi-layer data. CoRR, abs/1609.01641 (2016)
61. N. Linial, E. London, Y. Rabinovich, The geometry of graphs and some of its algorithmic applications. Combinatorica **15**(2), 215–245 (1995)

62. L. Lovász, M. Simonovits, Random walks in a convex body and an improved volume algorithm. RSA: Random Struct. Algorithms **4**, 359–412 (1993)
63. F. Masrour, I. Barjesteh, R. Forsati, A.-H. Esfahanian, H. Radha, Network completion with node similarity: a matrix completion approach with provable guarantees, in *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM'15 (ACM, 2015), pp. 302–307
64. J. Matoušek, Approximations and optimal geometric divide-and-conquer, in *Proceedings of the Twenty-Third Annual ACM Symposium on Theory of Computing*, STOC'91 (1991), pp. 505–511
65. J. Matoušek, M. Sharir, E. Welzl, A subexponential bound for linear programming, in *Proceedings of the Eighth Annual Symposium on Computational Geometry*, SCG'92 (1992), pp. 1–8
66. T.P. Michalak, K.V. Aadithya, P.L. Szczepanski, B. Ravindran, N.R. Jennings, Efficient computation of the Shapley value for game-theoretic network centrality. J. Artif. Int. Res. **46**(1), 607–650 (2013)
67. G.L. Miller, S.-H. Teng, W. Thurston, S.A. Vavasis, Separators for sphere-packings and nearest neighbor graphs. J. ACM **44**(1), 1–29 (1997)
68. G.L. Miller, S.-H. Teng, W. Thurston, S.A. Vavasis, Geometric separators for finite-element meshes. SIAM J. Sci. Comput. **19**(2), 364–386 (1998)
69. J. Nash, Equilibrium points in n-person games. Proc. Natl. Acad. USA **36**(1), 48–49 (1950)
70. J. Nash, Noncooperative games. Ann. Math. **54**, 289–295 (1951)
71. M. Newman, *Networks: An Introduction* (Oxford University Press, Inc., New York, 2010)
72. L. Page, S. Brin, R. Motwani, T. Winograd., The Pagerank citation ranking: bringing order to the Web, in *Proceedings of the 7th International World Wide Web Conference* (1998), pp. 161–172
73. S. Paul, Y. Chen, Community detection in multi-relational data with restricted multi-layer stochastic blockmodel. CoRR, abs/1506.02699v2 (2016)
74. L. Peel, D.B. Larremore, A. Clauset, The ground truth about metadata and community detection in networks. CoRR, abs/1608.05878, 2016
75. R. Peng, Approximate undirected maximum flows in O(mpolylog(n)) time, in *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA'16 (2016), pp. 1862–1867
76. M. Piraveenan, M. Prokopenko, L. Hossain, Percolation centrality: quantifying graph-theoretic impact of nodes during percolation in networks. PLoS ONE **8**(1) (2013)
77. Y. Rekhter, T. Li, A Border Gateway Protocol 4. *IETF RFC 1771* (1995)
78. M. Richardson, P. Domingos, Mining knowledge-sharing sites for viral marketing, in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'02 (2002), pp. 61–70
79. A.E. Roth, The evolution of the labor market for medical interns and residents: A case study in game theory. J. Politi. Econ. **92**, 991–1016 (1984)
80. G. Sabidussi, The centrality index of a graph. Psychometirka **31**, 581–606 (1996)
81. L.S. Shapley, A value for n-person games, in *Contributions to the Theory of Games II*, ed. by H. Kuhn, A.W. Tucker (Princeton University Press, Princeton, 1953), pp. 307–317
82. L.S. Shapley, Cores of convex games. Int. J. Game Theory **1**(1), 11–26 (1971)
83. J. Sherman, Nearly maximum flows in nearly linear time, in *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, FOCS'13 (2013), pp. 263–269
84. D.A. Spielman, S.-H. Teng, Spectral partitioning works: planar graphs and finite element meshes. Linear Algebra Appl. **421**(2–3), 284–305 (2007)
85. D.A. Spielman, S.-H. Teng, Spectral sparsification of graphs. SIAM J. Comput. **40**(4), 981–1025 (2011)
86. D.A. Spielman, S.-H. Teng, A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. SIAM J. Comput. **42**(1), 1–26 (2013)

