Twentieth Anniversary Volume

Discrete & Computational Geometry

Twentieth Anniversary Volume

Discrete & Computational Geometry

Editors
Jacob E. Goodman
János Pach
Richard Pollack

Springer

Editors

Jacob E. Goodman
The City College of The City University of New York
New York, NY 10031
USA
jgoodman@ccny.cuny.edu

János Pach
The City College of The City University of New York
New York, NY 10031
USA
pach@cims.nyu.edu

Richard Pollack
Courant Institute of Mathematical Sciences
New York University
New York, NY 10012
USA
pollack@cims.nyu.edu

ISBN: 978-0-387-87362-6 e-ISBN: 978-0-387-87363-3
DOI: 10.1007/978-0-387-87363-3

Library of Congress Control Number: 2008934108

Mathematics Subject Classification (2000): 52Bxx, 52Cxx, 68U05, 65D18, 51M20, 52Axx, 14Pxx

springer.com

Contents

Preface

While we were busy putting together the present collection of articles celebrating the twentieth birthday of our journal, *Discrete & Computational Geometry*, and, in a way, of the field that has become known under the same name, two more years have elapsed. There is no doubt that *DCG* has crossed the line between childhood and adulthood.

By the mid-1980s it became evident that the solution of many algorithmic questions in the then newly emerging field of computational geometry required classical methods and results from discrete and combinatorial geometry. For instance, visibility and ray shooting problems arising in computer graphics often reduce to Helly-type questions for line transversals; the complexity (hardness) of a variety of geometric algorithms depends on McMullen's upper bound theorem on convex polytopes or on the maximum number of "halving lines" determined by $2n$ points in the plane, that is, the number of different ways a set of points can be cut by a straight line into two parts of the same size; proximity questions stemming from several application areas turn out to be intimately related to Erdős's classical questions on the distribution of distances determined by n points in the plane or in space.

On the other hand, the algorithmic point of view has fertilized several fields of convexity and of discrete geometry which had lain fallow for some years, and has opened new research directions. Computing the convex hull or the diameter of a point set, or estimating the volume of a convex body or the maximum density of a packing of translates of a given convex body, has motivated a wide range of exciting new questions concerning classical concepts in discrete geometry. Motion planning problems have triggered the systematic study of the "combinatorial complexity" of the boundary of the union of geometric objects, and hence the development of Davenport–Schinzel theory, the use of epsilon-nets, Vapnik–Chervonenkis dimension, and probabilistic techniques. Similar methods have been needed for range searching, and this has also led to a renaissance of geometric discrepancy theory.

In the last two decades, *DCG* has provided a common platform for mathematicians working in the theory of packing and covering, and in convexity and combinatorial geometry, as well as for computer scientists interested in computational geometry, computational topology, geometric optimization, graph drawing, motion planning, and so on. In fact, exceeding all the expectations of its editors, the journal has served

as an effective catalyst in the creation of a new generation of researchers working on the common borderline between mathematics and computer science.

The present selection of 28 exceptionally strong articles, many of which solve longstanding open problems, reflects the current state of our subject, its many different facets, and its strong links to other important disciplines.

Nevo and Barvinok–Novik study problems related to Barnette's Lower Bound and McMullen's Upper Bound Theorem, respectively. Nagel gives a proof of the Kalai–Kleinschmidt–Lee conjecture for the maximum number of empty simplices in a simplicial polytope. Miller–Pak and Damian–Flatland–O'Rourke prove the existence of nonoverlapping unfoldings of certain manifolds. Billera–Hsiao–Provan construct nearly polytopal CW spheres with special properties. Khachiyan–Boros–Borys–Elbassioni–Gurvich show that generating all vertices of a polyhedron is a hard problem. Braun establishes improved estimates for the roots of Ehrhart polynomials of lattice polytopes. Przesławski–Yost give new conditions for the decomposability of polytopes as a Minkowski sum, while Richardson–Vu–Wu describe the asymptotic behavior of certain random polytopes.

Schreiber–Sharir design optimal shortest path algorithms on polytopes. Niyogi–Smale–Weinberger show how to find the homology of the underlying submanifold of a probability distribution with high confidence. Basu–Zell establish new bounds on Betti numbers of projections of semialgebraic sets. Efficient algorithms for snap rounding in pixel geometry and for computing optimal embeddings of paths, trees, and cycles in two and three dimensions are presented by Hershberger and by Agarwal–Klein–Knauer–Langerman–Morin–Sharir–Soss, respectively. Agarwal–Har–Peled–Yu apply coresets to design approximation algorithms to shape fitting. Shewchuk generalizes constrained Delaunay triangulations to higher dimensions, while Boissonat–Cohen–Steiner–Vegter find the first provably correct implicit surface meshing algorithm, where the mesh is isotopic to the surface.

Ackerman–Buchin–Knauer–Pinchasi–Rote and Gerken solve Murty's and Erdős's many-decade-old problems for finite point configurations. Pfender proves that every finite graph can be obtained as the visibility graph of a rational point set in the plane, while Pelsmajer–Schaefer–Štefankovič construct the first examples showing that the crossing number of a graph is not necessarily the same as its odd-crossing number. Aliev uses convex geometry to make progress on an old Erdős–Moser problem in additive number theory. Lee–Solomyak use dynamical systems to answer a question of Lagarias on Delone sets. Gronchi–Longinetti solve an extremal problem for polygons that plays a role in X-ray tomography. Bárány–Hubard–Jerónimo, Borcea–Goaoc–Petitjean, and Cheong–Goaoc–Holmsen–Petitjean solve various hyperplane- and line-transversal problems in Euclidean spaces.

Discrete & Computational Geometry saw the light of day in 1986, and this volume celebrates its majority. By now, the field, which has become inseparable from the journal, has acquired its own characteristics, its own methodology and toolbox. Deep connections have been discovered between its basic problems and many other fields of mathematics and computer science, such as additive combinatorics, topology, real algebraic geometry, randomized algorithms, and data structures. The field has its annual conferences: the ACM Symposia, the Fall Workshops, and the European Workshops on Computational Geometry, and a biennial meeting in Schloss Dagstuhl. Established research institutes such as DIMACS, MSRI, and IPAM regularly run special semester

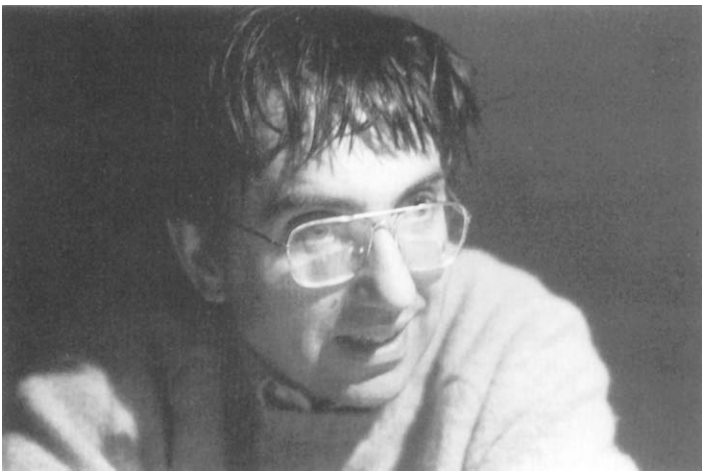
programs dedicated to the subject, and Oberwolfach sponsors a meeting every few years. We have several excellent textbooks for teaching discrete and computational geometry, not to mention two comprehensive handbooks. The "genie" has been let out of the bottle. Its movements and actions are now largely independent of the original intentions of its "creators," who include the founding editors of *DCG*. It has been a tremendous pleasure and honor to edit the journal, to watch it grow alongside the field proper, and to serve the community built around it. Our everlasting gratitude must also go to the late Walter Kaufmann-Bühler, who had the foresight to accept our original invitation to Springer-Verlag to publish a journal in this new field.

We dedicate the present volume to the members of the very active and gifted community of researchers who have taken part in the development of the field during the past more-than-two decades; many of them are represented in its pages.

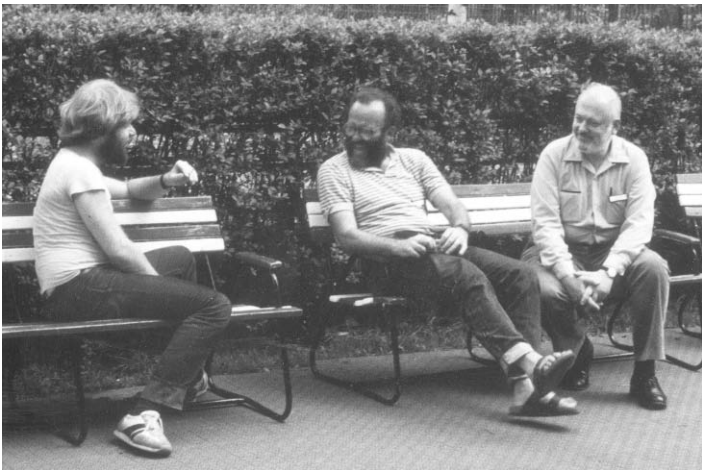
New York,
July 2008

Jacob E. Goodman
János Pach
Richard Pollack

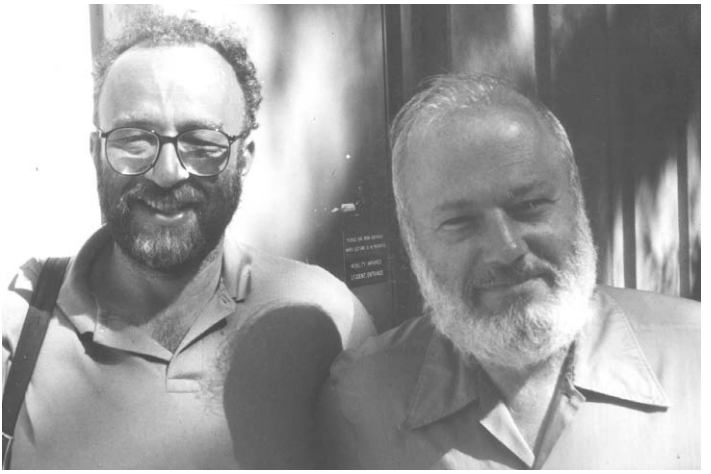
We would like to express our appreciation to those who helped us gather the photographs that appear in this volume: Ludwig Danzer, Wlodek Kuperberg, Ina Mette and Ute Motz (of Springer), Willy Moser, Lori Smith, Emo Welzl, and Jörg M. Wills.



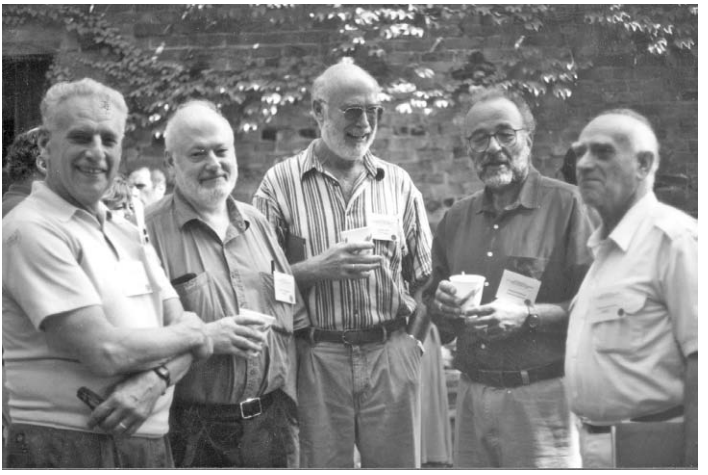
Walter Kaufmann-Bühler, Mathematics Editor, Springer-Verlag (1944–1986)



János Pach, Ricky Pollack, and Jacob E. Goodman—at the birth of *DCG*—Siofok, Hungary, 1985



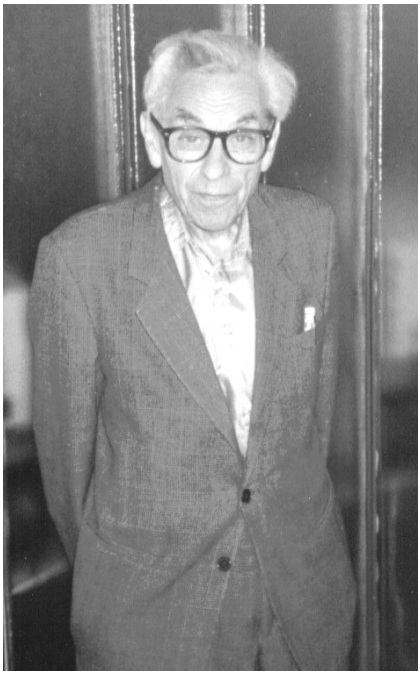
Ricky Pollack and Jacob E. Goodman at the AMS–IMS–SIAM Joint Summer Research Conference “Discrete and Computational Geometry,” Santa Cruz, California, 1986



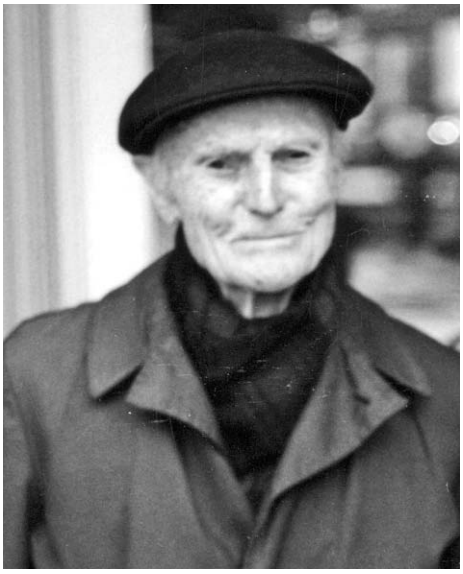
Branko Grünbaum, Jacob E. Goodman, Jörg M. Wills, Ricky Pollack, and Vladimir G. Boltyanski at the AMS–IMS–SIAM Joint Summer Research Conference “Discrete and Computational Geometry, Ten Years Later,” Mount Holyoke, Massachusetts, 1996

Summer Research Conference on Discrete and Computational Geometry, Monte Verita (Ascona), Switzerland, 1999

AMS–IMS–SIAM Joint Summer Research Conference “Discrete and Computational Geometry, Twenty Years Later,” Snowbird, Utah, 2006



Paul Erdős (1913–1996), Budapest, Hungary, 1992



H. S. M. Coxeter (1907–2003)

László Fejes Tóth (1915–2005), Tihany, Hungary, 1985

Victor Klee (1925–2007), Mount Holyoke, 1996

János Pach and Branko Grünbaum

Richard P. Stanley and Louis J. Billera, Mount Holyoke, 1996

Thomas C. Hales holding the *DCG* special issue on the Kepler conjecture, flanked by his editors, Gábor Fejes Tóth and Jeffrey C. Lagarias, Snowbird, 2006

Nina Amenta, Marie-Françoise Roy, and Ileana Streinu, Mount Holyoke, 1996

There Are Not Too Many Magic Configurations

Eyal Ackerman · Kevin Buchin · Christian Knauer · Rom Pinchasi · Günter Rote

Abstract A finite planar point set P is called a *magic configuration* if there is an assignment of positive weights to the points of P such that, for every line l determined by P, the sum of the weights of all points of P on l equals 1. We prove a conjecture of Murty from 1971 and show that if a set of n points P is a magic configuration, then P is in general position, or P contains $n-1$ collinear points, or P is a special configuration of 7 points.

Keywords Magic configuration · Euler's formula · Discharging method · Murty's conjecture · Points · Lines · Euclidean plane

1 Introduction

Let P be a finite set of points in the plane. P is called a *magic configuration* if there is an assignment of positive weights to the points of P such that, for every line l determined by P, the sum of the weights of all points of P on l equals 1. Figure 1 shows an example of a point set that is a magic configuration. This special point set (and any projective transformation of it) is called a *failed Fano* configuration.

The research by Rom Pinchasi was supported by a Grant from the G.I.F., the German-Israeli Foundation for Scientific Research and Development.

E. Ackerman
Computer Science Department, Technion—Israel Institute of Technology, Haifa 32000, Israel
e-mail: ackerman@cs.technion.ac.il

K. Buchin · C. Knauer · G. Rote
Institute of Computer Science, Freie Universität Berlin, Takustr. 9, 14195 Berlin, Germany

K. Buchin
e-mail: @inf.fu-berlin.de

R. Pinchasi (✉)
Mathematics Department, Technion—Israel Institute of Technology, Haifa 3200, Israel
e-mail: room@math.technion.ac.il

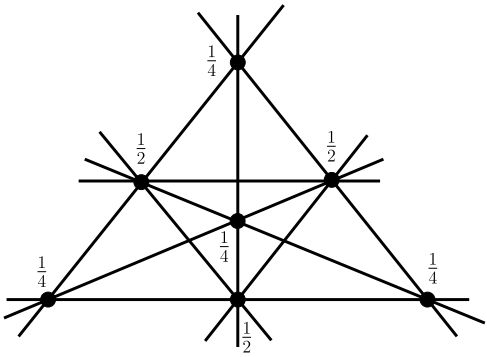


Fig. 1 Failed Fano configuration

We prove a conjecture of Murty [8] saying that apart from failed Fano configurations, every set of n points that is a magic configuration is either in general position, or contains $n-1$ collinear points. A few other remarks on the history of the problem can be found in *The Open Problems Project* [2].

Theorem 1 *There do not exist magic configurations of cardinality n, other than*

- *Configurations with $n-1$ collinear points, or*
- *Configurations in general position, that is, with no three points on a line, or*
- *A configuration with 7 points that up to a projective transformation is depicted in Figure* 1

We will now make some preliminary observations regarding magic configurations. Many of these observations can be found already in Murty's paper [8].

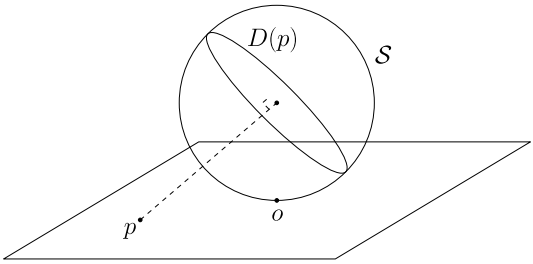
Assume that a configuration P of $n \geq 2$ points in the plane is magic and that its points are assigned positive weights that witness the fact that P is magic. Recall that a line determined by P is called *ordinary* if it includes precisely two points of P. By Gallai–Sylvester theorem [4, 10], the points of P must determine an ordinary line unless they are all collinear.

We claim that unless P has $n-1$ collinear points, then for every point $p \in P$ there is an ordinary line not passing through p. Indeed, otherwise, by the theorem of Kelly and Moser [6], the set $P \setminus \{p\}$ determines at least $\frac{3}{7}(n-1)$ ordinary lines (see [1] for the current best bound on the number of ordinary lines determined by n points). Clearly all these lines must be passing through p. It follows that at most $\frac{1}{7}(n-1)$ points of $P \setminus \{p\}$ lie on an ordinary line through p and these are all the ordinary lines determined by P, contradicting the Kelly–Moser theorem.

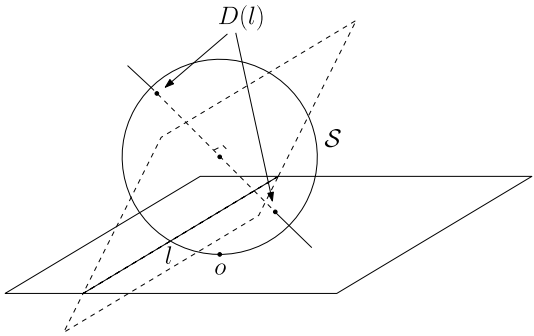
It is now easy to see that unless P has $n-1$ collinear points, every point through which there is an ordinary line must be assigned the weight $\frac{1}{2}$. To see this assume that p is such a point and assume without loss of generality that it is assigned a weight that is greater than $\frac{1}{2}$ (otherwise look at the other point on the ordinary line through p). Let q and r be two points different from p that constitute an ordinary line in P. One of q and r is assigned a weight greater than or equal to $\frac{1}{2}$. The sum of the weights assigned to the points on the line through that point and p will be strictly greater than 1, a contradiction.

Denote by A the set of all points in P through which there is an ordinary line, and assume that P does not have $n-1$ collinear points. Then each point in A is

Fig. 2 Duality between the plane and the unit sphere



(a) The dual $D(p)$ of a point p is the great circle that is the intersection of $\mathcal{S}$ with the plane through the center of $\mathcal{S}$ that is perpendicular to the line through p and the center of $\mathcal{S}$



(b) The dual $D(l)$ of a line l is the pair of antipodal points that are the intersection points of $\mathcal{S}$ and the line through the center of S that is perpendicular to the plane through l and the center of $\mathcal{S}$

assigned a weight of $\frac{1}{2}$. It follows that any line through two points in A must be ordinary. Observe that $|A| \geq 3$, as any noncollinear set of points determines at least 3 ordinary lines (see [6]) and this would be impossible if $|A| \leq 2$. Denote by B the set $P \setminus A$. Clearly, every point in B must be assigned a weight that is strictly smaller than $\frac{1}{2}$. Indeed, let $b \in B$ and $a \in A$. The line through a and b cannot be ordinary for otherwise $b \in A$. Therefore it must contain a third point c. As the weight assigned to a is $\frac{1}{2}$, it follows that the weight assigned to b can be at most $\frac{1}{2}$ minus the weight assigned to c.

Theorem 1 will therefore follow from the following theorem.