87. D.A. Spielman, S.-H. Teng, Nearly-linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. SIAM J. Matrix Anal. Appl. **35**(3), 835–885 (2014)
88. S.-H. Teng, Points, Spheres, and Separators: A Unified Geometric Approach to Graph Partitioning. PhD thesis, Advisor: Gary Miller, Carnegie Mellon University, Pittsburgh, 1991
89. S.-H. Teng, Scalable algorithms for data and network analysis. Found. Trends Theor. Comput. Sci. **12**(1–2), 1–261 (2016)
90. H. Tverberg, A generalization of Radon's theorem. J. Lond. Math Soc. **41**, 123–128 (1966)
91. V.N. Vapnik, *The Nature of Statistical Learning Theory* (Springer, New York, 1995)
92. V.N. Vapnik, A.Y. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities. Theory Probab. Appl. **16**, 264–280 (1971)
93. H.P. Young, An axiomatization of Borda's rule. J. Econ. Theory **9**(1), 43–52 (1974)
94. A.Y. Zhang, H.H. Zhou, Minimax rates of community detection in stochastic block models. CoRR, abs/1507.05313 (2015)

Anti-concentration Inequalities for Polynomials

Van Vu

In memory of Jirka Matoušek

Abstract In this short survey, we discuss the notion of anti-concentration and describe various ideas used to obtain anti-concentration inequalities, together with several open questions.

1 What is Anti-concentration?

Many arguments in probabilistic combinatorics (and probability in general) rely on bounding the probability of rare events. A frequently used tool for such arguments is large deviation (or strong concentration) inequalities. Let X be a real random variable; a typical large deviation inequality asserts that (under certain assumptions) for any interval I located far from the mean of X, $\mathbf{P}(X \in I)$ is small. The length of I is not important, one usually takes I to be a half-line. For many results of this type, we refer to [10].

Anti-concentration is a phenomenon in the opposite direction. A typical anti-concentration asserts that if an interval I has small length, then $\mathbf{P}(X \in I)$ is small, *regardless the location of I*. Inequalities of this type have recently found powerful applications in many branches of probability, most notably the theory of random matrices (see [19] for an example). It is obvious to extend the notion of anti-concentration to more abstract spaces.

V. Vu is supported by NSF grant DMS-1307797 and AFORS grant FA9550-12-1-0083.

V. Vu (✉)
Department of Mathematics, Yale University, 06520 New Haven, CT, USA
e-mail: van.vu@yale.edu

M. Loebl et al. (eds.), *A Journey Through Discrete Mathematics*,
DOI 10.1007/978-3-319-44479-6_32

In what follows, we use the central limit theorem to illustrate both phenomena. Let $\xi_i, i \geq 1$ be iid random variables with mean 0 and variance 1. Then the CLT asserts that

$$X := \frac{\xi_1 + \cdots + \xi_n}{\sqrt{n}} \to N(0,1).$$

In other words, for all fixed T,

$$P(X \geq T) = \frac{1}{\sqrt{2\pi}} \int_T^\infty e^{-t^2/2} dt + o(1).$$

For sufficiently large T, $\frac{1}{\sqrt{2\pi}} \int_T^\infty e^{-t^2/2} dt \leq \exp(-T^2/2)$, so with $I := (T, \infty)$, we obtain a large deviation bound

$$\mathbf{P}(X \in I) \leq \exp(-T^2/2).$$

On the other hand, if we assume further that $\mathbf{E}|\xi_i|^3$ is bounded, then Berry–Esseen theorem asserts that for any interval I

$$|\mathbf{P}(X \in I) - \mathbf{P}(N(0,1) \in I)| = O(n^{-1/2}).$$

If I has length $\Omega(n^{-1/2})$, then $\mathbf{P}(N(0,1) \in I) = O(n^{-1/2})$. In this case, we have an anti-concentration bound

$$\mathbf{P}(X \in I) = O(n^{-1/2}), \tag{1}$$

regardless the location of I.

In the rest of this article, we present several anti-concentration inequalities for functions which can be represented as polynomials in term of iid variables $\xi_1, \ldots, \xi_n$. We make an effort to describe the main ideas behind the proofs as we believe those are of independent interest.