Theorem 2 *Let A and B be two nonempty sets of distinct points in the Euclidean plane such that $|A| \geq 3$. Assume that all the ordinary lines determined by $A \cup B$ are precisely all the lines determined by two points of A. Assume further that every point in $A \cup B$ is assigned a positive weight such that the sum of the weights of all points on any given line determined by $A \cup B$ is 1. Then the configuration of points $A \cup B$ is a failed Fano configuration that is equal, up to a projective transformation, to the one shown in Fig. 1, where A consists of the points whose weight is $\frac{1}{2}$.*

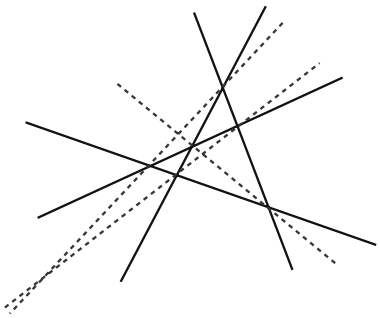


Fig. 3 A set of lines that corresponds to the exceptional configuration of Theorem 3

Instead of proving Theorem 2 we will prove its dual theorem on the sphere. We refer here to the standard duality under which the dual $D(p)$ of a point p in the plane is a great circle on the unit sphere $\mathcal{S}$ that touches the plane at the origin. The dual $D(l)$ of a line l in the plane is a pair of antipodal points on $\mathcal{S}$. For a point p in the plane, $D(p)$ is the great circle on $\mathcal{S}$ which is the intersection of $\mathcal{S}$ with the plane through the center of $\mathcal{S}$ that is perpendicular to the line through p and the center of $\mathcal{S}$ (see Fig. 2a). For a line l in the plane, $D(l)$ is the pair of antipodal points that are the intersection points of $\mathcal{S}$ and the line through the center of S that is perpendicular to the plane through l and the center of $\mathcal{S}$ (see Fig. 2b). This duality preserves incidence relations in the sense that if p is a point in the plane that is incident to a line l in the plane, then $D(p)$ is a great circle on $\mathcal{S}$ that is incident to the two points of $D(l)$. Recall that given an arrangement of curves, an *ordinary* intersection point is an intersection point through which precisely two curves pass.

Theorem 3 *Let A and B be two nonempty sets of distinct great circles on a sphere $\mathcal{S}$ such that $|A| \geq 3$. Assume that all the ordinary intersection points determined by $A \cup B$ are precisely all the intersection points determined by A. Assume further that every circle in $A \cup B$ is assigned a positive weight such that the sum of the weights of all circles incident to any given intersection point on $\mathcal{S}$ is 1. Then the configuration of circles $A \cup B$ is the sphere-dual of a failed Fano configuration that is equal, up to a projective transformation, to the one shown in Fig. 1. The set A consists of the circles dual to the points whose weight is $\frac{1}{2}$.*

Figure 3 shows a projection to the plane of the exceptional configuration of Theorem 3. The projection is a central projection through the center of $\mathcal{S}$ on a plane that touches $\mathcal{S}$. Under this projection every two antipodal points on $\mathcal{S}$ are projected to the same point in the plane.

2 Proof of Theorem 3

We refer to the circles in A as *red* circles and to the circles in B as *blue* circles. We remark that in all the next figures in this paper the solid lines represent the blue circles, while the dashed lines represent the red circles.

For every circle $s \in A \cup B$, let $W(s)$ denote the weight assigned to s. As we observed, for every $s \in A$ we have $W(s) = 1/2$, and for every $s \in B$, $0 < W(s) < 1/2$.

Table 1 Charge of objects of $\mathcal{B}$ before and after Steps 1–4

object of $\mathcal{B}$	$ch(\cdot)$	$ch_1(\cdot)$	$ch_2(\cdot)$	$ch_3(\cdot)$	$ch_4(\cdot)$
bad crossing point	-1	0	0	0	0
good crossing point	≥ 0	≥ 0	≥ 0	≥ 0	≥ 0
bad (but not evil) triangle	0	0	$-1/4$	≥ 0	≥ 0
evil triangle	0	0	$-1/4$	$-1/4$	0
0-quadrangle	1	1	≥ 0	≥ 0	≥ 0
1-quadrangle	1	≥ 0	≥ 0	≥ 0	≥ 0
good 2-quadrangle	1	≥ 0	≥ 0	≥ 0	≥ 0
bad 2-quadrangle	1	-1	0	0	0
0-pentagon	2	2	$\geq 3/4$	≥ 0	≥ 0
1-pentagon	2	≥ 1	$\geq 3/4$	$\geq 1/2$	$\geq 1/2$
2-pentagon	2	≥ 0	≥ 0	≥ 0	≥ 0
r-(k-gon) , $k \geq 6, r \leq \lfloor \frac{k}{2} \rfloor$	$k-3$	$\geq k-3-r$	$\geq \frac{3}{4}k-3-\frac{r}{2}$	$\geq \frac{1}{2}k-3$	$\geq \frac{1}{2}k-3$

We consider the arrangement $\mathcal{B}$ of the circles in B on the sphere $\mathcal{S}$. For every face f in $\mathcal{B}$, the *size* of f is the number of edges of the face f. We will use the term 'triangle' for a face of size three, the term 'quadrangle' for a face of size four, etc. Two faces in $\mathcal{B}$ are called *adjacent*, if they share an edge. Similarly, two edges in $\mathcal{B}$ are called adjacent, if they are incident to the same crossing point. A great circle $s \in B$ and a face f of $\mathcal{B}$ are called adjacent, if s includes an edge of f. We begin by assigning a *charge* $ch(\cdot)$ to the faces and vertices of the arrangement $\mathcal{B}$: The charge of a face of size k is $k-3$, while the charge of a crossing point of exactly k blue circles is $k-3$. For every $k \geq 2$ denote by f_k the number of faces in $\mathcal{B}$ of size k, and by t_k the number of crossing points of exactly k blue circles. It follows from Euler's formula that $\sum_k (k-3) f_k + \sum_k (k-3) t_k + 6 = 0$. Therefore, the overall charge is -6. Observe that any crossing point on a circle $b \in B$, even with a circle in A, is a crossing point in $\mathcal{B}$. Indeed, otherwise either it is an ordinary intersection point on b, or it is an intersection point that is not ordinary of at least two circles in A.

Our plan is to redistribute the charges (*discharge*) in four steps, such that finally every face and crossing point in $\mathcal{B}$ will have a nonnegative charge. Then it will follow that the total charge is nonnegative, hence a contradiction. For each $i = 1, 2, 3, 4$ we will denote by $ch_i(\cdot)$ the charge of an object (a face in $\mathcal{B}$ or a crossing point of blue circles) after the ith step. For convenience, Table 1 summarizes the charges of selected objects from $\mathcal{B}$ through the four steps of discharging.

Note that the only elements whose initial charge is negative are crossing points through which there are precisely two blue circles. We call such a crossing point *bad*. Observe that there are no faces of size two in $\mathcal{B}$. Indeed, otherwise all blue circles pass through the same two antipodal points p and p' on the sphere $\mathcal{S}$. As $|A| \geq 3$, there is a circle in A not passing through p, and hence also not through its antipodal point p'. This circle intersects the circles in B in ordinary intersection points, a contradiction.

The following claim and its corollary will be useful throughout the analysis of the discharging steps.

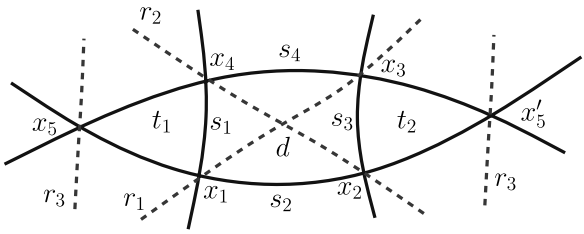


Fig. 4 d is adjacent to two triangles at two of its opposite edges

Claim 4 *Assume that there is a quadrangle d in $\mathcal{B}$ such that there are precisely two blue circles through every vertex of d, and d is adjacent to two triangles at two of its opposite edges. Then $A \cup B$ is the sphere-dual of a failed Fano configuration.*

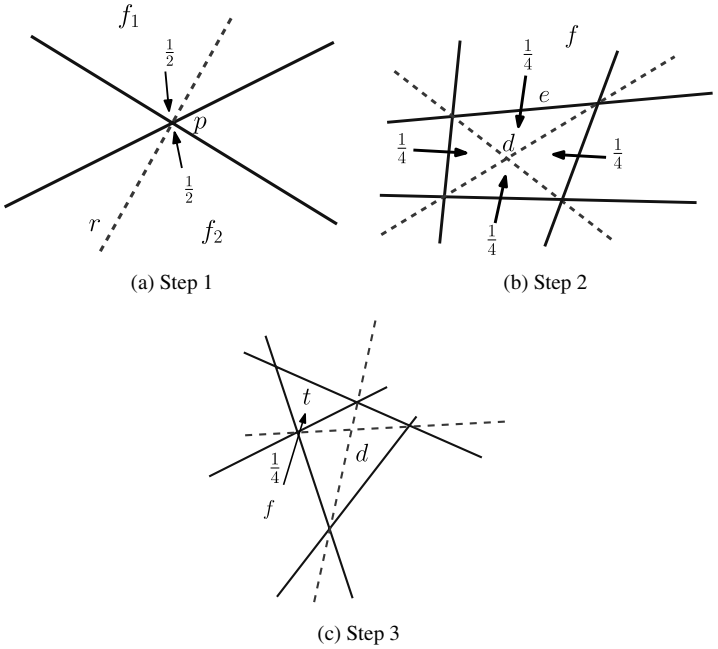
Proof Let t_1 and t_2 denote the two triangles adjacent to d at two of its opposite edges. Let $s_1, s_2, s_3.$ and s_4 denote the four blue circles that include the edges of d in the counterclockwise order so that s_1 and s_3 separate d from t_1 and t_2, respectively. Let x_1, x_2, x_3, and x_4 denote the four vertices of d listed in the counterclockwise order so that x_1 is the intersection point of s_1 and s_2. Since s_1 and s_2 are the only blue circles through x_1 and s_4 and s_1 are the only blue circles through x_4, it follows that s_2 and s_4 meet at a vertex of t_1 that we denote by x_5. Similarly, s_2 and s_4 meet at a vertex of t_2 that we denote by x_5'. x_5 and x_5' are therefore two antipodal points on the sphere $\mathcal{S}$. Therefore, x_1, x_2, x_5 and their antipodal points on $\mathcal{S}$ are the only intersection points on s_2.

Since there are precisely two blue circles through x_1, there must be a red circle passing through x_1. We denote this red circle by r_1. r_1 cannot cross t_1 and therefore it must cross d. Evidently, r_1 must pass through x_3. Similarly, there is a red circle r_2 passing through x_2 and x_4. As $|A| \geq 3$, there is a third red circle in A that we denote by r_3. r_3 and s_2 cannot cross at any other point but x_5 (and hence also x_5'). It follows that there are precisely three red circles in A since a fourth red circle would have to cross s_2 at a point through which one of r_1, r_2, or r_3 passes.

We claim that s_1, s_2, s_3, and s_4 are the only blue circles in B. Indeed, all other blue circles $s_5, \ldots, s_k$ must cross s_2 and s_4 at x_5 (and hence also at x_5'). None of $r_3, s_5, \ldots, s_k$ can cross s_1 at x_1 or x_4, and only one can cross s_1 at the intersection point of s_1 and s_3. Moreover, no two of $r_3, s_5, \ldots, s_k$ cross s_1 at a common point. It follows that one of $r_3, s_5, \ldots, s_k$ must cross s_1 at an ordinary intersection point, a contradiction.

Now it easily follows by inspection that $A \cup B$ must be the sphere-dual of a failed Fano configuration. More specifically, by looking at Fig. 3, we see that the sphere-dual of the failed Fano configuration has the properties of Claim 4. Moreover, from the assumption of Claim 4 we were led to conclude that there are no lines additional to those drawn in Fig. 4. It is easy to see that the only intersection point (modulo antipodals) not already indicated in Fig. 4 is the common intersection point of s_1, s_3, and r_3. Hence there is a unique arrangement satisfying the conditions of the claim and it is necessarily the sphere-dual of the failed Fano configuration. □

Corollary 5 *Assume that B consists of precisely four circles, then $A \cup B$ is the sphere-dual of a failed Fano configuration.*



(a) Step 1

(b) Step 2

(c) Step 3

Fig. 5 Discharging steps 1–3

Proof By previous arguments not all the circles in B are concurrent. If B has precisely three concurrent circles, then each of them has exactly four crossing points in $\mathcal{B}$. Since the circles in A cross the circles in B only at vertices of $\mathcal{B}$, and $|A| \geq 3$, there must be two circles of A crossing a circle of B at the same point, which is impossible. Therefore, B consists of four circles, no three of which are concurrent. Thus, the arrangement $\mathcal{B}$ satisfies the conditions of Claim 4. □

We proceed by describing the four discharging steps and analyzing their effect on the charges of the faces and intersection points of $\mathcal{B}$.

Step 1 (Charging bad crossing points) Let $\mathcal{C}$ denote the arrangement of all circles in $A \cup B$. Since we assume that no ordinary intersection point in $\mathcal{C}$ lies on a blue circle and that every pair of red circles cross at an ordinary point in $\mathcal{C}$, it follows that through each bad crossing point in $\mathcal{B}$ there is precisely one red circle. Let r be a red circle passing through a bad crossing point p, and let f_1 and f_2 be the two faces in $\mathcal{B}$ that are incident to p and are crossed by r (see Fig. 5a). Then, we take $1/2$ units of charge from each of f_1 and f_2 and charge it to p.

After Step 1 every crossing point of blue circles has a nonnegative charge. Let us now examine the remaining charge at the faces of the arrangement $\mathcal{B}$. A red circle can cross the boundary of a face in $\mathcal{B}$ only at its vertices, for otherwise we would have either an ordinary intersection point of $\mathcal{C}$ on a blue circle, or an intersection point of two (or more) red circles that is not ordinary in $\mathcal{C}$. Thus, every red circle that crosses a face f in $\mathcal{B}$ induces, in fact, a red diagonal in f. A face f with m such red diagonals

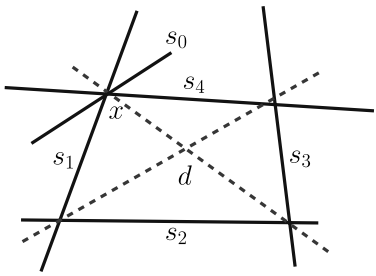


Fig. 6 A good 2-quadrangle cannot be incident to exactly one good crossing point

loses at most m units of charge in Step 1. We use an integer before the name of a face in $\mathcal{B}$ to denote the number of its red diagonals. For example, a 2-hexagon is a face of size six in $\mathcal{B}$ that has precisely two red diagonals. Since triangles cannot have a (red) diagonal, we refer to them simply as 'triangles' instead of 0-triangles. Thus, triangles do not lose charge in Step 1. Pentagons may have at most two red diagonals, and thus they remain with a nonnegative charge as well. The only elements whose charge might be negative after Step 1 are 2-quadrangles, as their charge might be -1, in case they are incident to four bad crossing points.

A crossing point x of circles from $\mathcal{C}$ is called *good*, if there is a (necessarily one) red circle through x and at least 3 blue circles through x. We call a 2-quadrangle *good*, if it is incident to a good crossing point. We call a 2-quadrangle that is not incident to any good crossing point a *bad* 2-quadrangle.

Claim 6 *Any good 2-quadrangle is incident to at least two good crossing points.*

Proof Assume to the contrary that d is a good 2-quadrangle that is incident to precisely one good crossing point x. Let s_1, s_2, s_3, and s_4 denote the four circles in $\mathcal{B}$ that constitute the edges of d in the counterclockwise order so that s_1 and s_4 are incident to x. As x is a good crossing point, there is another blue circle through x that we denote by s_0. (See Fig. 6.)

By our assumption, all the crossing points that are incident to d, with the exception of x, are incident to precisely two blue circles and one red circle. Considering the crossing point of s_1 and s_2, we see that $W(s_1) + W(s_2) = 1/2$. Similarly, considering the crossing point of s_2 and s_3, we see that $W(s_2) + W(s_3) = 1/2$, and in particular $W(s_1) = W(s_3)$. Considering the crossing point of s_3 and s_4, we see that $W(s_3) + W(s_4) = 1/2$. Therefore, $W(s_1) + W(s_4) = W(s_3) + W(s_4) = 1/2$. But this is a contradiction because considering the circles through x we see that $W(s_1) + W(s_4) \leq 1/2 - W(s_0) < 1/2$. □

As a corollary of Claim 6, we conclude that after Step 1 every good 2-quadrangle has a nonnegative charge, as it is incident to at most two bad crossing points. We still have to take care of the bad 2-quadrangles. This will be carried out in the next step.

Step 2 (Charging bad 2-quadrangles) In this step every bad 2-quadrangle compensates for its charge shortage by taking $1/4$ units of charge from each of its four neighboring faces. That is, let f be a face in $\mathcal{B}$ adjacent to a bad 2-quadrangle d, then d

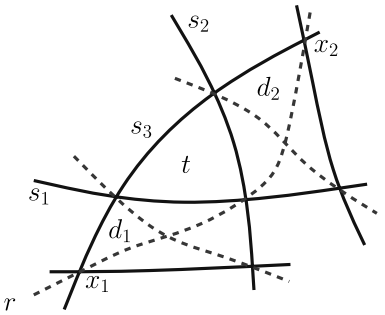


Fig. 7 A bad triangle cannot be adjacent to two bad quadrangles

takes the $1/4$ units of charge from the charge of f (see Fig. 5b). Note that in such a case f does not have red diagonals at the vertices of the edge common to f and d.

It is easy to check, by considering the different possibilities for f, that the only elements that might have a negative charge after Step 2 are triangles adjacent to bad 2-quadrangles. We refer the reader to the proof of Claim 7 for a proof of this observation. We call a triangle that is adjacent to a bad 2-quadrangle a *bad* triangle. Note that we may assume that a triangle might share an edge with at most one bad 2-quadrangle. Indeed, let t be a triangle adjacent to two bad 2-quadrangles d_1 and d_2. Let s_1, s_2, and s_3 denote the three blue circles that constitute the triangle t, such that s_1 and s_2 separate t from d_1 and d_2, respectively (see Fig. 7).

There is a red circle r through the intersection point of s_1 and s_2. r crosses s_3 at a vertex x_1 of d_1 and at a vertex x_2 of d_2, which are therefore antipodal points on the sphere $\mathcal{S}$. It follows that s_1, s_2, s_3, and another blue circle that passes through x_1 and x_2 are the only blue circles in B. By Corollary 5, $A \cup B$ is the sphere-dual of a failed Fano configuration.

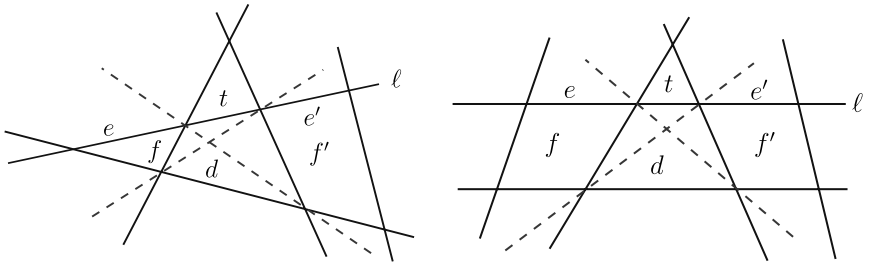
Step 3 (Charging some of the bad triangles) In this step we use the excess charge that exists at faces with at least five edges to charge part of the bad triangles.

Let f be a face in $\mathcal{B}$ with k edges, where $k \geq 5$. Let t be a bad triangle adjacent to a bad 2-quadrangle d. We transfer $1/4$ units of charge from f to t, if f and t share a vertex and f is adjacent to (that is, shares an edge with) d (see Fig. 5c).

Before continuing to the last step, we show that after Step 3, every face f with at least five edges remains with a nonnegative charge.

Claim 7 *Let f be a face with k edges, where $k \geq 5$. Then after Step 3 f has a nonnegative charge.*

Proof Let r be the number of red diagonals of f. Assume first that $k \geq 6$. Right after Step 1, the charge of f is at least $k - 3 - r$. f has exactly $k - 2r$ vertices that are not incident to a red diagonal, and hence at most $k - 2r$ edges none of whose vertices is incident to a red diagonal of f. It follows that f may be adjacent to at most $k - 2r$ (bad) 2-quadrangles. Therefore, the charge of f right after Step 2 is at least $k - 3 - r - \frac{k-2r}{4}$. As f may contribute $1/4$ units of charge to at most $k - 2r$ bad triangles, the charge of f right after Step 3 is at least $k - 3 - r - \frac{k-2r}{2} = \frac{k}{2} - 3 \geq 0$.



(a) An evil pair (ℓ, d) of degree 1 in H (b) An evil pair (ℓ, d) of degree 2 in H

Fig. 8 Evil pairs

It is left to consider the case where f is a pentagon. If f is a 2-pentagon, then f cannot be adjacent to any (bad) 2-quadrangle. Therefore, Step 2 as well as Step 3 do not affect the charge of f and it remains at least 0, as it is right after Step 1. If f is a 1-pentagon, then right after Step 1 the charge of f is at least 1. f may be adjacent to at most one 2-quadrangle. Therefore, right after Step 2 the charge of f is at least $3/4$. f contributes $1/4$ units of charge in Step 3 to at most one bad triangle and hence remains with a charge of at least $1/2$ after Step 3.

Finally, if f is a 0-pentagon, then after Step 1 the charge of f is 2. Observe that if f shares two adjacent edges e_1 and e_2 with bad 2-quadrangles d_1 and d_2, respectively, then the common vertex of e_1 and e_2 cannot be a vertex of a bad triangle t. Indeed, otherwise t is adjacent to two bad 2-quadrangles d_1 and d_2 which we have previously shown to be only possible in the case where $A \cup B$ is the sphere-dual of a failed Fano configuration. From this observation it follows that if f is adjacent to five bad 2-quadrangles, then it does not share a vertex with any bad triangle and hence the charge of f right after Step 3 is $3/4$. If f shares a vertex with five bad triangles, then it may be adjacent to at most two bad 2-quadrangles (in fact one could show that even that is not possible) and hence the charge of f after Step 3 is at least $1/4$. In all other cases f is adjacent to at most four bad 2-quadrangles and shares a vertex with at most four bad triangles and hence the charge of f after Step 3 is at least 0 (we remark that this last argument is by far suboptimal, yet suffices for our needs). □

Therefore, after Step 3 the only objects with a negative charge are those bad triangles who did not receive $1/4$ units of charge in Step 3. We call those triangles *evil*.

Step 4 (Charging evil triangles) After Step 3 of discharging, the only elements without the desired charge are evil triangles, as they are charged with $-1/4$ units of charge each. We will use the excess charge that exists at the 0-quadrangles to charge with $1/4$ units of charge each and every evil triangle.

For every 0-quadrangle q, consider the set E of edges of q that are not edges of bad 2-quadrangles. Then the charge of q after Step 3 is $|E|/4$. For every $e \in E$ let $\ell_e \in B$ be the great circle that includes e. We call the pair (ℓ_e, q) a *helping pair* and we designate $1/4$ unit from the charge of q to the pair (ℓ_e, q).

For any evil triangle t, let d be the bad 2-quadrangle adjacent to it, and let $\ell \in B$ be the great circle that separates t and d. We call the pair (ℓ, d) an *evil pair*. We will

show that there are at least as many helping pairs as there are evil pairs. Thus we will successfully charge each evil triangle with $1/4$ units of charge taken of the excess charge at the 0-quadrangles after step 3.

Define a bipartite graph H whose vertices are the evil pairs and the helping pairs. Let (ℓ, d) be an evil pair, let t be the evil triangle adjacent to d and ℓ, and let f and f' be the two faces in $\mathcal{B}$, other than t, that are adjacent to both ℓ and d. Let e and e' be the edges of f and f', respectively, on ℓ. Since t is evil, then f and f' can be either triangles or 0-quadrangles (see Fig. 8). Moreover, the edges e and e' cannot be edges of bad 2-quadrangles, as d is the only bad 2-quadrangle adjacent to t. Each of (ℓ, f) and (ℓ, f') is a helping pair, assuming f or f', respectively, are not triangles. If f is not a triangle, we connect (ℓ, d) in H to the helping pair (ℓ, f). Similarly, if f' is not a triangle, we connect (ℓ, d) in H to the helping pair (ℓ, f'). Observe that if both f and f' are triangles, then by Claim 4, $A \cup B$ is the sphere-dual of a failed Fano configuration. Therefore, we may assume that the degree in H of every evil pair is either 1 or 2 (see Fig. 8). The degree in H of every helping pair is at most 2, because a helping pair (ℓ, q) may be connected only to evil pairs (ℓ', d) such that $\ell = \ell'$ and d is adjacent to q. It follows that the connected components of H that include evil pairs are either paths alternating between evil pairs and helping pairs, or theoretically, even cycles alternating between evil pairs and helping pairs. Therefore, in order to show that there are at least as many helping pairs as there are evil pairs, it is enough to show that no connected component in H is a path both of whose end vertices are evil pairs.

Indeed, assume to the contrary that there is such a connected component in H. Let its vertices be $(\ell, d_1), (\ell, q_1), \ldots, (\ell, q_{k-1}), (\ell, d_k)$, so that for every $1 \leq i \leq k-1$, (ℓ, q_i) is a helping pair connected to both (ℓ, d_i) and (ℓ, d_{i+1}). It follows that there is a great circle $\ell' \in B$ that includes all edges of $d_1, \ldots, d_k$ and $q_1, \ldots, q_{k-1}$ that are opposite to those included in ℓ.

Since the degree in H of (ℓ, d_1) is 1, then the face in $\mathcal{B}$, other than q_1, adjacent to both ℓ, ℓ', and to d_1 must be a triangle which we denote by q_0. Similarly, the face in $\mathcal{B}$, other than q_{k-1}, adjacent to both ℓ, ℓ', and to d_k must be a triangle which we denote by q_k. Observe that ℓ and ℓ' meet at a vertex of q_0 and at a vertex of q_k (see Fig. 9).

We claim that the only triangles in $\mathcal{B}$ adjacent to ℓ' are q_0, q_k, and of course their antipodal triangles on the sphere $\mathcal{S}$. This is because for every $0 \leq i \leq k$, the face adjacent to ℓ' that shares an edge with q_i cannot be a triangle as it admits a red diagonal at least at one of its vertices. And moreover, we may assume that for every $1 \leq i \leq k$, the face adjacent to ℓ' that shares an edge with d_i is not a triangle. Indeed, otherwise by Claim 4, $A \cup B$ is the sphere-dual of a failed Fano configuration (recall that there is an evil triangle adjacent to d_i on the other side of ℓ on $\mathcal{S}$). This is a contradiction to a theorem of Levi [7] saying that in any nontrivial arrangement of lines in the real projective plane, every line must be adjacent to at least 3 triangular faces. (Here, we apply Levi's theorem after identifying antipodal points on the sphere $\mathcal{S}$ and thus reducing the great circles in $A \cup B$ to a set of lines in the projective plane.) Since the reference to Levi's theorem is not widely available we refer the reader also to [3, Sect. 5.4] and [5] for very short proofs of Levi's theorem.

We conclude that after Step 4, all the faces in the arrangement $\mathcal{B}$ have a nonnegative charge, and the same holds for every crossing point in $\mathcal{B}$. Thus, the overall charge is nonnegative, contradicting the fact the total charge in the beginning was -6.

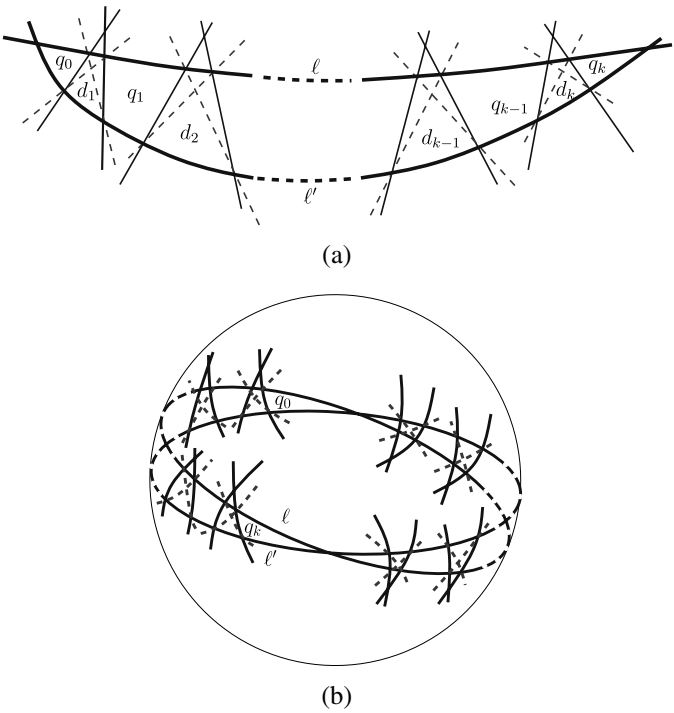


Fig. 9 A connected component in H both of whose endpoints are evil pairs

3 Notes and Concluding Remarks

If the arrangement $\mathcal{B}$ is in general position in the sense that no three blue circle from B pass through the same point, then Theorem 3 and hence also its dual Theorem 2 could be strengthened as follows leaving the proof almost as is:

Theorem 8 *Let A and B be two nonempty disjoint sets of points in the plane such that $|A| > 1$, and B is in general position. Assume that no line determined by A passes through a point of B. Then there is an ordinary line in $A \cup B$ through a point in B, unless $A \cup B$ is, up to a projective transformation, the configuration in Fig.* 1.

To see why Theorem 8 follows from the proof of Theorem 3, observe that if the circles in B are in general position, then there are no good crossing points in $\mathcal{C}$, and hence the assumptions in Theorem 3 on the weights assigned to the circles in $\mathcal{C}$ are not required. In Theorem 8 we allow more than two points of A to be collinear as long as they are not collinear with a point of B. Indeed, in the proof of Theorem 3 we did not really use the assumption that every intersection point determined by A is an ordinary intersection point with respect to $A \cup B$, but only that no intersection point determined by circles from A is incident to a circle from B.

Proving Theorem 8 for the case B is not required to be in general position would imply that the following conjecture[1] holds. Recall that an *ordinary point* in a point configuration P is a point $x \in P$ through which there is an ordinary line.

Conjecture 9 *Let $G = (V, E)$ be the* Sylvester Graph *of a finite set of points P. That is, $V = \{p \in P | p$ is an ordinary point in $P\}$ and $E = \{(p_1, p_2) | p_1$ and p_2 determine an ordinary line in $P\}$. Then G is a complete (nonempty) graph if and only if no three points in P are collinear, or P is a failed Fano configuration.*

We would like to note a corollary of Theorem 8. It is well known that the set of edges of a complete graph on $2n$ vertices can be partitioned into (necessarily $2n - 1$) edge-disjoint perfect matchings. A nice way to realize such a partitioning is to think about the vertices of K_{2n} as the vertices of a regular $(2n - 1)$-gon plus its center. Then every one of the $2n - 1$ directions of the edges of the $(2n - 1)$-gon induces a perfect matching in which two points are matched if the straight line they determine is parallel to the direction we choose, plus taking the center to be matched with the remaining point. These $2n - 1$ perfect matchings are edge-disjoint.

Now let G be a complete geometric graph on $2n$ vertices in general position in the plane. We call a matching in G *geometrically induced*, if the lines containing the edges of the matching are concurrent. If a matching of G is geometrically induced, then the meeting point of all lines that include an edge of the matching is called the *center* of the matching. The question is can we partition the set of edges of a complete geometric graph G on $2n$ vertices in general position in the plane into edge-disjoint geometrically induced perfect matchings. By Theorem 8, this is impossible unless $n = 1$ or $n = 2$. Indeed, assume it is possible and let B be the set of $2n$ vertices of G, and let A be the set of all points that are the centers of the geometrically induced perfect matchings. Then A and B satisfy the assumptions in Theorem 8.

It is an interesting open question of what is the maximum possible number of edge-disjoint geometrically induced perfect matchings of a complete geometric graph on $2n$ vertices in general position in the plane. It seems natural to conjecture that the answer should be $n + 1$. This number is attained for the set of vertices of a regular $2n$-gon in the plane when n is even. Here observe that the geometrically induced perfect matchings whose centers are the points at infinity that correspond to the n directions of the edges of the regular $2n$-gon plus the center of the $2n$-gon, are all pairwise edge-disjoint.

One can try to weaken the notion of a magic configuration and omit the restriction of all weights assigned to the points being positive. In this case there seem to be a much larger variety of magic configurations and yet not every configuration is magic. In this context it is interesting to note that given that a configuration is magic (even in the weak sense) it is very easy to assign the right weights (and in a unique way) to the points, just as a function of the number of lines determined by the set that pass through each of the points of the set. To this end let $p_1, \ldots, p_n$ denote the points of a magic configuration P. For every $1 \leq i \leq n$ let x_i denote the weight assigned to p_i and let k_i be the number of lines determined by P that pass through p_i. For convenience

[1] This conjecture is attributed to Sylvester according to Smyth [9].

denote $Y = \sum_{i=1}^{n} x_i$. Fix i and consider the point p_i. There are k_i lines determined by P that pass through p_i. The sum of the weights assigned to the points of P on each of these lines is 1. It follows that $Y = k_i - x_i(k_i - 1)$. Therefore, $x_i = \frac{k_i - Y}{k_i - 1}$. We can get an explicit expression for x_i just in terms of k_j $(j = 1, \ldots, n)$. Observe that $Y = \sum_{j=1}^{n} x_j = \sum_{j=1}^{n} \frac{Y - k_j}{1 - k_j}$. Therefore,

$$Y = \frac{\sum_{j=1}^{n} \frac{k_j}{k_j - 1}}{1 + \sum_{j=1}^{n} \frac{1}{k_j - 1}}, \quad \text{and hence,} \quad x_i = \frac{1}{k_i - 1}\left(k_i - \frac{\sum_{j=1}^{n} \frac{k_j}{k_j - 1}}{1 + \sum_{j=1}^{n} \frac{1}{k_j - 1}}\right).$$

Observe in particular that if $k_i = k_{i'}$, then $x_i = x_{i'}$. It is also clear from here that the weights assignment is unique, if exists.

Acknowledgements We would like to thank Micha Sharir for extremely helpful discussions on Murty's problem. We also thank anonymous referees for several helpful suggestions for improving the presentation of the paper.

References

1. Csima, J., Sawyer, E.T.: There exist $6n/13$ ordinary points. Discrete Comput. Geom. **9**(1), 187–202 (1993)
2. Demaine, E., Mitchell, J.S.B., O'Rourke, J.: Problem 65: Magic configurations. The Open Problems Project, http://maven.smith.edu/~orourke/TOPP/P65.html
3. Felsner, S.: Geometric Graphs and Arrangements. Some Chapters from Combinatorial Geometry, 1st edn. Advanced Lectures in Mathematics. Vieweg, Wiesbaden (2004)
4. Gallai, T.: Solution of problem 4065. Am. Math. Mon. **51**, 169–171 (1944)
5. Grünbaum, B.: Arrangements and Spreads. Conference Board of the Mathematical Sciences Regional Conference Series in Mathematics, vol. 10. American Mathematical Society, Providence (1972)
6. Kelly, L.M., Moser, W.O.J.: On the number of ordinary lines determined by n points. Can. J. Math. **1**, 210–219 (1958)
7. Levi, F.: Die Teilung der projectiven Ebene durch Gerade oder Pseudogerade. Ber. Math.-Phys. Kl. Sächs. Akad. Wiss. **78**, 256–267 (1926)
8. Murty, U.S.R.: How many magic configurations are there? Am. Math. Mon. **78**(9), 1000–1002 (1971)
9. Smyth, W.F.: Sylvester configurations. James Cook Math. Notes **5–49**, 5193–5196 (1989)
10. Sylvester, J.J.: Mathematical question 11851. Educational Times **59**, 98–99 (1893)

Computing the Detour and Spanning Ratio of Paths, Trees, and Cycles in 2D and 3D

Pankaj K. Agarwal · Rolf Klein · Christian Knauer · Stefan Langerman · Pat Morin · Micha Sharir · Michael Soss

Abstract The detour and spanning ratio of a graph G embedded in $\mathbb{E}^d$ measure how well G approximates Euclidean space and the complete Euclidean graph, respectively. In this paper we describe $O(n \log n)$ time algorithms for computing the detour and spanning ratio of a planar polygonal path. By generalizing these algorithms, we

This research was partly funded by CRM, FCAR, MITACS, and NSERC. P.A. was supported by NSF under grants CCR-00-86013 EIA-99-72879, EIA-01-31905, and CCR-02-04118, by ARO grants W911NF-04-1-0278 and DAAD19-03-1-0352, and by a grant from the U.S.-Israeli Binational Science Foundation. R.K. was supported by DFG grant Kl 655/14-1. M.S. was supported by NSF Grants CCR-97-32101 and CCR-00-98246, by a grant from the U.S.-Israeli Binational Science Foundation (jointly with P.A.), by a grant from the Israeli Academy of Sciences for a Center of Excellence in Geometric Computing at Tel Aviv University, and by the Hermann Minkowski–MINERVA Center for Geometry at Tel Aviv University.

Some of these results have appeared in a preliminary form in [2, 20].

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

R. Klein
Institut für Informatik I, Universität Bonn, Römerstraße 164, 53117 Bonn, Germany
e-mail: rolf.klein@uni-bonn.de

C. Knauer
Institut für Informatik, Freie Universität Berlin, Takustraße 9, 14195 Berlin, Germany
e-mail: knauer@inf.fu-berlin.de

S. Langerman
FNRS, Département d'Informatique, Université Libre de Bruxelles, ULB CP212, Boulevard du Triomphe, 1050 Bruxelles, Belgium
e-mail: sl@cgm.cs.mcgill.ca

P. Morin
School of Computer Science, Carleton University, 1125 Colonel By Drive, Ottawa, ON K1S 5B6, Canada
e-mail: morin@cs.carleton.ca

obtain $O(n \log^2 n)$-time algorithms for computing the detour or spanning ratio of planar trees and cycles. Finally, we develop subquadratic algorithms for computing the detour and spanning ratio for paths, cycles, and trees embedded in $\mathbb{E}^3$, and show that computing the detour in $\mathbb{E}^3$ is at least as hard as Hopcroft's problem.

1 Introduction

Suppose we are given an embedded connected graph $G = (V, E)$ in $\mathbb{E}^d$. Specifically, V consists of points in $\mathbb{E}^d$ and E consists of closed straight line segments whose endpoints are in V. For any two points p and q in $\bigcup_{e \in E} e$, let $d_G(p, q)$ be the shortest path between p and q along the edges of G. The *detour* between p and q in G is defined as

$$\delta_G(p, q) = \frac{d_G(p, q)}{\|pq\|}$$

where $\|pq\|$ denotes the Euclidean distance between p and q. The *detour of* G is defined as the maximum detour over all pairs of points in $\bigcup_{e \in E} e$, i.e.,

$$\delta(G) = \sup_{p \neq q} \delta_G(p, q).$$

The challenge is in computing the detour quickly. Several cases of this generic problem have been studied in the last few years. One variant results from restricting the points p, q in the above definition to a smaller set. For example, the *spanning ratio* or *stretch factor* of G is defined as the maximum detour over all pairs of *vertices* of G, i.e.,

$$\sigma(G) = \sup_{\substack{p \neq q \\ p,q \in V}} \delta_G(p, q).$$

Such restrictions influence the nature of the problem considerably. In this paper we study both, detour and spanning ratio.

The case of G being a planar polygonal chain is of particular interest. Alt et al. [6] proved that if the detour of two planar curves is at most κ, then their Fréchet distance is at most $\kappa + 1$ times their Hausdorff distance. The Fréchet and Hausdorff distances are two commonly used similarity measures for geometric shapes [5]. Although the Hausdorff distance works well for planar regions, the Fréchet distance is more suitable to measure the similarity of two curves [5]. However, the Fréchet distance is

M. Sharir
School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel
e-mail: michas@tau.ac.il

M. Sharir
Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA

M. Soss
Foreign Exchange Strategy Division, Goldman Sachs, New York, USA
e-mail: soss@cs.mcgill.ca

much harder to compute [6]. A relationship between the two measures suggests that one could use the Hausdorff distance when the detours of the two given curves are bounded and small. This is the only known condition (apart from convexity) under which a linear relationship between the two measures is known.

Analyzing on-line navigation strategies also often involves estimating the detour of curves [8, 17]. Sometimes the geometric properties of curves allow us to infer upper bounds on their detour [4, 18, 24], but these results do not lead to efficient computation of the detour of the curve.

Related Work Recently, researchers have become interested in computing the detour and spanning ratio of embedded graphs. The spanning ratio of a graph G embedded in $\mathbb{E}^d$ can be obtained by computing the shortest paths between all pairs of vertices of G. Similarly, the detour of G can be determined by computing the detour between every pair of edges $e_1 = (u_1, v_1)$ and $e_2 = (u_2, v_2)$. Although this seems to involve infinitely many pairs of points, this problem is of constant size: For each pair of points (p, q) in $e_1 \times e_2$, the *type* of the shortest connecting path $d_G(p, q)$ is determined by the two endpoints of e_1 and e_2 contained in this path. In the 2-dimensional rectangular parameter space of all positions of p and q on e_1 and e_2, classification by type induces at most four regions that are bounded by a constant number of line segments. For each region, the maximization problem can be solved in time $O(1)$, after having computed the shortest paths between all pairs of vertices of G. This approach, however, requires $\Omega(n^2)$ and $\Omega(m^2)$ time for computing the spanning ratio and detour, respectively, where n denotes the number of vertices and m is the number of edges. Surprisingly, these are the best known results for these problems for arbitrary crossing-free graphs in $\mathbb{E}^2$. Even if the input graph G is a simple path in $\mathbb{E}^2$, no subquadratic-time algorithm has previously been known for computing its detour or spanning ratio.

Narasimhan and Smid [23] study the problem of approximating the spanning ratio of an arbitrary geometric graph in $\mathbb{E}^d$. They give an $O(n \log n)$-time algorithm that computes an $(1 - \varepsilon)$-approximate value of the spanning ratio of a path, cycle, or tree embedded in $\mathbb{E}^d$. More generally, they show that the problem of approximating the spanning ratio can be reduced to answering $O(n)$ approximate shortest-path queries after $O(n \log n)$ preprocessing.