2 Anti-concentration for Linear Forms

Let us consider the linear form $L := a_1\xi_1 + \cdots + a_n\xi_n$, where a_i are real numbers and ξ_i are iid Rademacher random variables. In 1943, Littlewood and Offord [12] discovered the first anti-concentration result

Theorem 2.1 *There is a constant B such that the following holds for all n. If all coefficients a_i have absolute value at least 1, then for any open interval I of length 1,*

$$\mathbf{P}(L \in I) \leq B n^{-1/2} \log n.$$

Notice that in the case $a_i = 1$, this is almost (1), after a proper scaling. However, for arbitrary a_i, we do not even have a CLT for L, let alone the stronger result of Berry–Esseen.

The term $\log n$ was removed later by Erdős, who proved the following sharp form under the same condition

Theorem 2.2

$$\mathbf{P}(L \in I) \leq \frac{\binom{n}{\lfloor n/2 \rfloor}}{2^n} = O(n^{-1/2}).$$

Assuming that n is even, it is easy to see that for $a_i = 1$ (for all i), $\mathbf{P}(L = 0) = \frac{\binom{n}{\lfloor n/2 \rfloor}}{2^n}$, showing the sharpness of the result. Notice that this theorem implies that if all $a_i \neq 0$, then for any number x

$$\mathbf{P}(L = x) \leq \frac{\binom{n}{\lfloor n/2 \rfloor}}{2^n} = O(n^{-1/2}). \tag{2}$$

Erdős' proof uses a lovely combinatorial argument [5]. By symmetry, one can assume that all a_i are positive. For each instance of the variables ξ_i such that $L \in I$, let A be the set of indices i where $\xi_i = 1$. It is easy to see that the collection of those sets A form an anti-chain, namely no two A's are proper subset of each other. Sperner's lemma, which asserts that the size of an anti-chain is at most $\binom{n}{\lfloor n/2 \rfloor}$ concludes the proof. Theorems 2.1 and 2.2 are the starting points of a massive study in combinatorics and probability which goes through many decades; see [16] for a survey.

3 Esseen's Inequality

In the 1960s, Esseen [6] proved the following general anti-concentration inequality

Lemma 3.1 *For any fixed d there exists an absolute positive constant $C = C(d)$ such that for any random variable X with support in $\mathbb{R}^d$ and any unit ball $\mathbf{B} \subset \mathbb{R}^d$*

$$\mathbf{P}(X \in \mathbf{B}) \leq C \int_{\|t\|_2 \leq 1} |\mathbf{E}(\exp(i\langle t, X\rangle))| \, dt. \tag{3}$$

This lemma leads to a systematic approach to prove new anti-concentration inequalities; see [7, 16] for many examples. Let us use it to give another proof of Erdős bound $O(n^{-1/2})$ in Theorem 2.2, which is entirely different from the combinatorial one using Kneser's lemma. In view of Lemma 3.1, it suffices to show that

$$\int_{|t| \leq 1} |\mathbf{E}(\exp(it \sum_{j=1}^{n} a_j \xi_j)|) \, dt = O(1/\sqrt{n}).$$

By the independence of the ξ_j, we have

$$\left|\mathbf{E}(\exp(it\sum_{j=1}^{n} a_j\xi_j))\right| = \prod_{j=1}^{n} |\mathbf{E}(\exp(ita_j\xi_j)| = \left|\prod_{j=1}^{n} \cos(ta_j)\right|.$$

By Hölder's inequality

$$\int_{|t|\le 1} \left|\mathbf{E}(\exp(it\sum_{j=1}^{n} a_j\xi_j))\right| \, dt \le \prod_{j=1}^{n} (\int_{|t|\le 1} |\cos(ta_j)|^n \, dt)^{1/n}.$$

But since each a_j has magnitude at least 1, it is easy to check that $\int_{|t|\le 1} |\cos(ta_j)|^n \, dt = O(1/\sqrt{n})$, and the claim follows.

The analytic approach via Esseen's lemma and ideas from Additive Combinatorics provide essential tools to the development of Inverse Littlewood–Offord theory; see [16] for a survey.