Ebbers-Baumann et al. [10] have studied the problem of computing the detour of a planar polygonal chain G with n vertices. They have established several geometric properties, the most significant of which (restated in Lemma 2.1) is that the detour of G is always attained by two mutually visible points p, q, one of which is a vertex of G. Using these properties, they develop an ε-approximation algorithm that runs in $O((n/\varepsilon) \log n)$ time. However, the existence of a subquadratic exact algorithm has remained elusive.

New Results In this paper we present randomized algorithms with $O(n \log n)$ expected running time that compute the exact spanning ratio or detour of a polygonal path with n vertices embedded in $\mathbb{E}^2$. These are the first subquadratic-time algorithms for finding the exact spanning ratio or detour, and they solve open problems posed in at least two papers [10, 23]. Our algorithm for the spanning ratio is worst-case optimal, as shown in [23], and we suspect that the algorithm for the detour is

also optimal, although we are not aware of a published $\Omega(n \log n)$ lower bound. By extending these algorithms, we present $O(n \log^2 n)$ expected time randomized algorithms for computing the detour and spanning ratio of planar cycles and trees. We can also obtain deterministic versions of our algorithms. They are more complicated and a bit slower—they run in $O(n \log^c n)$ time, for some constant c.

We also consider the problem of computing the detour and spanning ratio of 3-dimensional polygonal chains, and show that the first problem can be solved in randomized expected time $O(n^{16/9+\varepsilon})$, for any $\varepsilon > 0$ (where the constant of proportionality depends on ε), and the second problem can be solved in randomized expected time $O(n^{4/3+\varepsilon})$, for any $\varepsilon > 0$. Using the same extensions as in the planar case, this leads to subquadratic time algorithms for 3-dimensional trees and cycles. We also show that it is unlikely that an $o(n^{4/3})$-time algorithm exists for computing the detour of 3-dimensional chains, since this problem is at least as hard as Hopcroft's problem, for which a lower bound of $\Omega(n^{4/3})$, in a special model of computation, is given in [12].

Preliminary versions of this work appeared in [2, 20]; the 2-dimensional algorithm described in [20] is significantly different from the one presented here.

2 Polygonal Chains in the Plane

Let the graph $P = (V, E)$ be a simple polygonal chain in the plane with n vertices. That is, $V = \{p_0, \ldots, p_{n-1}\}$ is a set of n points in $\mathbb{E}^2$, and $E = \{[p_{i-1}, p_i] \mid i = 1, \ldots, n-1\}$. Throughout the paper, we write P when referring to the set $\bigcup_{e \in E} e$. We extend the definition of the detour from points to any two subsets A and B of P, by putting

$$\delta_P(A, B) = \sup_{\substack{a \in A, b \in B \\ a \neq b}} \delta_P(a, b),$$

which we call the *P-detour* between A and B. We also write $\delta_P(A) = \delta_P(A, A)$. Thus, $\delta(P) = \delta_P(P) = \delta_P(P, P)$ and $\sigma(P) = \delta_P(V, V)$. Since P will be fixed throughout this section, we will omit the subscript P from δ.

2.1 Overall Approach

Since computing the detour is more involved than computing the spanning ratio, we present below the algorithm for solving the detour problem. Certain modifications and simplifications, noted on the fly, turn the algorithm into one that computes the spanning ratio.

We first describe an algorithm for the decision problem for the detour: "Given a parameter $\kappa \geq 1$, determine whether $\delta(P) \leq \kappa$." Our algorithm makes crucial use of the following properties established in [10]. The proof of property (iii) is straightforward. It implies that the maximum detour is attained by a pair of co-visible points. Property (ii) ensures that one of them can be assumed to be a vertex. Together, (ii) and (iii) imply property (i).

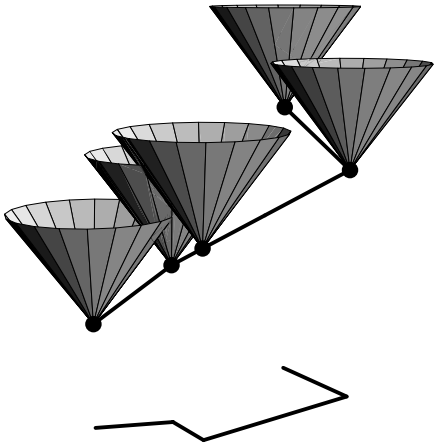


Fig. 1 Transforming P into a 3-dimensional chain

Lemma 2.1 (Ebbers-Baumann et al. [10])

(i) *Let V be the set of vertices in the polygonal chain P, and let $\kappa \geq 1$. There is a pair $(p, q) \in P \times P$ so that $\delta(p, q) > \kappa$ if and only if there is a pair $(p', q') \in P \times V$ so that $\delta(p', q') > \kappa$ and p' is visible from q'.*

(ii) *Assume that the detour attains a local maximum at two points, q, q' that are interior points of edges e, e' of P, correspondingly. Then the line segment qq' forms the same angle with e and e', and the detour of q, q' does not change as both points move, at the same speed, along their corresponding edges.*

(iii) *Let q, q' be two points on P, and assume that the line segment connecting them contains a third point, r, of P. Then $\max\{\delta(q, r), \delta(r, q')\} \geq \delta(q, q')$. Moreover, if the equality holds, then $\delta(q, r) = \delta(r, q') = \delta(q, q')$.*

We observe that a claim analogous to property (i) does not hold for the spanning ratio: while it is always attained by two vertices, by definition, these vertices need not be co-visible. As an immediate corollary of Lemma 2.1, we always have $\delta(P) = \delta(P, V)$. It thus suffices to describe an algorithm for the decision problem: *Given a parameter $\kappa \geq 1$, determine whether $\delta(P, V) \leq \kappa$*. We will then use a randomized technique by Chan [9] to compute the actual value of $\delta(P) = \delta(P, V)$.

2.2 Decision Algorithm

We orient P from p_0 to p_{n-1}. For a given parameter $\kappa \geq 1$, we describe an algorithm that determines whether for all pairs $(p, q) \in V \times P$, so that p lies before q, the inequality $\delta(p, q) \leq \kappa$ holds. By reversing the orientation of P and repeating the same algorithm once more, we can also determine whether for all pairs $(p, w) \in V \times P$ so that p lies after q the property $\delta(q, p) \leq \kappa$ is fulfilled.

For a point $p \in P$, we define the *weight* of p to be

$$\omega(p) = d_P(p_0, p)/\kappa.$$

Let C denote the cone $z = \sqrt{x^2 + y^2}$ in $\mathbb{E}^3$. We map each point $p = (p_x, p_y) \in V$ to the cone $C_p = C + (p_x, p_y, \omega(p))$. That is, we translate the apex of C (i.e., the origin)

to the point $\hat{p} = (p_x, p_y, \omega(p))$. If we regard C_p as the graph of a bivariate function, which we also denote by C_p, then for any point $q \in \mathbb{E}^2$, $C_p(q) = \|qp\| + \omega(p)$ holds. Let $\mathfrak{C} = \{C_p \mid p \in V\}$. We map a point $q = (q_x, q_y) \in P$ to the point $\hat{q} = (q_x, q_y, \omega(q))$ in $\mathbb{E}^3$. For any subchain π of P, we define $\hat{\pi} = \{\hat{q} \mid q \in \pi\}$.

Lemma 2.2 *For any point $q \in P$ and a vertex $p \in V$ that lies before q on P, $\delta(p, q) \leq \kappa$ if and only if $\hat{q}$ lies below the cone C_p.*

Proof

$$\begin{aligned}
\delta(p,q) \leq \kappa &\iff \frac{d_P(p,q)}{\|qp\|} \leq \kappa \\
&\iff \frac{d_P(p_0,q) - d_P(p_0,p)}{\|qp\|} \leq \kappa \\
&\iff \frac{d_P(p_0,q)}{\kappa} \leq \|qp\| + \frac{d_P(p_0,p)}{\kappa} \\
&\iff \omega(q) \leq \|qp\| + \omega(p) \\
&\iff \omega(q) \leq C_p(q).
\end{aligned}$$

That is, $\delta(p, q) \leq \kappa$ if and only if $\hat{q}$ lies below the cone C_p. □

Since the cones C_p are erected on the chain $\hat{P}$, the point $\hat{q}$, for any $q \in P$, always lies below all the cones erected on vertices appearing after q on P. Therefore, if we denote by V_q the set of all vertices $p \in V$ that precede q along P, Lemma 2.2 implies that $\delta(\{q\}, V_q) \leq \kappa$ if and only if $\hat{q}$ lies on or below each of the cones in $\mathfrak{C}$, i.e., if and only if $\hat{q}$ lies on or below the lower envelope of $\mathfrak{C}$.

The minimization diagram of $\mathfrak{C}$, the projection of the lower envelope of $\mathfrak{C}$ onto the xy-plane, is the additive-weight Voronoi diagram $\mathrm{Vor}_\omega(V)$ of V, under the weight function ω. For a point $p \in V$, let $\mathrm{Vor}_\omega(p)$ denote the Voronoi cell of p in $\mathrm{Vor}_\omega(V)$. $\mathrm{Vor}_\omega(V)$ can be computed in $O(n \log n)$ time [13].

We first test whether $\mathrm{Vor}_\omega(p)$ is nonempty for every vertex $p \in V$. If not, we obtain a pair of vertices that attain a detour larger than κ, namely a vertex p that has an empty Voronoi cell, and a vertex q whose cone C_q passes below $\hat{p}$.

Note that if $\mathrm{Vor}_\omega(p)$ is empty for some vertex $p \in V$, then we also know that the spanning ratio of P is larger than κ. Conversely, if the spanning ratio is larger than κ, then some Voronoi cell $\mathrm{Vor}_\omega(p)$ must be empty. Thus, the decision procedure for the spanning ratio terminates after completing this step.

We can therefore assume, for the case of detour, that $\mathrm{Vor}_\omega(p)$ is nonempty for every vertex $p \in V$. To check whether $\hat{P}$ lies below the lower envelope of $\mathfrak{C}$, we proceed as follows. We partition P into a family E of maximal connected subchains so that each subchain lies within a single Voronoi cell of $\mathrm{Vor}_\omega(V)$. Since $\mathrm{Vor}_\omega(p)$ is nonempty for every vertex $p \in V$, p is the only vertex of P that lies in $\mathrm{Vor}_\omega(p)$. Therefore every subchain in E is either a segment or consists of two connected segments with p as their common endpoint. For each such segment $e \in E$, if e lies in $\mathrm{Vor}_\omega(p)$, we can determine in $O(1)$ time whether $\hat{e}$ lies fully below C_p. If this is true

for all segments, then $\hat{P}$ lies below $\mathfrak{C}$. The total time spent is $O(n)$ plus the number of segments. Unfortunately, the number of segments may be quadratic in the worst case, so we cannot afford to test them all.

We circumvent the problem of having to test all segments by using the observation (i) from Lemma 2.1 that it is sufficient to test all $q \in P$ that are visible from p. More precisely, let $\mathfrak{A}$ denote the planar subdivision obtained by overlaying $\mathrm{Vor}_\omega(V)$ with P. Each edge of $\mathfrak{A}$ is a portion of an edge of P or of $\mathrm{Vor}_\omega(V)$. For a vertex $p \in V$, let f_p denote the set of (at most *two*) faces of $\mathfrak{A}$ containing p, and let E_p denote the set of edges of $\mathfrak{A}$ that are portions of P and that bound the faces in f_p. The discussion so far implies the following lemma.

Lemma 2.3 *$\hat{P}$ lies below all the cones of $\mathfrak{C}$ if and only if $\bigcup\{\hat{e} \mid e \in E_p\}$ lies below all the cones of $\mathfrak{C}$.*

The algorithm thus proceeds as follows: We compute the Voronoi diagram $\mathrm{Vor}_\omega(V)$ in $O(n \log n)$ time [7]. By using the red-blue-merge algorithm of Guibas et al. [15] (see also [11, 25]), we compute the sets of faces f_p for all $p \in V$, which in turn gives us the sets E_p for all $p \in V$. By the Combination Lemma of Guibas et al. [15], $\sum_{p \in V} |E_p| = O(n)$, and the set $\{E_p \mid p \in V\}$ can be computed in $O(n \log n)$ time. Finally, for each edge $e \in E_p$, we determine whether $\hat{e}$ lies below C_p in $O(1)$ time. The overall running time of the algorithm is $O(n \log n)$.

As mentioned in the beginning, we next reverse the orientation of P and repeat the algorithm to determine whether for each vertex $p \in V$ lying after a point $q \in P$ the inequality $\delta(p, q) \leq \kappa$ holds. (Note that this reversal is not required in the decision procedure for the spanning ratio.) Putting everything together, we obtain the following.

Lemma 2.4 *Let P be as polygonal chain with n vertices embedded in $\mathbb{E}^2$, and let $\kappa \geq 1$ be a parameter. We can decide in $O(n \log n)$ time whether $\delta(P) \leq \kappa$ or $\sigma(P) \leq \kappa$.*

Let $W \subseteq V$ be a subset of vertices of P, and let Q be a subchain of P; set $m = |W| + |Q|$. Assuming that the weights of all vertices in W have been computed, the decision algorithm described above can be used to detect in $O(m \log m)$ time whether $\sigma(W, Q) \leq \kappa$. However, unlike $\delta(V, P)$, the detour of the entire chain P, $\delta(W, Q)$ need not be realized by a co-visible pair of points in $W \times Q$, so it is not clear how to detect in $O(m \log m)$ time whether $\delta(W, Q) \leq \kappa$. Instead we can make a weaker claim. Let $\delta^*(W, Q) = \sup_{(p,q) \in W \times Q} \delta(p, q)$, where the supremum is taken over all pairs of points such that the interior of the segment pq does not intersect the interior of an edge of Q. Obviously, $\delta^*(W, Q) \leq \delta(W, Q)$. Clearly, the above decision algorithm can detect in $O(m \log m)$ time whether $\delta^*(W, Q) \leq \kappa$. Lemma 2.1(iii) implies that if $\delta(W, Q) = \delta(P)$, then $\delta^*(W, Q) = \delta(W, Q)$, and in this special case we can detect in $O(m \log m)$ time whether $\delta(W, Q) \leq \kappa$. Hence, we obtain the following.

Corollary 2.5 *Let P be a polygonal chain with n vertices in $\mathbb{E}^2$. After $O(n)$ preprocessing, for a given subset W of vertices of P, a subchain Q of P, and a given parameter $\kappa \geq 1$, we can decide, in $O(m \log m)$ time, whether $\delta^*(W, Q) \leq \kappa$ or*

$\sigma(W, Q) \leq \kappa$, where $m = |W| + |Q|$. Moreover, if $\delta(W, Q) = \delta(P)$, then we can also detect in $O(m \log m)$ time whether $\delta(W, Q) \leq \kappa$.

2.3 Computing $\delta(P)$ and $\sigma(P)$

So far we have shown how to solve the decision problems associated with finding the detour and spanning ratio of a path. Now we apply a randomized technique of Chan [9], which does not affect the asymptotic running time of our decision algorithms, to compute the actual detour $\delta(P)$ or spanning ratio $\sigma(P)$. Suppose we have precomputed the weights of all vertices in P. Let W be a subset of vertices of P, and let Q be a subchain of P; set $m = |W| + |Q|$. We describe an algorithm that computes a pair $(\xi, \eta) \in W \times Q$ so that $\delta^*(W, Q) \leq \delta(\xi, \eta) \leq \delta(W, Q)$.

If $|W|$ or $|Q|$ is less than a prespecified constant, then we compute $\delta(W, Q)$ using a naive approach and report a pair (ξ, η) that attains it. Otherwise, we partition W into two subsets W_1, W_2 of roughly the same size, and partition Q into two subchains Q_1, Q_2 of roughly the same size. We have four subproblems (W_i, Q_j), $1 \leq i, j \leq 2$, at our hand. Note that

$$\delta(W, Q) = \max\{\delta(W_1, Q_1), \delta(W_2, Q_1), \delta(W_1, Q_2), \delta(W_2, Q_2)\}, \tag{1}$$

$$\delta^*(W, Q) \leq \max\{\delta^*(W_1, Q_1), \delta^*(W_2, Q_1), \delta^*(W_1, Q_2), \delta^*(W_2, Q_2)\}, \tag{2}$$

where (2) is an easy consequence of the visibility constraints in the definition of δ^*.

Following Chan's approach [9], we process the four subproblems in a random order and maintain a pair of points $(\xi, \eta) \in W \times Q$. Initially, we set (ξ, η) to be an arbitrary pair of points in $W \times Q$. While processing a subproblem (W_i, Q_j), for $1 \leq i, j \leq 2$, we first check in $O(m \log m)$ time whether $\delta^*(W_i, Q_j) > \delta(\xi, \eta)$, using Corollary 2.5. If the answer is yes, we solve the subproblem (W_i, Q_j) recursively and update the pair (ξ, η); otherwise, we ignore this subproblem. By (1), (2), and induction hypothesis, the algorithm returns a pair (ξ, η) such that $\delta^*(W, Q) \leq \delta(\xi, \eta) \leq \delta(W, Q)$. Moreover, if $\delta(W, Q) = \delta(P)$, then $\delta^*(W, Q) = \delta(W, Q)$, so the algorithm returns the value of $\delta(W, Q)$. Chan's analysis [9] (cf. proof of Lemma 2.1) shows that the expected running time of the algorithm on an input of size m is $O(m \log m)$. Hence, by invoking this algorithm on the pair (V, P), $\delta(V, P) = \delta(P)$ can be computed in $O(n \log n)$ expected time.

The case of the spanning ratio is handled in a similar and simpler manner, replacing (1) and (2) by

$$\sigma(W, Q) = \max\{\sigma(W_1, Q_1), \sigma(W_2, Q_1), \sigma(W_1, Q_2), \sigma(W_2, Q_2)\} \tag{3}$$

and applying Chan's technique using this relationship. Hence, we obtain the following main result of this section.

Theorem 2.6 *The detour or spanning ratio of a polygonal chain P with n vertices embedded in $\mathbb{E}^2$ can be computed in $O(n \log n)$ randomized expected time.*

Remark One can obtain an alternative *deterministic* solution that uses parametric search [22], and runs in time $O(n \log^c n)$, for some constant c. However, the resulting

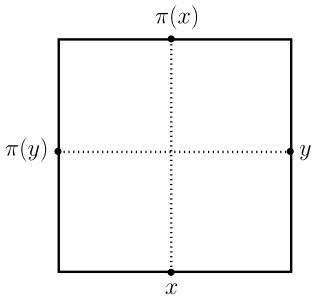


Fig. 2 *Dotted lines* indicate (the only two) pairs of points that attain the maximum detour

algorithm is considerably more involved on top of being slightly less efficient. We therefore omit its description.

We extend the definition of $\delta^*(\cdot,\cdot)$ to two disjoint subchains L and R of P as follows. Let V_L (resp. V_R) be the set of vertices in L (resp. R). Define $\delta^*(L,R) = \max\{\delta^*(V_L,R), \delta(V_R,L)\}$. Using the same argument as in the proof of Lemma 2.1, we can argue that if $\delta(L,R) = \delta(P)$, then $\delta(L,R) = \delta^*(L,R)$. The following corollary, which will be useful in the next section, is an obvious generalization of the above algorithm.

Corollary 2.7 *Let L and R be two disjoint subsets of a polygonal chain P in $\mathbb{E}^2$, with a total of n vertices, preprocessed to report weights in $O(1)$ time. Then $\sigma(L,R)$ can be computed in $O(n\log n)$ randomized expected time. We can also compute within the same time a pair $(p,q) \in L \times R$ such that $\delta^*(L,R) \le \delta(p,q) \le \delta(L,R)$. Moreover, if $\delta(L,R) = \delta(P)$, then $\delta(p,q) = \delta(L,R)$.*

As to lower bounds, it was shown by Narasimhan and Smid [23] that computing the spanning ratio of a planar polygonal chain requires $\Omega(n\log n)$ time if self-overlapping chains are allowed as input. Grüne [14] has shown that the same lower bound holds if the input is restricted to polygonal chains that are monotonic, hence simple. It is unknown whether the $\Omega(n\log n)$ lower bound also holds for computing the detour of a polygonal curve.

3 Planar Cycles and Trees

In this section we show that the tools developed for planar paths can be used for solving the detour and spanning ratio problems on more complicated graphs. Again, we consider only the problem of computing the detour, because the resulting algorithms can easily be adapted (and simplified) so as to compute the spanning ratio.

3.1 Polygonal Cycles in the Plane

Let us now consider the case in which $P = (V,E)$ is a closed (simple) polygonal curve. This case is more difficult because there are two paths along P between any two points of P. As a result, the detour of P might occur at a pair of points neither

of which is a vertex of P. For example, the detour in a unit square occurs at the midpoints of two opposite edges; in this case the lengths of the two paths between the points must be equal.

For two points $p, q \in P$, let $P[p, q]$ denote the subsets of P from p to q in counterclockwise direction. We use here the notation $d_P(p, q)$ to denote the length of $P[p, q]$; thus, in general, $d_P(p, q) \neq d_P(q, p)$ and $d_P(p, q) + d_P(q, p)$ is the length $|P|$ of the entire curve P. For a point $p \in P$, let $\pi(p)$ denote the point on P such that $d_P(p, \pi(p)) = d_P(\pi(p), p) = |P|/2$; obviously, $\pi(\pi(p)) = p$. Let P_p denote the polygonal chain $P[p, \pi(p)]$.

Lemma 3.1 *Let p be a point on P, and let A, B be two portions of P_p, then $\delta_P(A, B) = \delta_{P_p}(A, B)$.*

This follows from the fact that the shortest path along P between any two points $a, b \in A \times B$ is contained in the polygonal chain P_p.

Now the P-detour between two points $p, q \in P$ is defined as

$$\delta_P(p, q) = \frac{\min\{d_P(p, q), d_P(q, p)\}}{\|pq\|},$$

and the detour of the whole of P is defined as

$$\delta(P) = \max_{\substack{p,q \in P \\ p \neq q}} \delta_P(p, q).$$

Lemma 3.2 *The detour $\delta(P)$ of P is attained by a pair of points $p, q \in P$, such that either one of them is a vertex of P, or $q = \pi(p)$.*

Proof Suppose $\delta(P) = \delta_P(p, q)$, where neither p nor q is a vertex, and $q \neq \pi(p)$. Suppose $|P|/2 - d_P(p, q) = a > 0$. We extend, on either end, $P[p, q]$ by subpaths $P[p', p]$ and $P[q, q']$ of P, each of length $a/2$, and thereby obtain a polygonal subchain $P' = P[p', q'] \subset P$ of length $|P|/2$. Since a shortest path in P between any two points of P' is contained in P', we have

$$\delta(P) = \delta_P(p, q) \leq \delta(P') \leq \delta(P).$$

Thus, the maximum detour of P' is attained at p and q. By Lemma 2.1(ii), the detour does not change as we simultaneously move p toward p' and q toward q' at equal speed, along their edges in P'. This motion continues until one of the two points reaches a vertex of P'—which must be a vertex of P, too—or both endpoints $p', q' = \pi(p')$ of P' are reached. □

By using a rotating-caliper approach, we can compute $\max_{p \in P} \delta_P(p, \pi(p))$ in $O(n)$ time, so we focus on the case in which one of the points attaining the detour is a vertex of P. We present a different divide-and-conquer algorithm, which will use the algorithm described in Sect. 2.2 repeatedly. We can preprocess P in $O(n)$ time, so that, for any two points $p, q \in P$, $d_P(p, q)$ can be computed in $O(1)$ time.

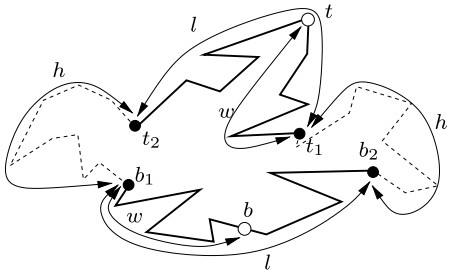


Fig. 3 An instance of the recursive problem; $d_P(t_1, t_2) = d_P(b_1, b_2) = l$, $d_P(t_2, b_1) = d_P(b_2, t_1) = h$, $|P| = 2(l + h)$, $d_P(t_1, t) = d_P(b_1, b) = w$

Let t_1, t_2, b_1, b_2 be four points of P appearing in this counterclockwise order along P, so that the following condition is satisfied.

$$b_1 = \pi(t_1) \quad \text{and} \quad b_2 = \pi(t_2). \tag{4}$$

We observe that condition (4) implies $d_P(t_1, t_2) = d_P(b_1, b_2)$ and $d_P(t_2, b_1) = d_P(b_2, t_1)$. Let m, m' be the number of edges in $P[b_1, b_2]$ and $P[t_1, t_2]$, respectively. Define

$$\rho(t_1, t_2, b_1, b_2) = \delta_P(P[t_1, t_2], P[b_1, b_2]).$$

We describe a recursive algorithm that computes a pair of points $(p, q) \in P[b_1, b_2] \times P[t_1, t_2]$ such that $\delta(p, q) = \rho(t_1, t_2, b_1, b_2)$ if $\rho(t_1, t_2, b_1, b_2) = \delta(P)$. If $\rho(t_1, t_2, b_1, b_2) < \delta(P)$, it returns an arbitrary pair of points in $P[b_1, b_2] \times P[t_1, t_2]$.

If $\min\{m, m'\} = 1$, then we can compute $\rho(t_1, t_2, b_1, b_2)$ in $O(m + m')$ time. Otherwise, suppose, without loss of generality, that $m' \geq m$, and let t be the middle vertex of $P[t_1, t_2]$ (i.e., the vertex for which each of $P[t_1, t]$, $P[t, t_2]$ has $m'/2$ edges), and let $b = \pi(t)$. It is easily seen that $b \in P[b_1, b_2]$ (by condition (4)). Clearly,

$$\rho(t_1, t_2, b_1, b_2) = \max\{\rho(t_1, t, b, b_2), \rho(t, t_2, b_1, b), \rho(t_1, t, b_1, b), \rho(t, t_2, b, b_2)\}.$$

Since $P[t_1, t]$ and $P[b, b_2]$ lie in $P[b, t] = P[\pi(t), t]$, using Corollary 2.7, we can compute in $O((m' + m)\log(m' + m))$ randomized expected time a pair $(p, q) \in P[t_1, t] \times P[b, b_2]$ so that $\delta(p, q) = \rho(t_1, t, b, b_2)$ if $\rho(t_1, t, b, b_2) = \delta(P)$. We can compute a similar pair in $P[t, t_2] \times P[b_1, b]$ within the same time bound. Each of the two 4-tuples (t_1, t, b_1, b) and (t, t_2, b, b_2) satisfies condition (4), and we solve the problem recursively for them. Among the pairs computed by the four subproblems, we return the one with the largest detour. The correctness of the algorithm is straightforward.

Let m_1 be the number of edges in $P[b_1, b]$. Then $P[b, b_2]$ contains at most $m - m_1 + 1$ edges. Let $T(m', m)$ denote the maximum expected time of computing $\rho(t_1, t_2, b_1, b_2)$, with the relevant parameters m' and m. Then we obtain the following recurrence:

$$T(m', m) \leq T\left(\frac{m'}{2}, m_1\right) + T\left(\frac{m'}{2}, m - m_1 + 1\right) + O((m' + m)\log(m' + m)),$$

for $m' \geq m$,

with a symmetric inequality for $m \geq m'$, and $T(m', 1) = O(m')$, $T(1, m) = O(m)$. The solution to the above recurrence is easily seen to be

$$T(m', m) = O((m' + m)\log^2(m' + m)).$$

Returning to the problem of computing $\delta(P)$, we choose a vertex $v \in P$. Let $P_1 = P[v, \pi(v)]$ and $P_2 = P[\pi(v), v]$. Then

$$\begin{aligned}\delta(P) &= \max\Big\{\max_{x,y\in P_1} \delta_P(x, y), \max_{x,y\in P_2} \delta_P(x, y), \delta_P(P_1, P_2)\Big\} \\ &= \max\{\delta(P_1), \delta(P_2), \rho(v, \pi(v), \pi(v), v)\}.\end{aligned}$$

The last equality follows from the fact that the 4-tuple $(v, \pi(v), \pi(v), v)$ satisfies (4). We can compute $\delta(P_1), \delta(P_2)$ in $O(n \log n)$ randomized expected time, using Theorem 2.6. Next we invoke the above algorithm on the 4-tuple $(v, \pi(v), \pi(v), v)$. We return the maximum of these values. If $\rho(v, \pi(v), \pi(v), v) = \delta(P)$, then the above recursive algorithm computes $\rho(v, \pi(v), \pi(v), v)$. Hence, the total expected time spent in computing $\delta(P)$ is $O(n \log^2 n)$.

The same method also applies to the computation of the spanning ratio of P, and we thus obtain:

Theorem 3.3 *The detour or spanning ratio of a polygonal cycle P with n edges in $\mathbb{E}^2$ can be computed in $O(n \log^2 n)$ randomized expected time.*

3.2 Planar Trees

Let $T = (V, E)$ be a tree embedded in $\mathbb{E}^2$. With a slight abuse of notation, we will use T to denote the embedding of the tree as well. We describe a randomized algorithm for computing $\delta(T)$. Without loss of generality, assume T is rooted at a vertex v_0 so that if we remove v_0 and the edges incident upon v_0, each component in the resulting forest has at most $n/2$ vertices; v_0 can be computed in linear time; refer to Fig. 4. We partition the children of v_0 into two sets A and B. Let T_A (resp., T_B), denote the tree induced by v_0 and all vertices having ancestors in A (resp., B). The partition A, B is chosen so that

$$\frac{1}{4}n \leq \|T_A\|, \|T_B\| \leq \frac{3}{4}n.$$

Since no descendent of v_0 is the root of a subtree with size more than $n/2$, such a partition can be found with a linear-time greedy algorithm.

We recursively compute $\delta(T_A)$ and $\delta(T_B)$. Let $\kappa^* = \max\{\delta(T_A), \delta(T_B)\}$. If $\delta(T_A, T_B) > \kappa^*$, then we need to compute $\delta(T_A, T_B)$. The following lemma, whose proof is identical to that of Lemma 2.1 given in [10], will be useful.

Lemma 3.4 *Let T_A and T_B be two subtrees of T, and let V_A (resp. V_B) be the set of vertices in T_A (resp. T_B). There exists a pair of points $(p, q) \in (V_A \times T_B) \cup (V_B \times T_A)$ such that $\delta(p, q) = \delta(T_A, T_B)$. Moreover, if $\delta(T_A, T_B) = \delta(T)$ then p is visible from q with respect to $T_A \cup T_B$.*

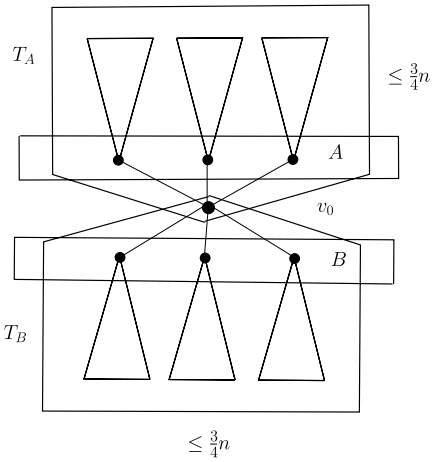


Fig. 4 Partitioning T into subtrees T_A and T_B

By Lemma 3.4, it suffices to compute $\delta(V_A, T_B)$ and $\delta(V_B, T_A)$, where V_A and V_B are the sets of vertices in T_A and T_B, respectively. As in Sect. 2, we first describe a decision algorithm that determines whether $\delta(T_A, T_B) \leq \kappa$ for some parameter $\kappa \geq \kappa^*$. We define the weight $\omega(p)$ of a point $p \in T$ to be

$$\omega(p) = \frac{d_T(p, v_0)}{\kappa}.$$

Let C be the cone $z = \sqrt{x^2 + y^2}$. To determine whether $\delta(V_A, T_B) \leq \kappa$, we map each point $u = (u_x, u_y) \in V_A$ to the cone $C_u = C + (u_x, u_y, -\omega(u))$, and map each point $v = (v_x, v_y) \in T_B$ to the point $\hat{v} = (v_x, v_y, \omega(v))$. Let $\hat{T}_B = \{\hat{v} \mid v \in T_B\}$ be the resulting tree embedded in $\mathbb{E}^3$. Following the same argument as in Lemma 2.2, we can argue that, for any $(u, v) \in V_A \times T_B$, $\delta(u, v) \leq \kappa$ if and only if $\hat{v}$ lies below the cone C_u. If $\delta(T_A, T_B) > \kappa \geq \kappa^*$, then $\delta(T_A, T_B) = \delta(T)$ and, by Lemma 3.4, there is a co-visible pair of points in $V_A \times T_B$ whose detour is greater than κ. So we can restrict our attention to co-visible pairs in $V_A \times T_B$. Using this observation and Lemma 3.4, we can determine whether $\delta(V_A, T_B) \leq \kappa$, in $O(n \log n)$ time, by the same approach as in Sect. 2. Similarly, we can determine whether $\delta(V_B, T_A) \leq \kappa$ in $O(n \log n)$ time.

Finally, returning to the problem of computing $\delta(T)$, we first use the decision algorithm to determine whether $\delta(T_A, T_B) > \kappa^*$. If the answer is no, we return κ^* and a pair of points, both from T_A or both from T_B, realizing this detour. Otherwise, $\delta(T) = \delta(T_A, T_B)$. Since each of T_A, T_B can be decomposed into two subtrees, each of size at most $3/4$ the size of T_A or T_B, respectively, we can plug this decision algorithm into Chan's technique, with the same twist as in Sect. 2, to obtain an algorithm that computes $\delta(V_A, T_B)$ in $O(n \log n)$ randomized expected time.

Putting everything together, the expected running time of the above algorithm is given by the recurrence

$$T(n) = T(n - k + 1) + T(k) + O(n \log n),$$

with $n/4 \leq k \leq 3n/4$. The recurrence solves to $O(n \log^2 n)$. (As in the case of chains, we need one preliminary global pass that computes the distances along T from v_0 to each of the vertices.)

The algorithm for computing the spanning ratio proceeds in a similar but simpler manner, as in the case of chains, and has the same randomized expected running time bound. We thus conclude the following.

Theorem 3.5 *The detour or spanning ratio of a planar tree with n vertices can be computed in $O(n \log^2 n)$ randomized expected time.*

4 Polygonal Chains, Cycles, and Trees in $\mathbb{E}^3$

Let P be a polygonal chain with n vertices embedded in $\mathbb{E}^3$. We describe subquadratic algorithms for computing the detour and spanning ratio of P, and a reduction showing that the problem of computing the detour is at least as hard as Hopcroft's problem.

4.1 Computing the Spanning Ratio

We begin with the simpler problem of computing the spanning ratio $\sigma(P)$ of P. We solve this problem by adapting the technique for computing spanning ratios in the plane, as described in Sect. 2. Specifically, consider the decision problem, where we want to determine whether $\sigma(P) \leq \kappa$. We take the set V of vertices of P, and map each $p \in V$ to the point $\hat{p} = (p, \omega(p)) \in \mathbb{R}^4$, where $\omega(p) = d_P(p_0, p)/\kappa$ and p_0 is the starting point of P. We take the cone

$$C : x_4 = \sqrt{x_1^2 + x_2^2 + x_3^2},$$

and define, for each $p \in V$, the cone C_p to be $\hat{p} + C$. As in the planar case, $\sigma(P) \leq \kappa$ if and only if each point $\hat{p}$, for $p \in V$, lies *on* the lower envelope of $\mathfrak{C} = \{C_q \mid q \in V\}$.

Let $p = (a_1, a_2, a_3)$ be a point in V, and let $\omega(p) = a_4$. A point $\xi = (\xi_1, \xi_2, \xi_3, \xi_4)$ lies below the cone

$$C_p : x_4 - a_4 = \sqrt{(x_1 - a_1)^2 + (x_2 - a_2)^2 + (x_3 - a_3)^2}$$

if and only if the point

$$\varphi(\xi) = (\xi_1, \xi_2, \xi_3, \xi_4, \xi_4^2 - \xi_1^2 - \xi_2^2 - \xi_3^2)$$

in $\mathbb{E}^5$ lies in the halfspace

$$h_p : x_5 \leq -2a_1x_1 - 2a_2x_2 - 2a_3x_3 + 2a_4x_4 + (a_1^2 + a_2^2 + a_3^2 - a_4^2).$$

Therefore a point $\xi \in \mathbb{E}^4$ lies in the lower envelope of $\mathfrak{C}$ if and only if $\varphi(\xi)$ lies in the convex polyhedron $\bigcap_{p \in V} h_p$. Hence, the problem of determining whether $\sigma(P) \leq \kappa$ reduces to locating n points in a 5-dimensional convex polyhedron defined by the intersection of n halfspaces. This problem can be solved in $O(n^{4/3+\varepsilon})$ time using a data structure for halfspace-emptiness queries [1]. Using Chan's technique, as in the planar case, we can compute $\sigma(P)$ itself within the same asymptotic time bound. Finally, as for the planar case, the algorithm can be extended to compute the spanning ratio of a polygonal cycle or tree embedded in $\mathbb{E}^3$. That is, we have shown:

Theorem 4.1 *The spanning ratio of a polygonal chain, cycle, or tree with n vertices embedded in* $\mathbb{E}^3$ *can be computed in randomized expected time* $O(n^{4/3+\varepsilon})$, *for any* $\varepsilon > 0$.

4.2 Computing the Detour

We next consider the problem of computing the detour $\delta(P)$ of P. Here the algorithm becomes considerably more involved and less efficient, albeit still subquadratic. As in some of the preceding algorithms, we use a divide-and-conquer approach to compute $\delta(P)$. That is, we partition P into two connected portions, P_1, P_2, each consisting of $n/2$ edges, recursively compute $\delta(P_1)$ and $\delta(P_2)$, and then compute explicitly the detour between P_1 and P_2, as follows. Let o be the common endpoint of P_1 and P_2. For any point x in P, let $\omega(x) = d_P(o, x)$ be the arc length of P (that is, either of P_1 or of P_2) between o and x. For any $x \in P_1$, $y \in P_2$, we have

$$\delta_P(x, y) = \frac{\omega(x) + \omega(y)}{\|xy\|}.$$

For a pair of edges $e \in P_1$ and $e' \in P_2$, define, as above,

$$\delta(e, e') = \delta_P(e, e') = \max_{x \in e, x' \in e'} \delta_P(x, x');$$

as in Sect. 2, we drop the subscript P in the function δ. Then

$$\delta(P) = \max\left\{\delta(P_1), \delta(P_2), \max_{e \in P_1, e' \in P_2} \delta(e, e')\right\}.$$

Let A, B denote the set of edges of P_1 and P_2, respectively. It suffices to compute the third term,

$$\delta(A, B) = \max_{a \in A, b \in B} \delta(a, b).$$

Unlike the planar case, the detour of P is not necessarily attained at a vertex of P (for example, there P might contain two long edges that orthogonally pass near each other at a very small distance, and the detour could then be obtained between the two points that realize the distance between the segments). This makes the 3-dimensional algorithm considerably more complicated, and less efficient, than its 2-dimensional counterpart. Consider first the decision problem, in which we wish to determine whether $\delta(A, B) \leq \kappa$, for some given $\kappa \geq 1$.

For an edge $e \in A \cup B$, let e^+ denote the ray that emanates from the endpoint, z^+, of e closer to o along P and that contains e; see Fig. 5. Similarly, let e^- denote the ray emanating from the point z^- of e farther from o and containing e. We extend the definition of $\omega(\cdot)$ for points on the rays e^+, e^- even though these points might not lie on P. For a point $x \in e^+$ (resp., $x \in e^-$), we define $\omega(x) = \omega(z^+) + \|z^+x\|$ (resp., $\omega(x) = \omega(z^-) - \|xz^-\|$). Note that these definitions of ω are consistent with the earlier definition, in the sense that all of them assume the same value for the points on e. We can now define $\delta(\cdot, \cdot)$ for points lying on the rays supporting the edges of P_1 and P_2. Namely, for a given pair a, b, where a, b are either edges of P or the rays supporting the edges, $\delta(a, b) = \max_{x \in a, y \in b}(\omega(x) + \omega(y))/\|xy\|$.

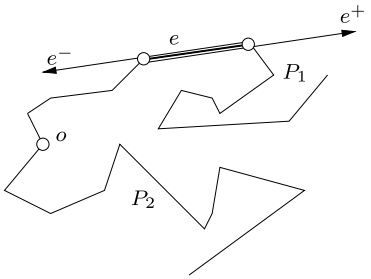


Fig. 5 Decomposition of P and rays e^+, e^-

Lemma 4.2 *Let $a \in A$ and $b \in B$ be a pair of edges. The following four conditions are equivalent*:

(i) $\delta(a, b) > \kappa$;
(ii) $\delta(a^+, b) > \kappa$ *and* $\delta(a^-, b) > \kappa$;
(iii) $\delta(a, b^+) > \kappa$ *and* $\delta(a, b^-) > \kappa$;
(iv) $\delta(a^+, b^+) > \kappa$, $\delta(a^+, b^-) > \kappa$, $\delta(a^-, b^+) > \kappa$, *and* $\delta(a^-, b^-) > \kappa$.

Proof Let a^* (resp., b^*) be the line supporting the edge a (resp., b) oriented in the direction of the ray a^+ (resp., b^+). Parametrize the lines a^* and b^* by the signed distances along these lines from appropriate respective initial points $\xi \in a, \eta \in b$, and denote these distances by t and s, respectively. Regard $a^* \times b^*$ as the parametric ts-plane. Let u, v denote the positively oriented unit vectors along a^* and b^*, respectively. For $x = \xi + tu \in a^*$ and $y = \eta + sv \in b^*$, the condition $\delta(x, y) > \kappa$ can be written as:

$$\delta(x, y) = \frac{\omega(\xi) + \omega(\eta) + t + s}{\|(\xi - \eta) + tu - sv\|} > \kappa,$$

or

$$\kappa\|(\xi - \eta) + tu - sv\| - \omega(\xi) - \omega(\eta) - t - s < 0. \tag{5}$$

The left-hand side of (5) is a *convex* function on the st-parametric plane, being the difference of a convex function and a linear function. The lemma is then an easy consequence of this convexity property. Indeed (i) implies (ii)–(iv) because $a = a^+ \cap a^-$ and $b = b^+ \cap b^-$. For the converse implications, consider the implication (ii) $\Rightarrow$ (i). Suppose that $\delta(x^+, y^+) > \kappa$ for $x^+ \in a^+$, $y^+ \in b$ and $\delta(x^-, y^-) > \kappa$ for $x^- \in a^-$, $y^- \in b$. By construction, $x^+x^- \cap a \neq \emptyset$. Moreover, by convexity of (5), $\delta(x', y') > \kappa$ for all $x' \in x^+x^-$, $y' \in y^+y^-$, thereby implying that $\delta(a, b) > \kappa$. Similar arguments imply that (iii) or (iv) implies (i). □

Using Lemma 4.2(iv) and the standard random-sampling technique [16], we construct a four-level data structure to decide whether $\delta(A, B) > \kappa$. The first level constructs a complete bipartite decomposition for the set $\{(a, b) \in A \times B \mid \delta(a^+, b^+) > \kappa\}$. The second level processes each bipartite clique $A_i \times B_i$ in the decomposition, and represents the set $\{(a, b) \in A_i \times B_i \mid \delta(a^-, b^+) > \kappa\}$ as the union of complete bipartite subgraphs. The third level then refines further this decomposition, to collect pairs that also satisfy $\delta(a^+, b^-) > \kappa$, and the fourth level finally tests whether $\delta(a^-, b^-) > \kappa$ for any of the surviving pairs.