4 Anti-concentration for Quadratic Forms

Consider the quadratic form $Q := \sum_{1\le i,j\le n} a_{ij}\xi_i\xi_j$, where a_{ij} are real coefficients. Anti-concentration for quadratic forms was first studied by Costello et al. [3], as a step to the solution of Weiss' conjecture (that a random symmetric ±1 matrix typically has full rank). The leading idea in [3] was to reduce to the linear case, using the following *decoupling* lemma

Lemma 4.1 (Decoupling lemma) *Let Y and Z be independent random variables and $E = E(Y, Z)$ be an event depending on Y and Z. Then*

$$\mathbf{P}(E(Y,Z)) \le \mathbf{P}(E(Y,Z) \wedge E(Y',Z) \wedge E(Y,Z') \wedge E(Y',Z'))^{1/4}$$

where Y' and Z' are independent copies of Y and Z, respectively. Here we use $A \wedge B$ to denote the event that A and B both hold.

Let us show how to use this lemma to get a bound on $\mathbf{P}(Q \in I)$. For simplicity, we consider the discrete version, namely bounding $\mathbf{P}(Q = x)$ for any value x. Take U_1 to be the first half of the indices and U_2 to be the second half. Define $Y := (\xi_i)_{i\in U_1}$ and $Z := (\xi_i)_{i\in U_2}$. We can write $Q(x) = Q(Y, Z)$. Let ξ_i' be an independent copy of ξ_i and set $Y' := (\xi_i')_{i\in U_1}$ and $Z' := (\xi_i')_{i\in U_2})$. By Lemma 4.1, for any number x

$$\mathbf{P}(Q(Y,Z) = x) \le \mathbf{P}(Q(Y,Z) = Q(Y,Z') = Q(Y',Z) = Q(Y',Z') = x)^{1/4}.$$

On the other hand, if $Q(Y,Z) = Q(Y,Z') = Q(Y',Z) = Q(Y',Z') = x$ then regardless the value of x

$$R := Q(Y,Z) - Q(Y',Z) - Q(Y,Z') + Q(Y',Z') = 0.$$

Furthermore, we can write R as

$$R = \sum_{i\in U_1}\sum_{j\in U_2} a_{ij}(\xi_i - \xi_i')(\xi_j - \xi_j') = \sum_{i\in U_1} R_i w_i,$$

where w_i is the random variable $w_i := \xi_i - \xi_i'$, and R_i is the random variable $\sum_{j\in U_2} a_{ij}w_j$.

We now can conclude the proof by applying Theorem 2.2 (or more precisely (2)) twice. First, combining this theorem with a combinatorial argument, one can show that (with high probability) many R_i are non-zero. Next, one can condition on the non-zero R_i and apply Theorem 2.2 for the linear form $\sum_{i\in U_1} R_i w_i$ to obtain a bound on $\mathbf{P}(R = 0)$.

The proof of Lemma 4.1 is worth discussing, as it reveals a surprising connection to *extremal combinatorics*. Without loss of generality, we can assume that the probability space Ω is discrete, with uniform weight on atoms $v_1, \dots, v_m$, where m is very large compared to n. We build a bipartite graph G whose color classes are copies of Ω, by connecting (a copy of) v_i to (a copy) of v_j if (v_i, v_j) belongs to the support of $E(Y,Z)$. Thus, $\mathbf{P}(E(Y,Z))$ is exactly the edge density p of G. A similar consideration shows that $\mathbf{P}(E(Y,Z) \wedge E(Y',Z) \wedge E(Y,Z') \wedge E(Y',Z'))$ is the density p_{C_4} of C_4 (cycle of length 4). Thus, the claimed bound is equivalent to the well known fact that $p_{C_4} \ge p^4$. Bounds of this type are abundant in extremal combinatorics and perhaps one will find more applications of this type.

5 Anti-concentration for Polynomials: Iteration

The argument in the last section gives bound $n^{-1/8}$. Later, Costello [2] obtained the optimal bound $n^{-1/2+o(1)}$ for quadratic forms. One can use the decoupling idea repeatedly to obtain bounds for higher degree polynomials. We consider a multilinear polynomial P of degree d of the form

$$P := \sum_{S\subset\{1,\dots,n\};|S|=d} a_S \prod_{i\in S}\xi_i, \qquad (4)$$

where a_i are real coefficients. Costello et al. proved

Theorem 5.1 *There is a constant B such that the following holds for all d, n. If there are mn^{d-1} coefficients a_S with absolute value at least 1, then for any open interval I of length 1,*

$$\mathbf{P}(P(\xi_1,\dots,\xi_n) \in I) \le Bm^{-\frac{1}{2^{(d^2+d)/2}}}.$$

Notice that in each iteration, the degree goes down by 1, but one loses a 4th root in the probability bound. Thus, the exponential loss in the final result is expected. In [17], Razborov and Viola, motivated by an application in complexity theory, found a more efficient way to apply Lemma 4.1. They first introduced the following definition.