We compute the first-level decomposition of $\{(a, b) \in A \times B \mid \delta(a^+, b^+) > \kappa\}$, as follows. (Similar procedures are then applied at each of the three other levels of the data structure.) For each edge $a \in A$, we map the ray a^+ to a point $\zeta(a^+) = (\zeta_1, \ldots, \zeta_6)$ in $\mathbb{R}^6$, where $(\zeta_1, \zeta_2, \zeta_3)$ are the coordinates of the endpoint z^+ of a^+, (ζ_4, ζ_5) is an appropriate parametrization of the orientation of a^+, and $\zeta_6 = \omega(z^+)$. A similar parametrization will be used for the rays a^-. Next, we map each edge $b \in B$ to a surface $\gamma(b^+)$ that represents the locus of all rays a^+ for which $\delta(a^+, b^+) = \kappa$. Since δ increases as the parameter ζ_6 increases and each 5-tuple $(\zeta_1, \ldots, \zeta_5)$ defines a unique ray in $\mathbb{E}^3$, it follows that $\gamma(b^+)$ is the graph of a totally defined 5-variate function and $\delta(a^+, b^+) > \kappa$ (resp., $\delta(a^+, b^+) < \kappa$) if and only if $\zeta(a^+)$ lies above (resp., below) $\gamma(b^+)$. We can thus regard the problem at hand as that of collecting, in compact form, all pairs $(\zeta(a^+), \gamma(b^+))$ for which $\zeta(a^+)$ lies above $\gamma(b^+)$. Abusing the notation slightly, set $|A| = n$ and $|B| = m$.

We fix a sufficiently large constant r, draw a random sample R of $cr \log r$ edges of B, where c is a sufficiently large constant independent of r, and compute the vertical decomposition $\mathfrak{A}^{\|}$ of the arrangement $\mathfrak{A}$ of the surfaces $\{\gamma(b^+) \mid b \in R\}$. It is easily verified that these surfaces are all semi-algebraic of constant description complexity. Hence, we can apply the result of Koltun [19], to conclude that $\mathfrak{A}^{\|}$ has $O(r^{8+\varepsilon})$ cells, for any $\varepsilon > 0$. For each cell $\tau \in \mathfrak{A}^{\|}$, let $A_\tau = \{e \in A \mid \zeta(e^+) \in \tau\}$, let $B_\tau \subseteq B$ be the set of edges b for which the surface $\gamma(b^+)$ crosses τ, and let $B^*_\tau \subseteq B$ be the set of edges b for which the surface $\gamma(b^+)$ lies completely below τ. The sets A_τ, B_τ can be computed in $O(m + n)$ time under an appropriate model of computation, in which we assume that the roots of a constant degree polynomial can be computed in $O(1)$ time; see [25].

Set $n_\tau = |A_\tau|$ and $m_\tau = |B_\tau|$. Obviously, $\sum_\tau n_\tau = n$ and $|B^*_\tau| \leq m$. By the theory of random sampling [16, 25] (where we use the fact that the VC-dimension of the underlying range space is finite), $m_\tau \leq m/r$ for all τ, with probability at least $1 - \eta$, where $\eta = \eta(r)$ is a constant that can be made arbitrarily small by choosing the value of r sufficiently large. If $m_\tau > m/r$ for a cell, we choose another random sample and restart the above step. Since the probability of this event is a sufficiently small constant, it does not affect the asymptotic expected running time of the algorithm and we can ignore this step. Moreover, by splitting the cells into subcells, if needed, we may also assume that $n_\tau \leq n/r^8$ for each τ; the number of cells remains $O(r^{8+\varepsilon})$. By construction, $\delta(a^+, b^+) > \kappa$ for any pair $e \in A_\tau$ and $b \in B^*_\tau$. We use the second-level data structure, sketched below, to determine whether $\delta(A_\tau, B^*_\tau) > \kappa$. If m_τ or n_τ is less than a prespecified constant, then we use a naive procedure to determine whether $\delta(A_\tau, B_\tau) > \kappa$. Otherwise, we recursively determine (using the first-level data structure) whether $\delta(A_\tau, B_\tau) > \kappa$. For an edge $a \in A_\tau$ and for an edge $b \in B$ such that $\gamma(b^+)$ lies above τ, $\delta(a^+, b^+) < \kappa$, so there is no need to compare A_τ with such edges.

To exploit the symmetry in the condition $\delta(a^+, b^+) > \kappa$ between A and B, we next switch the roles of A_τ and B_τ, by mapping the rays b^+, for $b \in B_\tau$, to points in $\mathbb{R}^6$, and the rays a^+, for $a \in A_\tau$, to surfaces $\gamma(a^+)$, as above. We take a random sample of $cr \log r$ of these surfaces, and construct the vertical decomposition of their arrangement, as above. Repeating this for each cell τ, we end up with $O(r^{16+\varepsilon})$ subproblems, each involving at most n/r^9 segments of A and at most m/r^9 segments of B, which we proceed to solve recursively, using the first-level data structure. In

addition, we have subproblems involving pairs of sets of the form A_τ, B^*_τ, or $B_{\tau'}$, $A^*_{\tau'}$, which we pass to the second level of the structure.

The second-level structure is constructed in an analogous manner, with the only difference that we use the rays a^- instead of the rays a^+. Thus, starting with a pair of subsets A_τ, B_τ, we obtain a decomposition into $O(r^{16+\varepsilon})$ subproblems, each involving at most $|A_\tau|/r^9$ segments of A_τ and at most $|B_\tau|/r^9$ segments of B_τ, which we process recursively using the second-level structure, and a collection of other subproblems that we pass to the third level. The third level is again constructed in complete analogy, using the rays a^+ for the segments in A and the rays b^- for the segments in B. The fourth-level structure is constructed for the rays a^-, b^-, and is a little simpler than the preceding levels, in the sense that whenever we detect a cell that lies fully below a surface ($\gamma(a^-)$ or $\gamma(b^-)$), we stop and report that $\delta(A, B) > \kappa$. Otherwise, we continue the processing recursively, as in the preceding levels.

For $i = 1, \ldots, 4$ and for integers $m, n > 0$, let $T^{(i)}(n, m)$ denote the maximum running time of the ith level data structure on a set of n edges of P_1 and a set of m edges of P_2. Then

$$T^{(4)}(n, m) = O(r^{16+\varepsilon}) \cdot T^{(4)}\left(\frac{n}{r^9}, \frac{m}{r^9}\right) + O(m + n),$$

and

$$T^{(i)}(n, m) = O(r^{16+\varepsilon}) \cdot \left[T^{(i)}\left(\frac{n}{r^9}, \frac{m}{r^9}\right) + T^{(i+1)}(n, m)\right] + O(m + n),$$

for $i \leq 3$. The solutions to the above recurrences are easily seen to be $T^{(i)}(n, m) = O((mn)^{8/9+\varepsilon})$, for any $\varepsilon > 0$ and for each i.

Hence, we obtain the following.

Lemma 4.3 *Given a polygonal chain in $\mathbb{E}^3$, two disjoint subchains A and B of P with a total of m vertices, and a parameter $\kappa \geq 1$, we can determine, in $O(n^{16/9+\varepsilon})$ randomized expected time, whether $\delta(A, B) > \kappa$.*

As in the planar case, we can use the randomized technique of Chan [9] to compute the actual $\delta(A, B)$ within the same asymptotic expected running time bound. The algorithm extends to polygonal cycles and trees in $\mathbb{E}^3$.

In conclusion, we obtain the following.

Theorem 4.4 *The detour of a polygonal chain, cycle, or tree with n edges in $\mathbb{E}^3$ can be computed in randomized expected time $O(n^{16/9+\varepsilon})$, for any $\varepsilon > 0$.*

Remark We remark that it is also possible to use the parametric search technique [22], as in [3], to obtain a deterministic alternative solution. This however (a) results in a considerably more involved algorithm, and (b) requires us to derandomize the decision algorithm, i.e., its vertical decomposition step. This too is doable, but is considerably more complicated.

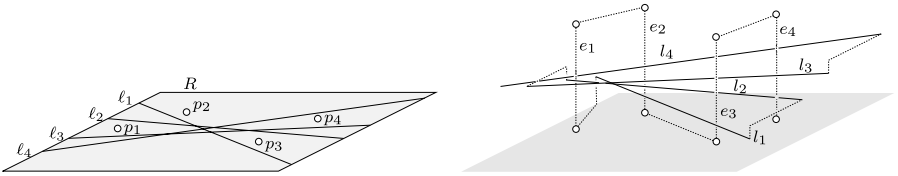


Fig. 6 Reducing Hopcrofts's problem to computing the detour of a 3-dimensional path. (**i**) An instance of Hopcroft's problem. (**ii**) Construction of the polygonal chain Π

4.3 Lower Bound

Finally, we show that computing the detour of a 3-dimensional path is as hard as Hopcroft's problem: Given a set $L = \{\ell_1, \ldots, \ell_n\}$ of n lines in $\mathbb{R}^2$ and a set $P = \{p_1, \ldots, p_n\}$ of n points in $\mathbb{R}^2$, determine whether any line of L contains any point of P. There is an abundance of evidence to suggest that Hopcroft's problem has an $\Omega(n^{4/3})$ lower bound [12]. The best known upper bound in any reasonable model of computation is $O(n^{4/3}2^{O(\log^* n)})$ [21].

To reduce an instance of Hopcroft's problem to that of computing the detour of a 3-dimensional path, we will first build a 3-dimensional path Π that is self-intersecting, i.e., has infinite detour, if and only if the answer to Hopcroft's problem is affirmative. Then we show how the proof can be modified to cover the case where we know *a priori* that the polygonal chains we are given as input do not self-intersect. The construction uses techniques presented in Erickson [12].

Without loss of generality, we may assume that none of the given lines is y-vertical. We begin by sorting the lines in L in increasing order of their slopes and the points in P in increasing lexicographic order. Let $\langle \ell_1, \ldots, \ell_n \rangle$ be the resulting sequence of lines, and let $\langle p_1, \ldots, p_n \rangle$ be the resulting sequence of points. We compute a bounding rectangle R so that each line of L intersects the two y-vertical edges of R, and all the points of P, as well as all the intersection points of lines in L, lie inside R. These steps require $O(n \log n)$ time.

By construction, the ordering of L along the left edge of R in $-y$-direction is $\ell_1, \ldots, \ell_n$, and its ordering along the right edge of R is $\ell_n, \ldots, \ell_1$. For each $1 \le i \le n$, we lift the segment $R \cap \ell_i$ orthogonally to the plane $z = i$, to obtain a line segment l_i. Next, we transform each input point $p_j \in P$ to a line segment e_j that is parallel to the z-axis, whose endpoints are $(p_j, 0)$ and $(p_j, n + 1)$; see Fig. 6.

This gives us a set of line segments so that the answer to Hopcroft's problem for the original lines and points is "yes" if and only if some segment l_i intersects some segment e_j. It remains to construct a polygonal chain that contains all these segments without introducing any additional crossings. To do this, we first form a chain containing all segments l_j. It starts at the left endpoint of l_1. The right endpoint of l_1 is connected to the right endpoint of l_2. This connection consists of two segments; the first one is parallel to the z-axis and leads from the plane $z = 1$ to the plane $z = 2$, and the second one, contained in $z = 2$, is parallel to the y-axis. Next, l_2 is traversed, and its left endpoint is connected to the left endpoint of l_3 in an analogous way. We continue until the last endpoint of l_n is reached. Clearly, the resulting chain is simple.

Next, we connect the segments $e_1, \ldots, e_n$ into a simple polygonal chain by connecting the upper endpoints of e_i to e_{i+1} if i is odd and the lower endpoints if i is even. This chain is clearly not self-intersecting since its xy-projection is monotone in the lexicographic order. Finally, we connect the left endpoint of l_1 in $z = 1$ to the free endpoint of e_1 in $z = 0$ by two additional segments. The resulting concatenation of the two chains has the desired property. See Fig. 6.

One might state the problem of computing the detour of a 3-dimensional chain in such a way that the input chains are known apriori not to have self-intersections. The above lower bound proof can be adapted to this situation in the following way. First, we move each of the original lines ℓ_i a distance of ϵ to the right, where ϵ is a formal infinitesimal, i. e., ϵ is positive, but smaller than any real number. Then we construct the polygonal chain in the same way as before. It will always be non-intersecting, but its detour is bigger than c/ϵ, for some appropriate constant $c > 0$, if and only if there was a point-line incidence in the original instance of Hopcroft's problem. Reductions using infinitesimals were formally shown to be correct, in the algebraic decision tree model, by Erickson [12].

In conclusion, we have shown:

Theorem 4.5 *An algorithm with running time $f(n)$ for computing the detour of 3-dimensional polygonal chains with n vertices implies an $O(n \log n + f(n))$ time algorithm for Hopcroft's problem.*

Remark It is interesting to note that we have almost matched this lower bound with the algorithm in Theorem 4.1 for computing the spanning ratio of P. We do not know whether the preceding construction can be extended to yield a lower bound argument for computing spanning ratios.

5 Conclusions

We have given $O(n \log n)$-time randomized algorithms for computing the detour and spanning ratio of planar polygonal chains. These algorithms lead to an $O(n \log^2 n)$-time algorithms for computing the detour and spanning ratio of planar trees and cycles. In three dimensions, we have given subquadratic algorithms for computing the detour and spanning ratio of polygonal chains, cycles, and trees. Previously, no subquadratic-time (exact) algorithms were known for any of these problems.

There are many open problems in this new area. The most obvious is: Which other classes of graphs admit subquadratic-time algorithms for computing their detour or spanning ratio? Also, it remains open to prove an $\Omega(n \log n)$ lower bound for computing the detour of a simple planar polygonal chain of n vertices; at present, such a bound is only known for computing the spanning ratio. Finally, it seems likely that the algorithm for computing the detour in $\mathbb{E}^3$ can be improved.

Acknowledgement We would like to thank Günter Rote for interesting discussions related to the problems studied in the paper.

References

1. Agarwal, P.K., Erickson, J.: Geometric range searching and its relatives. In: Chazelle, B., Goodman, J.E., Pollack, R. (eds.) Advances in Discrete and Computational Geometry. Contemporary Mathematics, vol. 223, pp. 1–56. American Mathematical Society, Providence (1999)
2. Agarwal, P.K., Klein, R., Knauer, C., Sharir, M.: Computing the detour of polygonal curves. Technical Report B 02-03, Freie Universität Berlin, Fachbereich Mathematik und Informatik (2002)
3. Agarwal, P.K., Sharir, M., Toledo, S.: Applications of parametric searching in geometric optimization. J. Algorithms **17**, 292–318 (1994)
4. Aichholzer, O., Aurenhammer, F., Icking, C., Klein, R., Langetepe, E., Rote, G.: Generalized self-approaching curves. Discrete Appl. Math. **109**, 3–24 (2001)
5. Alt, H., Guibas, L.J.: Discrete geometric shapes: Matching, interpolation, and approximation. In: Sack, J.-R., Urrutia, J. (eds.) Handbook of Computational Geometry, pp. 121–153. Elsevier, Amsterdam (2000)
6. Alt, H., Knauer, C., Wenk, C.: Comparison of distance measures for planar curves. Algorithmica **38**, 45–58 (2004)
7. Aurenhammer, F., Klein, R.: Voronoi diagrams. In: Sack, J.-R., Urrutia, J. (eds.) Handbook of Computational Geometry, pp. 201–290. Elsevier, Amsterdam (2000)
8. Bose, P., Morin, P.: Competitive online routing in geometric graphs. Theor. Comput. Sci. **324**, 273–288 (2004)
9. Chan, T.M.: Geometric applications of a randomized optimization technique. Discrete Comput. Geom. **22**(4), 547–567 (1999)
10. Ebbers-Baumann, A., Klein, R., Langetepe, E., Lingas, A.: A fast algorithm for approximating the detour of a polygonal chain. Comput. Geom. Theory Appl. **27**, 123–134 (2004)
11. Edelsbrunner, H., Guibas, L.J., Sharir, M.: The complexity and construction of many faces in arrangements of lines and of segments. Discrete Comput. Geom. **5**, 161–196 (1990)
12. Erickson, J.: New lower bounds for Hopcroft's problem. Discrete Comput. Geom. **16**, 389–418 (1996)
13. Fortune, S.J.: A sweepline algorithm for Voronoi diagrams. Algorithmica **2**, 153–174 (1987)
14. Grüne, A.: Umwege in Polygonen. Master's thesis, Institut für Informatik I, Universität Bonn (2002)
15. Guibas, L.J., Sharir, M., Sifrony, S.: On the general motion planning problem with two degrees of freedom. Discrete Comput. Geom. **4**, 491–521 (1989)
16. Haussler, D., Welzl, E.: Epsilon-nets and simplex range queries. Discrete Comput. Geom. **2**, 127–151 (1987)
17. Icking, C., Klein, R.: Searching for the kernel of a polygon: a competitive strategy. In: Proceedings of the 11th Annual Symposium on Computational Geometry, pp. 258–266 (1995)
18. Icking, C., Klein, R., Langetepe, E.: Self-approaching curves. Math. Proc. Camb. Philos. Soc. **125**, 441–453 (1999)
19. Koltun, V.: Almost tight upper bounds for vertical decompositions in four dimensions. J. ACM **51**, 699–730 (2004)
20. Langerman, S., Morin, P., Soss, M.: Computing the maximum detour and spanning ratio of planar chains, trees and cycles. In: Proceedings of the 19th International Symposium on Theoretical Aspects of Computer Science (STACS 2002). Lecture Notes in Computer Science, vol. 2285, pp. 250–261. Springer, Berlin (2002)
21. Matoušek, J.: Range searching with efficient hierarchical cuttings. Discrete Comput. Geom. **10**(2), 157–182 (1993)
22. Megiddo, N.: Applying parallel computation algorithms in the design of serial algorithms. J. ACM **30**(4), 852–865 (1983)
23. Narasimhan, G., Smid, M.: Approximating the stretch factor of Euclidean graphs. SIAM J. Comput. **30**(3), 978–989 (2000)
24. Rote, G.: Curves with increasing chords. Math. Proc. Camb. Philos. Soc. **115**, 1–12 (1994)
25. Sharir, M., Agarwal, P.K.: Davenport-Schinzel Sequences and Their Geometric Applications. Cambridge University Press, New York (1995)

Robust Shape Fitting via Peeling and Grating Coresets

Pankaj K. Agarwal · Sariel Har-Peled · Hai Yu

Abstract Let P be a set of n points in $\mathbb{R}^d$. A subset $\mathcal{S}$ of P is called a (k, ε)-*kernel* if for every direction, the directional width of $\mathcal{S}$ ε-approximates that of P, when k "outliers" can be ignored in that direction. We show that a (k, ε)-kernel of P of size $O(k/\varepsilon^{(d-1)/2})$ can be computed in time $O(n + k^2/\varepsilon^{d-1})$. The new algorithm works by repeatedly "peeling" away $(0, \varepsilon)$-kernels from the point set.

We also present a simple ε-approximation algorithm for fitting various shapes through a set of points with at most k outliers. The algorithm is incremental and works by repeatedly "grating" critical points into a working set, till the working set provides the required approximation. We prove that the size of the working set is independent of n, and thus results in a simple and practical, near-linear ε-approximation algorithm for shape fitting with outliers in low dimensions.

We demonstrate the practicality of our algorithms by showing their empirical performance on various inputs and problems.

Keywords Shape fitting · Coresets · Geometric approximation algorithms

A preliminary version of this paper appeared in *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 182–191. P.A. and H.Y. are supported by NSF under grants CCR-00-86013, EIA-01-31905, CCR-02-04118, and DEB-04-25465, by ARO grants W911NF-04-1-0278 and DAAD19-03-1-0352, and by a grant from the U.S.–Israel Binational Science Foundation. S.H.-P. is supported by a NSF CAREER award CCR-0132901.

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

H. Yu
e-mail: fishhai@cs.duke.edu

S. Har-Peled
Department of Computer Science, University of Illinois, Urbana, IL 61801, USA
e-mail: sariel@cs.uiuc.edu

1 Introduction

In many areas such as computational geometry, computer graphics, machine learning, and data mining, considerable work has been done on computing descriptors of the extent of a set P of n points in $\mathbb{R}^d$. These descriptors, called *extent measures*, either compute certain statistics of P itself such as diameter and width, or compute some geometric shape enclosing P with respect to a certain optimization criterion, such as computing the smallest radius of a sphere or cylinder, the minimum volume or surface area of a box, and the smallest spherical or cylindrical shell that contain P. Motivated by more recent applications, there has also been work on maintaining extent measures of a set of moving points, e.g., using the kinetic data structure framework [6, 11].