Definition 5.2 For a degree d multi-linear polynomial of the form (4), the *rank* of P, denoted by *rank*(P), is the largest integer r such that there exist disjoint sets $S_1, \dots, S_r \subseteq [n]$ of size d with $|a_{S_j}| \geq 1$, for $j \in [r]$.

Then they proved

Theorem 5.3 *There is a constant B such that the following holds for all d, n. If P has rank r, then for any open interval I of length 1,*

$$\mathbf{P}(P(\xi_1, \dots, \xi_n) \in I) \leq B r^{-\frac{1}{d 2^{d+1}}}.$$

In the next two sections, we discuss two ideas to improve upon this bound.

6 Anti-concentration for Polynomials: The Switching Method

The next improvement makes use of Lindeberg's switching idea. In the 1920s, Lindeberg [11] found a new method to prove the central limit theorem. Let us consider

$$X := \frac{\xi_1 + \cdots + \xi_n}{\sqrt{n}},$$

where ξ_i are iid random variables with mean 0 and variance 1 with bounded third moment.

Let $\tilde{\xi}_i$ be iid standard normal variables. In this special case, $\tilde{X} := \frac{\tilde{\xi}_1 + \cdots + \tilde{\xi}_n}{\sqrt{n}}$ itself is standard normal. To conclude the proof, we are going to show that X and $\tilde{X}$ have approximately the same distribution. It suffices to show that for any smooth function G with compact support

$$\mathbf{E}G(X) = \mathbf{E}G(\tilde{X}) + o(1).$$

We write $G(X) - G(\tilde{X}) = \sum_{i=1}^n F_i - \tilde{F}_i$, where

$$F_i = G\left(\frac{\tilde{\xi}_1 + \cdots + \tilde{\xi}_{i-1} + \xi_i + \cdots + \xi_n}{\sqrt{n}}\right)$$

and

$$\tilde{F}_i = G\left(\frac{\tilde{\xi}_1 + \cdots + \tilde{\xi}_{i-1} + \tilde{\xi}_i + \xi_{i+1} + \cdots + \xi_n}{\sqrt{n}}\right).$$

By conditioning, we view $F_i := F(\xi_i)$ as a function in ξ_i only and use Taylor expansion

$$F(\xi_i) = F(0) + \xi_i F'(0) + \frac{1}{2}\xi_i^2 F''(0) + \frac{1}{6}\xi_i^3 F'''(z_i),$$

where z_i is between 0 and ξ_i.

Do the same with $F(\tilde{\xi}_i)$, it follows that

$$\mathbf{E}F(\xi_i) - \mathbf{E}F(\tilde{\xi}_i) = F'(0)\mathbf{E}(\xi_i - \tilde{\xi}_i) + \frac{1}{2}F''(0)\mathbf{E}(\xi_i^2 - \tilde{\xi}_i^2) + \frac{1}{6}\mathbf{E}\Big(\xi_i^3 F'''(z_i) - \tilde{\xi}_i^3 F'''(\tilde{z}_i)\Big).$$

As ξ_i and $\tilde{\xi}_i$ have the same mean and variance, the first two terms cancel and one can bound the absolute value of RHS by

$$\frac{1}{6}\mathbf{E}(|\xi_i|^3 + |\tilde{\xi}_i|^3)K,$$

where $K := \sup_x |F'''(x)|$. Notice that the lth derivative of F involve a term $n^{-l/2}$ (coming from the normalization by $n^{-1/2}$). Asuming that ξ_i has bounded third moment, then $\frac{1}{6}\mathbf{E}(|\xi_i|^3 + |\tilde{\xi}_i|^3)K = O(n^{-3/2})$. On the other hand, we need to add only n terms and thus the final bound is $O(n^{-1/2}) = o(1)$, concluding the proof.