The existing exact algorithms for computing extent measures are generally expensive. For example, the best known algorithm for computing the smallest enclosing cylindrical shell in $\mathbb{R}^3$ requires $O(n^5)$ time [5]. Consequently, attention has shifted to developing faster approximation algorithms; see, e.g., [4, 5, 10, 12]. Agarwal et al. [7] proposed a unified framework for computing numerous extent measures approximately in low dimensions. Their approach is to first extract a small subset from the input, known as a *coreset*, and then return the extent measure of this subset as an approximation to that of the original input. The running time of their algorithm, substantially improving upon many previous results, is typically of the form $O(n+1/\varepsilon^c)$, where n is the input size, c is a constant that may depend on the dimension d, and ε is the approximation error.

Most of the existing work assumes that the input does not contain noisy data. However in the real world, noise may come from different sources during data acquisition, transmission, storage and processing, and is unavoidable in general. Meanwhile, most extent measures are very sensitive to noise; a small number of inaccurate data points (i.e., the so-called *outliers*) may substantially affect extent measures of the entire input. In order to compute more reliable extent measures on the input, it is thus natural to require that the outliers should be excluded from consideration. For example, the smallest enclosing cylinder problem with k outliers is formulated as finding the smallest cylinder that covers all but at most k of the input points.

Following up the work in [7, 19], we consider the problem of finding *robust coresets* for various extent measures that are able to handle outliers. Assuming there are at most k outliers in the input, our goal is to compute a coreset of small size, so that the best solution on the coreset with at most k outliers would provide an ε-approximation to the original input with at most k outliers. We are mainly concerned with the case in which the number k of outliers is small compared to the input size n. Otherwise, random-sampling techniques have been effective in handling outliers [9].

Problem Statement Let P be a set of n points in $\mathbb{R}^d$. For a direction $u \in \mathbb{S}^{d-1}$ and an integer $0 \leq k < n$, the *level* of a point $a \in \mathbb{R}^d$ in direction u is the size of the set $\{p \in P \mid \langle u, p\rangle > \langle u, a\rangle\}$, i.e., the number of points in P that lie in the open halfspace $\langle u, x - a\rangle > 0$. This notion of level is the dual of the level of a point in an arrangement of hyperplanes [23]. In this paper, when we refer to a direction $u \in \mathbb{S}^{d-1}$, we always assume for the sake of simplicity that no two points in P lie on the same level in that direction; more careful but similar arguments would work for directions that do not satisfy this assumption.

Let $P^k[u]$ (resp. $P_k[u]$) denote the point of P whose level is k (resp. $n-k-1$) in a direction $u \in \mathbb{S}^{d-1}$. Let $\mathbf{U}_k(u, P) = \langle u, P^k[u] \rangle$ denote the *k-level of P* in direction u. Let $\mathbf{L}_k(u, P) = \langle u, P_k[u] \rangle = -\mathbf{U}_k(-u, P)$. For parameters k and ℓ, the *(k, ℓ)-directional width* of P in direction u, denoted by $\mathcal{E}_{k,\ell}(u, P)$, is defined as

$$\mathcal{E}_{k,\ell}(u, P) = \mathbf{U}_k(u, P) - \mathbf{L}_\ell(u, P).$$

For simplicity, we denote $\mathcal{E}_{k,k}(u, P)$ by $\mathcal{E}_k(u, P)$ and $\mathcal{E}_0(u, P)$ by $\mathcal{E}(u, P)$. Similarly, we denote $\mathbf{U}_0(u, P)$ by $\mathbf{U}(u, P)$ and $\mathbf{L}_0(u, P)$ by $\mathbf{L}(u, P)$ respectively.

Given a set P of n points in $\mathbb{R}^d$, a parameter $\varepsilon > 0$ and an integer $0 \leq k < n/2$, a subset $\mathcal{S} \subseteq P$ is called a *(k, ε)-kernel* of P if for every $u \in \mathbb{S}^{d-1}$ and every $0 \leq a, b \leq k$,

$$(1 - \varepsilon) \cdot \mathcal{E}_{a,b}(u, P) \leq \mathcal{E}_{a,b}(u, \mathcal{S}) \leq \mathcal{E}_{a,b}(u, P).$$

It implies that

$$\mathbf{U}_a(u, \mathcal{S}) \geq \mathbf{U}_a(u, P) - \varepsilon \cdot \mathcal{E}_{a,b}(u, P),$$
$$\mathbf{L}_b(u, \mathcal{S}) \leq \mathbf{L}_b(u, P) + \varepsilon \cdot \mathcal{E}_{a,b}(u, P).$$

Note that $(0, \varepsilon)$-kernel is the same as the notion of *ε-kernel* defined by Agarwal et al. [7].

We are interested in computing a (k, ε)-kernel of small size for any given point set $P \subset \mathbb{R}^d$ and parameters k and ε. Once we can compute small (k, ε)-kernels efficiently, we will immediately be able to compute robust coresets for various extent measures, using the standard linearization and duality transforms; see [7] for details.

Related Results The notion of ε-kernels was introduced by Agarwal et al. [7] and efficient algorithms for computing an ε-kernel of a set of n points in $\mathbb{R}^d$ were given in [7, 14, 24]. Yu et al. [24] also gave a simple and fast incremental algorithm for fitting various shapes through a given set of points. See [8] for a review of known results on coresets.

Although there has been much work on approximating a level in an arrangement of hyperplanes using the random-sampling and ε-approximation techniques [15, 21], this line of work has focused on computing a piecewise-linear surface of small complexity that lies within levels $(\pm\varepsilon)k$ for a given integer $k \geq 0$. These algorithms do not extend to approximating a level in the sense defined in this paper.

Perhaps the simplest case in which one can easily show the existence of a small (k, ε)-kernel is when all points of P are collinear in $\mathbb{R}^d$. One simply returns the first and last $k+1$ points along this line as the desired (k, ε)-kernel. In fact, this kernel has exactly the same k-level directional width as P, for all directions. Note that the size of this kernel is $2k + 2$, which is independent of the input size. Generalizing this simple example, Har-Peled and Wang [19] showed that for any point set $P \subset \mathbb{R}^d$, one can compute a (k, ε)-kernel of size $O(k/\varepsilon^{d-1})$. Their algorithm is based on a recursive construction, and runs in $O(n + k/\varepsilon^{d-1})$ time. Their result led to approximation algorithms for computing various extent measures with k outliers, whose running times are of the form $O(n + (k/\varepsilon)^c)$.

Our Results In Sect. 2 we prove that there exists a (k, ε)-kernel of size $O(k/\varepsilon^{(d-1)/2})$ for any set P of n points in $\mathbb{R}^d$. This result matches the lower bound $\Omega(k/\varepsilon^{(d-1)/2})$. Our construction is relatively simple and intuitive: it works by repeatly *peeling* away $(\varepsilon/4)$-kernels from the input point set P. The running time is bounded by $O(n + k^2/\varepsilon^{d-1})$. The algorithm also leads to a one-pass algorithm for computing (k, ε)-kernels. We tested our algorithm on a variety of inputs for $d \leq 8$; the empirical results show that it works well in low dimensions in terms of both the size of the kernel and the running time.

Our result immediately implies improved approximation algorithms on a wide range of problems discussed in [19]. To name a few, we can compute an ε-approximation of the diameter with k outliers in $O(n + k^2/\varepsilon^{d-1})$ time, an ε-approximation of the minimum-width spherical shell with k outliers in $O(n + k^{2d+1}/\varepsilon^{2d(d+1)})$ time, and a subset of size $O(k/\varepsilon^d)$ for a set of linearly moving points in $\mathbb{R}^d$ so that at any time the diameter (width, smallest-enclosing box, etc.) with k outliers of this subset is an ε-approximation of that of the original moving point set.

In Sect. 3 we present an incremental algorithm for shape fitting with k outliers, which is an extension of the incremental algorithm by Yu et al. [24]. The algorithm works by repeatedly *grating* points from the original point set into a working set; the points that violate the current solution for the working set the most are selected by the algorithm. We prove that the number of iterations of the algorithm is $O((k^2/\varepsilon^{d-1})^{d-1})$, which is independent of n. Our empirical results show that the algorithm converges fairly quickly in practice. Interestingly, while the algorithm itself does not make explicit use of (k, ε)-kernels at all, its analysis crucially relies on the properties of our new algorithm for constructing (k, ε)-kernels.

2 Construction of (k, ε)-Kernel

In this section we describe an iterative algorithm for constructing a (k, ε)-kernel for a set P of n points in $\mathbb{R}^d$. Without loss of generality, we assume that $\varepsilon \leq 1/2$.

2.1 Algorithm

Set $\delta = \varepsilon/4$. Our algorithm consists of $2k + 1$ iterations. At the beginning of the ith iteration, for $0 \leq i \leq 2k$, we have a set $P_i \subseteq P$; initially $P_0 = P$. We compute a δ-kernel $\mathcal{T}_i$ of P_i, using an existing algorithm for computing δ-kernels [7, 14, 24]. We set $P_{i+1} = P_i \setminus \mathcal{T}_i$. After $2k + 1$ iterations, the algorithm returns $\mathcal{S} = \bigcup_{i=0}^{2k} \mathcal{T}_i$ as the desired (k, ε)-kernel.

Intuitively, $\mathcal{T}_i$ approximates the extent measure of P_i. By peeling away $\mathcal{T}_i$ from P_i, important points (in the sense of approximating the extent measures) on the next level of P get "exposed" and can then be subsequently captured in the next iteration of the algorithm. By repeating this peeling process enough times, the union of these point sets approximates the extents of all the first k levels. Similar peeling ideas have been used for halfspace range searching [3, 16, 17] and computing k-hulls [18]. However, unlike our approach, in which we peel away only a small number of points, these algorithms peel away all points of level 0 in each step.

2.2 Proof of Correctness

Let $u \in \mathbb{S}^{d-1}$ be an arbitrary direction. For $0 \leq j < n/2$, let

$$\mathcal{V}_j(u, P) = \langle P^0[u], P^1[u], \ldots, P^j[u], P_j[u], P_{j-1}[u], \ldots, P_0[u] \rangle$$

denote the *ordered* sequence of points realizing the top/bottom j levels of P in direction u. We call the ith iteration of the algorithm *successful* with respect to direction u if $\mathcal{V}_{k-1}(u, P) \cap \mathcal{T}_i \neq \emptyset$ or *unsuccessful* otherwise. Since $|\mathcal{V}_{k-1}(u, P)| = 2k$ and the algorithm consists of $2k + 1$ iterations, at least one of them is unsuccessful with respect to u.

Lemma 2.1 *If the ith iteration is unsuccessful with respect to direction u, then $\mathcal{E}(u, P_i) \leq (1 + \varepsilon/2)\, \mathcal{E}_k(u, P)$. In fact,*

$$\mathbf{U}(u, P_i) \leq \mathbf{U}_k(u, P) + (\varepsilon/2)\mathcal{E}_k(u, P),$$
$$\mathbf{L}(u, P_i) \geq \mathbf{L}_k(u, P) - (\varepsilon/2)\mathcal{E}_k(u, P).$$

Proof Since $\mathcal{T}_i \cap \mathcal{V}_{k-1}(u, P) = \emptyset$, we have $\mathcal{E}(u, \mathcal{T}_i) \leq \mathcal{E}_k(u, P)$; see Fig. 1. By construction, $\mathcal{T}_i$ is a δ-kernel of P_i. Therefore,

$$\mathcal{E}(u, P_i) \leq \mathcal{E}(u, \mathcal{T}_i)/(1 - \delta) \leq (1 + \varepsilon/2)\, \mathcal{E}_k(u, P),$$

proving the first inequality of the lemma. Note that $\mathbf{U}(u, \mathcal{T}_i) \leq \mathbf{U}_k(u, P)$. We have

$$\begin{aligned} \mathbf{U}(u, P_i) &\leq \mathbf{U}(u, \mathcal{T}_i) + (\mathcal{E}(u, P_i) - \mathcal{E}(u, \mathcal{T}_i)) \\ &\leq \mathbf{U}_k(u, P) + \delta \mathcal{E}(u, P_i) \\ &\leq \mathbf{U}_k(u, P) + (\varepsilon/2)\mathcal{E}_k(u, P). \end{aligned}$$

The third claim in this lemma can be proved in a similar manner. □

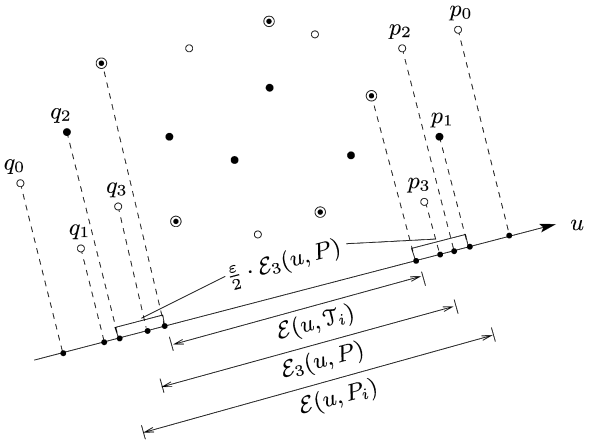


Fig. 1 Illustration of Lemma 2.1 (for $k = 3$). *Double circles* represent points in $\mathcal{T}_i$. The *union of double circles* and *solid circles* represent points in P_i. *Hollow circles* represent points in $P \setminus P_i = \bigcup_{j=0}^{i-1} \mathcal{T}_j$. Here $p_j = P^j[u]$ and $q_j = P_j[u]$

Lemma 2.2 *$\mathcal{S}$ is a (k, ε)-kernel of P.*

Proof To prove the claim, we fix an arbitrary direction $u \in \mathbb{S}^{d-1}$ and argue that

$$\mathcal{E}_{a,b}(u, \mathcal{S}) \geq (1 - \varepsilon)\mathcal{E}_{a,b}(u, P)$$

for all $0 \leq a, b \leq k$. We only discuss the case $a = b = k$; other cases can be handled by slightly modifying the argument given below.

We first show that

$$\mathbf{U}_k(u, \mathcal{S}) \geq \mathbf{U}_k(u, P) - (\varepsilon/2)\,\mathcal{E}_k(u, P). \tag{1}$$

If $P^\ell[u] \in \mathcal{S}$ for all $0 \leq \ell \leq k$, then $\mathbf{U}_k(u, \mathcal{S}) \geq \mathbf{U}_k(u, P)$ and hence (4) is clearly true. So let us assume that there exists $\ell \leq k$ such that $P^\ell[u] \notin \mathcal{S}$. Observe that for any iteration i, we must have $P^\ell[u] \in P_i$.

We define

$$Q = \big\{p \in P \mid \langle u, p\rangle \geq \mathbf{U}_k(u, P) - (\varepsilon/2)\,\mathcal{E}_k(u, P)\big\}.$$

Then (1) is equivalent to $|\mathcal{S} \cap Q| \geq k + 1$.

Consider the ith iteration of the algorithm that is unsuccessful with respect to direction u. Since $P^\ell[u] \in P_i$ and $\mathcal{T}_i$ is a δ-kernel of P_i,

$$\begin{aligned}\mathbf{U}(u, \mathcal{T}_i) &\geq \mathbf{U}(u, P_i) - \delta\,\mathcal{E}(u, P_i)\\ &\geq \mathbf{U}_\ell(u, P) - (\varepsilon/4)(1 + \varepsilon/2)\,\mathcal{E}_k(u, P) \quad \text{(by Lemma 2.1)}\\ &\geq \mathbf{U}_k(u, P) - (\varepsilon/2)\,\mathcal{E}_k(u, P).\end{aligned}$$

Hence ${\mathcal{T}_i}^0[u] \in Q$. Furthermore, since this iteration is unsuccessful, ${\mathcal{T}_i}^0[u] \notin \{P^0[u], \ldots, P^{k-1}[u]\}$, implying that $|\mathcal{T}_i \cap (Q \setminus \{P^0[u], \ldots, P^{k-1}[u]\})| \geq 1$.

Let m be the total number of successful iterations of the algorithm. Then $|\mathcal{S} \cap \mathcal{V}_{k-1}(u, P)| \geq m$ and therefore $|\mathcal{S} \cap \{P^0[u], \ldots, P^{k-1}[u]\}| \geq m - k$. Furthermore, as there are $2k + 1 - m$ unsuccessful iterations, the preceding argument implies that $|\mathcal{S} \cap (Q \setminus \{P^0[u], \ldots, P^{k-1}[u]\})| \geq 2k + 1 - m$. Hence, $|\mathcal{S} \cap Q| \geq (m - k) + (2k + 1 - m) = k + 1$, which in turn implies (1).

Using a similar argument, we can prove that

$$\mathbf{L}_k(u, \mathcal{S}) \leq \mathbf{L}_k(u, P) + (\varepsilon/2)\,\mathcal{E}_k(u, P). \tag{2}$$

Putting (1) and (2) together, we get that

$$\mathcal{E}_k(u, \mathcal{S}) = \mathbf{U}_k(u, \mathcal{S}) - \mathbf{L}_k(u, \mathcal{S}) \geq (1 - \varepsilon)\mathcal{E}_k(u, P).$$

Since u is an arbitrary direction, $\mathcal{S}$ is indeed a (k, ε)-kernel of P. □

2.3 Time Complexity

Chan [14] has shown that a δ-kernel of size $O(1/\delta^{(d-1)/2})$ can be computed in $O(n + 1/\delta^{d-1})$ time. Using this result, we obtain an algorithm for computing (k, ε)-kernels of size $O(k/\delta^{(d-1)/2})$ with running time $O(nk + k/\varepsilon^{d-1})$. We can improve

the running time to $O(n + k^2/\varepsilon^{d-1})$, using the following observation: for any point set $P \subset \mathbb{R}^d$, if $\mathcal{R} \subset P$ is a (k, ε_1)-kernel of P, and $\mathcal{S} \subset \mathcal{R}$ is a (k, ε_2)-kernel of $\mathcal{R}$, then $\mathcal{S}$ is a $(k, \varepsilon_1 + \varepsilon_2)$-kernel of P.

We first invoke the $O(n + k/\varepsilon^{d-1})$-time algorithm of Har-Peled and Wang [19] to compute a $(k, \varepsilon/2)$-kernel $\mathcal{R}$ of P of size $O(k/\varepsilon^{d-1})$, and then apply the above $O(nk + k/\varepsilon^{d-1})$-time algorithm on $\mathcal{R}$ to compute a $(k, \varepsilon/2)$-kernel $\mathcal{S}$ of $\mathcal{R}$ of size $O(k/\varepsilon^{(d-1)/2})$. The resulting set $\mathcal{S}$ is the desired (k, ε)-kernel of P, and the total running time is bounded by $O(n + k^2/\varepsilon^{d-1})$. We conclude with the following theorem.

Theorem 2.3 *Given a set P of n points in $\mathbb{R}^d$ and parameters $k, \varepsilon > 0$, one can compute, in $O(n + k^2/\varepsilon^{d-1})$ time, a (k, ε)-kernel of P of size $O(k/\varepsilon^{(d-1)/2})$.*

It is easy to verify that for a point set P' which is an $\Omega(\sqrt{\varepsilon})$-net of the unit hypersphere (i.e., the minimum distance in P' is $\Omega(\sqrt{\varepsilon})$), all points of P' must be in every $(0, \varepsilon)$-kernel of P'. By replicating $k + 1$ times every point of P' and perturbing slightly, the resulting point set P has the property that any (k, ε)-kernel of P must contain $\Omega(k/\varepsilon^{(d-1)/2})$ points. Thus, in the worst case, a (k, ε)-kernel for P is of size $\Omega(k/\varepsilon^{(d-1)/2})$, matching the upper bound given in Theorem 2.3.

We also note that performing $2k + 1$ iterations in the above algorithm is not only sufficient but also necessary to compute a (k, ε)-kernel. For example, in $\mathbb{R}^2$, consider $\Omega(n)$ (slightly perturbed) copies of the two points $(-1, 0)$ and $(1, 0)$ on the x-axis, together with the following $2k + 2$ points on the y-axis: $(0, 1/\varepsilon^{k-1}), (0, 1/\varepsilon^{k-2}), \ldots, (0, 1)$; $(0, -\varepsilon), (0, -\varepsilon^2), \ldots, (0, -\varepsilon^k)$; $(0, -\varepsilon^{k+1})$, $(0, \varepsilon^{k+1})$. If the number of iterations is $2k$, the algorithm may only output the first $2k$ points listed above along the y-axis together with a set of other points on the x-axis, which is clearly not a (k, ε)-kernel in the y-direction.

2.4 Extensions

One-pass Algorithms In many applications it is desirable to compute a certain function in a single pass over the input data, using a small working memory and processing each point quickly. Agarwal et al. [7], Agarwal and Yu [2], and Chan [14] described such one-pass algorithms for computing ε-kernels. Our (k, ε)-kernel algorithm suggests how to develop a one-pass algorithm for computing a (k, ε)-kernel by using such an algorithm for ε-kernel as a subroutine. Suppose there is a one-pass algorithm $\mathcal{A}$ that computes ε-kernels using $N(\varepsilon)$ space and $T(\varepsilon)$ time per point. To compute a (k, ε)-kernel of a point set P in one pass, we proceed as follows. We simultaneously run $2k + 1$ instances of $\mathcal{A}$, namely $\mathcal{A}_0, \mathcal{A}_1, \ldots, \mathcal{A}_{2k}$, each of which maintains an $(\varepsilon/4)$-kernel $\mathcal{T}_i$ $(0 \le i \le 2k)$ of its own input seen so far. The input of $\mathcal{A}_0$ is $P_0 = P$, and the input P_i of $\mathcal{A}_i$, for $i \ge 1$, is initially empty. The algorithm $\mathcal{A}_0$ processes each point in P_0 in turn. For $i \ge 1$, we insert a point $p \in P_{i-1}$ into P_i whenever any of the following two events happens:

1. p is not added into $\mathcal{T}_{i-1}$ after being processed by $\mathcal{A}_{i-1}$;
2. p is deleted from $\mathcal{T}_{i-1}$ by $\mathcal{A}_{i-1}$.