Lindeberg's idea is very flexible and can be repeated for high degree polynomials (and infact, for any smooth functions). Let P be a polynomial of degree d as above. It is not true that P satisfies the CLT, even when ξ_i are normal gaussian. However, Carbery and Wright [1] showed

Theorem 6.1 *There is a constant C such that for any interval I of length 1,*

$$\mathbf{P}(P \in I) \le Cd(\frac{1}{\sqrt{\mathbf{Var}P}})^{1/d}.$$

Now let us consider ξ_i being iid Rademacher variables. The *influence* of the i-th variable on P is defined to be $\mathbf{Inf}_i = \mathbf{Inf}_i(P) = \sum_{i\in S} a_S^2$. Since $\mathbf{Var}(P) = \sum_{S\neq\emptyset} a_S^2$, we have

$$\mathbf{Var}(P) \le \sum_{i=1}^{n} \mathbf{Inf}_i \le d\mathbf{Var}(P). \qquad (5)$$

Define $\tau := \frac{\max_{1\le i\le n} \mathbf{Inf}_i}{\sum_{i=1}^n \mathbf{Inf}_i}$. It is clear that $1/n \le \tau \le 1$. Using Lindeberg's method, Mossel et al. proved [14]

Theorem 6.2 *Let P be a non-constant polynomial of the form* (4)*. Then for any interval I of length 1,*

$$\mathbf{P}(P(\xi_1,\dots,\xi_n)\in I) \le \frac{Cd}{(\mathbf{Var}(P))^{1/2d}} + Cd\tau^{1/(4d+1)}.$$

The role of τ in the inequality is to replace the assumption of bounded third moment in the original argument of Lindeberg. For a detailed argument, we refer to [14].

7 Anti-concentration for Polynomials: Reducing τ by Conditioning

If τ is large, the bound in Theorem 6.2 loses its strength. In [13], we developed a method to deal with this case, following the proof of [4, Theorem 1.1] and also [8, 9]. The main idea is to condition on the random variables with large influence. Doing this properly, we obtain, with high probability, a polynomial which has small τ or is dominated by its constant part (the latter case is easy to deal with by an adhoc argument). Using this method, Meka et al. [13] improved Theorem 6.2 as follows

Theorem 7.1 *There is an absolute constant B such that the following holds for all d, n. Let P be a polynomial of the form* (4) *whose rank $r \ge 2$. Then for any interval I of length 1,*

$$\mathbf{P}(P(\xi_1,\dots,\xi_n)\in I) \le \min\left(\frac{Bd^{4/3}\sqrt{\log r}}{r^{\frac{1}{4d+1}}}, \frac{\exp(Bd^2(\log\log r)^2)}{\sqrt{r}}\right).$$

Notice that for the case $d = O(1)$ and $r = \Theta(n)$, the second term on the RHS dominates and is of order $n^{-1/2+o(1)}$. This bound is sharp, up to the $o(1)$ term. To see this, consider $\mathbf{P} = (\xi_1+\dots+\xi_n)^d$. In this case, $\mathbf{P}(P=0) = \mathbf{P}(\xi_1+\dots+\xi_n = 0)$ and the latter can be as large as $\Omega(n^{-1/2})$ (see Theorem 2.2). The first bound is better in the case d tends to infinity with n. (In particular, this bound was used to settle the problem of Razborov and Viola in complexity theory, mentioned earlier; see [13] for details.) There is another proof of Theorem 7.1 (by Kane), which directly links anti-concentration to the notion of average sensitivity and the Gotsman–Linial conjecture; see [13].

8 Remarks and Open Problems

An interesting question is to find an exact bound as in Theorem 2.2. Notice that if n is even and

$$P := \prod_{j=0}^{l} \left(\sum_{i=1}^{n} \xi_i \pm 2j \right)$$

then P is a polynomial of degree $d = 2l + 1$ such that

$$\mathbf{P}(P = 0) = \frac{S_d}{2^n}$$

where S_d is the sum of the largest d binomial coefficients. Could this be a upper bound for $\mathbf{P}(P = 0)$ for all polynomials P with all monomials having non-zero coefficients ? It would be already interesting to remove the $o(1)$ in the exponent of the bound $n^{-1/2+o(1)}$ discussed at the end of the last section.

In the linear case, a theory of Inverse Littlewood-Offord inequalities has been worked out by many researchers (see [16] for a survey). A typical result in this theory asserts that in Theorem 2.1, we can have a much better bound (say n^{-100}) unless the coefficients a_i satisfy certain strong additive properties. (In fact, one can more or less give a full characterization of all sets $\{a_1, \ldots, a_n\}$ such that $\mathbf{P}(L \in I) \geq n^{-100}$.) Results of this kind proved very useful in the studies of random matrices. leading to the solutions of various long standing problems (see [16, 18, 19]).

It is desirable to extend this theory for higher degree polynomials. Such extension will certainly has a large number of applications. Even for quadratic form, the situation is not absolutely clear; for partial results see [15].

References

1. A. Carbery, J. Wright, Distributional and L^q norm inequalities for polynomials over convex bodies in R^n. Math. Res. Lett. **8**(3), 233–248 (2001)
2. K. Costello, Bilinear and quadratic variants on the Littlewood-Offord problem. Israel J. Math. **194.1**, 359–394 (2013)
3. K. Costello, T. Tao, V. Vu, Random symmetric matrices are almost surely nonsingular. Duke Math. J. **135.2**, 395–413 (2006)
4. I. Diakonikolas, R. Servedio, L. Tan, A. Wan, A regularity lemma, and low-weight approximators, for low-degree polynomial threshold functions, in *2010 IEEE 25th Annual Conference on Computational Complexity (CCC)* (2010), pp. 211–222
5. P. Erdös, On a lemma of Littlewood and Offord. Bull. Am. Math. Soc. **51.12**, 898–902 (1945)
6. C. Esséen, On the Kolmogorov-Rogozin inequality for the concentration function. Z. Wahrsch. Verw. Gebiete **5**, 210–216 (1966)
7. G. Halász, Estimates for the concentration function of combinatorial number theory and probability. Period. Math. Hungar. **8**(3–4), 197–211 (1977)

8. D. Kane, The Correct Exponent for the Gotsman-Linial Conjecture, in *2013 IEEE 28th Conference on Computational Complexity (CCC)* (IEEE, 2013), pp. 56–64
9. D. Kane, A pseudorandom generator for polynomial threshold functions of gaussian with subpolynomial seed length, in *2014 IEEE 29th Conference on Computational Complexity (CCC)* (IEEE, 2014), pp. 217–228
10. M. Ledoux, *The Concentration of Measure Phenomenon*. Volume 89 of Mathematical Surveys and Mono-graphs. (American Mathematical Society, 2001)
11. J.W. Lindeberg, Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. Math. Z. **15**, 211–225 (1922)
12. J.E. Littlewood, A.C. Offord, On the number of real roots of a random algebraic equation. III. Rec. Math. [Mat. Sbornik] N.S. **54**, 277–286 (1943)
13. R. Meka, O. Nguyen, V. Vu, Anti-concentration for polynomials of independent random variables. Theory of Comput. **12**(11), 1–17 (2016)
14. E. Mossel, R. O'Donnell, K. Oleszkiewicz, Noise stability of functions with low influences: invariance and optimality. Ann. Math. **171**(1), 295–341 (2010)
15. H. Nguyen, Inverse Littlewood-Offord problems and the singularity of random symmetric matrices. Duke Math. J. **161**(4), 545–586 (2012)
16. H. Nguyen, V. Vu, *Small Ball Probability, Inverse Theorems, and Applications,* Erdös Centennial (Springer, Berlin/Heidelberg, 2013), pp. 409–463
17. A. Razborov, E. Viola, Real advantage. ACM Trans. Comput. Theory (TOCT) **5**(4), 17 (2013)
18. M. Rudelson, R. Vershynin, Non-asymptotic theory of random matrices: extremal singular values, in *Proceeding of the International Congress of Mathematicians*, vol. III (Hindustan Book Agency, New Delhi, 2010), pp. 1576–1602
19. T. Tao, V. Vu, From the Littlewood-Offord problem to the circular law: universality of the spectral distribution of random matrices. Bull. Am. Math. Soc. (N.S.) **46**(3), 377–396 (2009)