It is easy to see that in the end $P_i = P_{i-1} \setminus \mathcal{T}_{i-1}$ and $\mathcal{T}_i$ is an $(\varepsilon/4)$-kernel of P_i, for each $0 \le i \le 2k$. Therefore, $\mathcal{S} = \bigcup_{i=0}^{2k} \mathcal{T}_i$ is a (k, ε)-kernel of P as desired. The total

space needed is $O(k \cdot N(\varepsilon/4))$ and the amortized time to process each point in P is $O(k \cdot T(\varepsilon/4))$. Thus we obtain the following result.

Theorem 2.4 *Given a one-pass algorithm for computing ε-kernels in $N(\varepsilon)$ space and $T(\varepsilon)$ time per point, there is a one-pass algorithm for computing (k, ε)-kernels in $O(k \cdot N(\varepsilon/4))$ space and $O(k \cdot T(\varepsilon/4))$ amortized time per point.*

Polynomials Let $\mathcal{F}$ be a family of d-variate polynomials. The (k, ℓ)-*extent* of $\mathcal{F}$ at $x \in \mathbb{R}^d$, denoted by $\mathcal{E}_{k,\ell}(x, \mathcal{F})$, is defined by

$$\mathcal{E}_{k,\ell}(x, \mathcal{F}) = f_i(x) - f_j(x),$$

where f_i (resp. f_j) is the function in $\mathcal{F}$ that has the k-th largest (resp. ℓ-th smallest) value in the set $\mathcal{F}(x) = \{f(x) \mid f \in \mathcal{F}\}$. A subset $\mathcal{G} \subseteq \mathcal{F}$ is a (k, ε)-*kernel* of $\mathcal{F}$ if for any $0 \le a, b \le k$ and any $x \in \mathbb{R}^d$,

$$(1 - \varepsilon)\mathcal{E}_{a,b}(x, \mathcal{F}) \le \mathcal{E}_{a,b}(x, \mathcal{G}) \le \mathcal{E}_{a,b}(x, \mathcal{F}).$$

We say that the *dimension of linearization* of $\mathcal{F}$ is m if there exists a map $\varphi : \mathbb{R}^d \to \mathbb{R}^m$ so that each function $f \in \mathcal{F}$ maps to a linear function $h_f : \mathbb{R}^m \to \mathbb{R}$ in the sense that $f(x) = h_f(\varphi(x))$ for all $x \in \mathbb{R}^d$. Using Theorem 2.3 together with the standard linearization and duality transforms as described in [7], we immediately have the following.

Theorem 2.5 *Let $\mathcal{F}$ be a family of n polynomials, and let m be the dimension of linearization of $\mathcal{F}$. Given parameters $k, \varepsilon > 0$, one can compute a (k, ε)-kernel of $\mathcal{F}$ of size $O(k/\varepsilon^{m/2})$ in $O(n + k^2/\varepsilon^m)$ time.*

Roots of Polynomials To compute (k, ε)-kernels of fractional powers of polynomials, we need the following observation from [24] (see also [7]):

Lemma 2.6 *Let $0 < \varepsilon < 1$, $0 \le a \le A \le B \le b$, and r be a positive integer. If $B^r - A^r \ge (1 - \varepsilon^r)(b^r - a^r)$, then $B - A \ge (1 - \varepsilon)(b - a)$.*

Hence, in order to compute a (k, ε)-kernel of $\{f_1^{1/r}, \ldots, f_n^{1/r}\}$, where each f_i is a polynomial and r is a positive integer, it is sufficient to compute a (k, ε^r)-kernel of $\{f_1, \ldots, f_n\}$. Applying Theorem 2.5, we then have the following.

Theorem 2.7 *Let $\mathcal{F} = \{f_1^{1/r}, \ldots, f_n^{1/r}\}$ be a family of n functions, where each f_i is a polynomial and r is a positive integer. Let m be the dimension of linearization of $\{f_1, \ldots, f_n\}$. Given parameters $k, \varepsilon > 0$, one can compute a (k, ε)-kernel of $\mathcal{F}$ of size $O(k/\varepsilon^{rm/2})$ in $O(n + k^2/\varepsilon^{rm})$ time.*

Theorems 2.5 and 2.7 immediately imply improved results for various shape-fitting problems mentioned in [19], some of which have been listed in Introduction. The details can be found in [19].

3 Incremental Shape-Fitting Algorithm

In this section we present a simple incremental algorithm for shape fitting with k outliers. Compared to the shape-fitting algorithms derived directly from Theorems 2.5 and 2.7, the incremental algorithm does not enjoy a better bound on the running time, but usually performs faster in practice. The algorithm does not make explicit use of (k, ε)-kernels. However, by exploiting the construction of (k, ε)-kernels from the previous section, we show that the number of iterations performed by the algorithm is independent of the input size n. We first describe and analyze the algorithm for the special case in which we wish to find a minimum-width slab that contains all but at most k points of a point set. We then show that the same approach can be extended to a number of other shapes, including cylinders, spherical shells, and cylindrical shells.

3.1 Algorithm

A *slab* $\sigma \subseteq \mathbb{R}^d$ is the region bounded by two parallel hyperplanes. The *width* of σ is the distance between the two hyperplanes. The hyperplane passing through the middle of σ is called the *center hyperplane* of σ. For a given parameter $c > 0$, we will use $c \cdot \sigma$ to denote the slab obtained by scaling σ by the factor of c with respect to its center hyperplane. Let $u_\sigma \in \mathbb{S}^{d-1}$ denote the direction in the upper hemisphere normal to the hyperplanes bounding σ.

Let $A_{\text{opt}}(R, k)$ be an algorithm that returns a slab of the minimum width that contains all but at most k points of R. The incremental algorithm proceeds as follows. We start with an arbitrary subset $R \subseteq P$ of constant size and compute $\sigma = A_{\text{opt}}(R, k)$. If $\frac{1}{1-\varepsilon} \cdot \sigma$ can cover all but at most k points of P, then we stop because we have found an ε-approximation of $A_{\text{opt}}(P, k)$. Otherwise, we add the points of $\mathcal{V}_k(u_\sigma, P)$ (as defined in Sect. 2.2) to R and repeat the above step.

Note that the algorithm always terminates. If the number of iterations of the algorithm is small, its running time would also be small. We next prove a bound on the number of iterations that is independent of n.

3.2 Analysis

We will show that there exists a family $\mathcal{H}$ of $O(k^2/\varepsilon^{d-1})$ great hyperspheres on $\mathbb{S}^{d-1}$ with the following property: the algorithm stops as soon as it computes a slab σ_1 such that, for some slab σ_2 computed in an earlier iteration, u_{σ_1} and u_{σ_2} lie in the same cell of the arrangement $\mathcal{A}(\mathcal{H})$ of $\mathcal{H}$. This would immediately imply an

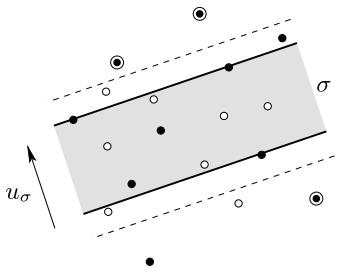


Fig. 2 One iteration of the incremental algorithm (for $k = 1$). *Solid circles* represent points in R, and *double circles* represent points to be added into R in this iteration

$O((k^2/\varepsilon^{d-1})^{d-1})$ bound on the number of iterations. First we prove a useful property of affine transforms. We then describe how to choose the great hyperspheres. Finally we prove the desired property of the chosen hyperspheres and the convergence of the algorithm.

We write an affine transform $\tau : \mathbb{R}^d \to \mathbb{R}^d$ as $\tau(x) = \boldsymbol{A}^T \cdot x + q_0$ for $x \in \mathbb{R}^d$, where $\boldsymbol{A}$ is a $d \times d$ nonsingular matrix and $q_0 \in \mathbb{R}^d$ is a fixed vector. Given τ, let $\widetilde{\tau} : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ be the map defined by $\widetilde{\tau}(u) = \boldsymbol{A}^{-1}u/\|\boldsymbol{A}^{-1}u\|$ for $u \in \mathbb{S}^{d-1}$. If the transform τ is clear from the context, we simply use $\widetilde{u}$ to denote $\widetilde{\tau}(u)$.

Lemma 3.1 *Let $\tau : \mathbb{R}^d \to \mathbb{R}^d$ be an affine transform. For any direction $u \in \mathbb{S}^{d-1}$, any four points $p, q, r, s \in \mathbb{R}^d$, and any parameter $c \in \mathbb{R}$,*

$$\langle u, p - q\rangle \leq c \cdot \langle u, r - s\rangle \quad \Longleftrightarrow \quad \langle \widetilde{u}, \tau(p) - \tau(q)\rangle \leq c \cdot \langle \widetilde{u}, \tau(r) - \tau(s)\rangle.$$

In particular, for any point set P and $\widetilde{P} = \tau(P)$, we have $\widetilde{P}^i[\widetilde{u}] = \tau(P^i[u])$ for $0 \leq i < |P|$.

Proof Suppose $\tau(x) = \boldsymbol{A}^T \cdot x + q_0$ for $x \in \mathbb{R}^d$. Then

$$\begin{aligned}\langle \widetilde{u}, \tau(p) - \tau(q)\rangle &= \langle \boldsymbol{A}^{-1}u/\|\boldsymbol{A}^{-1}u\|, \boldsymbol{A}^T(p - q)\rangle \\ &= u^T(\boldsymbol{A}^T)^{-1}\boldsymbol{A}^T(p - q)/\|\boldsymbol{A}^{-1}u\| \\ &= u^T(p - q)/\|\boldsymbol{A}^{-1}u\| = \langle u, p - q\rangle/\|\boldsymbol{A}^{-1}u\|.\end{aligned}$$

Similarly, we have $\langle \widetilde{u}, \tau(r) - \tau(s)\rangle = \langle u, r - s\rangle/\|\boldsymbol{A}^{-1}u\|$. Hence the first claim of the lemma follows. Setting $c = 0$ in the first claim, we know that $\langle u, p\rangle \leq \langle u, q\rangle$ if and only if $\langle \widetilde{u}, \tau(p)\rangle \leq \langle \widetilde{u}, \tau(q)\rangle$. The second claim then follows. □

We need the following two lemmas to describe how to choose the desired family of great hyperspheres.

Lemma 3.2 *Let S be a set of n points in $\mathbb{R}^d$. There exists a set H of $O(n^2)$ great hyperspheres in $\mathbb{S}^{d-1}$ so that for any $u, v \in \mathbb{S}^{d-1}$ lying in the same cell of $\mathcal{A}(H)$, we have $S^i[u] = S^i[v]$, for $i = 0, \ldots, n - 1$.*

Proof For any pair of points $p, q \in S$, let h_{pq} be the great hypersphere in $\mathbb{S}^{d-1}$, defined by the equation

$$\langle u, p\rangle = \langle u, q\rangle, \quad u \in \mathbb{S}^{d-1}.$$

We let $H = \{h_{pq} \mid p, q \in S\}$. Clearly $|H| = O(n^2)$. Consider any cell $\Delta \in \mathcal{A}(H)$. By construction, it is easy to see that the relative ordering of the elements in $\{\langle u, p\rangle \mid p \in S\}$ is the same for all $u \in \Delta$. Hence, $S^i[u] = S^i[v]$ for any $u, v \in \Delta$, as desired. □

Lemma 3.3 *Let S be a set of n points in $\mathbb{R}^d$ whose affine hull spans $\mathbb{R}^d$, and let $\delta > 0$ be a parameter. There exist an affine transform τ and a set H of $O(1/\delta)$ great*

hyperspheres in $\mathbb{S}^{d-1}$ so that for any $u, v \in \mathbb{S}^{d-1}$ lying in the same cell of $\mathcal{A}(H)$ and for any two points $p, q \in S$,

$$|\langle \widetilde{u} - \widetilde{v}, \tau(p) - \tau(q)\rangle| \leq \delta \cdot \mathcal{E}(\widetilde{v}, \tau(S)).$$

Proof Let $\mathbb{B}^d \subseteq \mathbb{R}^d$ be a unit ball centered at the origin. By John's Ellipsoid Theorem [23], there exists an affine transform τ so that $(1/d) \cdot \mathbb{B}^d \subseteq \text{conv}(\tau(S)) \subseteq \mathbb{B}^d$.

Agarwal et al. [7] proved that there exists a set H of $O(1/\delta)$ great hyperspheres in $\mathbb{S}^{d-1}$, such that for any $u, v \in \mathbb{S}^{d-1}$ lying in the same cell of $\mathcal{A}(H)$, $\|\widetilde{u} - \widetilde{v}\| \leq \delta/d$. Note that $\mathcal{E}(\widetilde{v}, \tau(S)) \geq 2/d$, and for any $p, q \in S$, $\|\tau(p) - \tau(q)\| \leq 2$. Thus,

$$\begin{aligned} |\langle \widetilde{u} - \widetilde{v}, \tau(p) - \tau(q)\rangle| &\leq \|\widetilde{u} - \widetilde{v}\| \cdot \|\tau(p) - \tau(q)\| \\ &\leq 2\delta/d \leq \delta \cdot \mathcal{E}(\widetilde{v}, \tau(S)), \end{aligned}$$

as claimed. □

We assume $\varepsilon \leq 1/2$. Fix $\delta = \varepsilon/6 \leq 1/12$. Let $\mathcal{S}$ be a (k, δ)-kernel of P computed by the algorithm in Sect. 2.1, and let $\mathcal{X} = P \setminus \mathcal{S}$. Using Lemmas 3.2 and 3.3, we construct a decomposition of $\mathbb{S}^{d-1}$ as follows. For each point $p \in \mathcal{S}$, let H_p be a family of $O(1/\delta)$ great hyperspheres that satisfy Lemma 3.3 for $\mathcal{X} \cup \{p\}$, and let τ_p be the corresponding affine transform. Let $\Gamma = \{\tau_p \mid p \in \mathcal{S}\}$. Let G be the set of $O(|\mathcal{S}|^2)$ great hyperspheres that satisfy Lemma 3.2 for the set $\mathcal{S}$. Set

$$\mathcal{H} = G \cup \Big(\bigcup_{p \in \mathcal{S}} H_p \Big).$$

Note that $|\mathcal{H}| = O(|\mathcal{S}|^2 + |\mathcal{S}|/\delta) = O(k^2/\varepsilon^{d-1})$. The number of cells in $\mathcal{A}(\mathcal{H})$ is $O(|\mathcal{H}|^{d-1}) = O((k^2/\varepsilon^{d-1})^{d-1})$.

Next we prove crucial properties of the decomposition $\mathcal{A}(\mathcal{H})$.

Lemma 3.4 *Let $u \in \mathbb{S}^{d-1}$ be a direction. Set $Q = \mathcal{X} \cup \{\mathcal{S}^k[u]\}$. For any affine transform τ, we have*

$$\mathcal{E}(\widetilde{u}, \tau(Q)) \leq (1 + \delta) \cdot \mathcal{E}_k(\widetilde{u}, \tau(P)). \tag{3}$$

Proof Since the algorithm described in Sect. 2.1 performs $2k + 1$ iterations, at least one of them, say the iteration i, was unsuccessful with respect to direction u. By Lemma 2.1 and the fact $\mathcal{X} \subseteq P_i$, we know that

$$\mathbf{U}(u, \mathcal{X}) \leq \mathbf{U}(u, P_i) \leq \mathbf{U}_k(u, P) + (\delta/2) \cdot \mathcal{E}_k(u, P), \tag{4}$$

$$\mathbf{L}(u, \mathcal{X}) \geq \mathbf{L}(u, P_i) \geq \mathbf{L}_k(u, P) - (\delta/2) \cdot \mathcal{E}_k(u, P). \tag{5}$$

Therefore,

$$\begin{aligned} \mathcal{E}\big(u, \mathcal{X} \cup \{\mathcal{S}^k[u]\}\big) &\leq \max\big(\langle u, \mathcal{S}^k[u]\rangle, \mathbf{U}(u, \mathcal{X})\big) - \min\big(\langle u, \mathcal{S}_k[u]\rangle, \mathbf{L}(u, \mathcal{X})\big) \\ &= \max\big(\mathbf{U}_k(u, \mathcal{S}), \mathbf{U}(u, \mathcal{X})\big) - \min\big(\mathbf{L}_k(u, \mathcal{S}), \mathbf{L}(u, \mathcal{X})\big) \end{aligned}$$

$$\begin{aligned}
&\leq \big(\mathbf{U}_k(u,P) + (\delta/2)\cdot\mathcal{E}_k(u,P)\big)\\
&\quad - \big(\mathbf{L}_k(u,P) - (\delta/2)\cdot\mathcal{E}_k(u,P)\big) \quad \text{(by (4) and (5))}\\
&\leq (1+\delta)\mathcal{E}_k(u,P).
\end{aligned}$$

Hence by Lemma 3.1, $\mathcal{E}(\widetilde{u},\tau(Q)) \leq (1+\delta)\cdot\mathcal{E}_k(\widetilde{u},\tau(P))$. □

Lemma 3.5 *Let $u, v \in \mathbb{S}^{d-1}$ be any two directions lying in the same cell of $\mathcal{A}(\mathcal{H})$. Then, for any $0 \leq a, b \leq k$, we have*

$$\mathcal{E}_{a,b}(v, \mathcal{V}_k(u,P)) \geq (1-\varepsilon)\cdot\mathcal{E}_{a,b}(v,P).$$

Proof We prove the claim for the case $a, b = k$; the argument easily adapts to other cases. To this end, we show that for $\ell \leq k$,

$$\langle v, P^\ell[u]\rangle \geq \mathbf{U}_k(v,P) - (\varepsilon/2)\cdot\mathcal{E}_k(v,P).$$

We start by considering the case $P^\ell[u] \in \mathcal{S}$. Observe that $\mathcal{V}_k(u,\mathcal{S}) = \mathcal{V}_k(v,\mathcal{S})$ (we remind the reader that $\mathcal{V}$ is an ordered set, as such equality here means also identical ordering by level). In particular, since $P^\ell[u]$ is clearly at level $\leq \ell$ of $\mathcal{S} \subseteq P$ in direction u, $P^\ell[u]$ is also at level $\leq \ell$ of $\mathcal{S}$ in direction v. Hence,

$$\langle v, P^\ell[u]\rangle \geq \mathbf{U}_\ell(v,\mathcal{S}) \geq \mathbf{U}_k(v,\mathcal{S}) \geq \mathbf{U}_k(v,P) - \delta\cdot\mathcal{E}_k(v,P),$$

where the last inequality follows from the fact that $\mathcal{S}$ is a (k,δ)-kernel of P.

Now consider the case $P^\ell[u] \in \mathcal{X}$. Set $Q = \mathcal{X} \cup \{\mathcal{S}^k[u]\}$, and let $\tau \in \Gamma$ be the affine transform that satisfies Lemma 3.3 for the set Q. Since $\ell \leq k$, we have

$$\langle u, \mathcal{S}^k[u]\rangle \leq \langle u, P^k[u]\rangle \leq \langle u, P^\ell[u]\rangle,$$

implying that $\langle u, \mathcal{S}^k[u] - P^\ell[u]\rangle \leq 0$, or equivalently by applying Lemma 3.1 with $c = 0$,

$$\langle \widetilde{u}, \tau(\mathcal{S}^k[u]) - \tau(P^\ell[u])\rangle \leq 0.$$

Therefore,

$$\begin{aligned}
&\langle \widetilde{v}, \tau(\mathcal{S}^k[u]) - \tau(P^\ell[u])\rangle\\
&\quad = \langle \widetilde{u}, \tau(\mathcal{S}^k[u]) - \tau(P^\ell[u])\rangle + \langle \widetilde{v} - \widetilde{u}, \tau(\mathcal{S}^k[u]) - \tau(P^\ell[u])\rangle\\
&\quad \leq \langle \widetilde{v} - \widetilde{u}, \tau(\mathcal{S}^k[u]) - \tau(P^\ell[u])\rangle. \qquad (6)
\end{aligned}$$

Note that u, v lie in the same cell of $\mathcal{A}(\mathcal{H})$, and $P^\ell[u], \mathcal{S}^k[u] \in Q = \mathcal{X} \cup \{\mathcal{S}^k[u]\}$. By applying Lemma 3.3 to the right-hand side of (6), we obtain

$$\begin{aligned}
&\langle \widetilde{v}, \tau(\mathcal{S}^k[u]) - \tau(P^\ell[u])\rangle\\
&\quad \leq \delta\cdot\mathcal{E}(\widetilde{v}, \tau(Q)) \leq \delta(1+\delta)\cdot\mathcal{E}_k(\widetilde{v}, \tau(P)) \leq 2\delta\cdot\mathcal{E}_k(\widetilde{v}, \tau(P)), \qquad (7)
\end{aligned}$$

where the second inequality follows from Lemma 3.4. By Lemma 3.1, (7) implies

$$\langle v, \mathcal{S}^k[u] - P^\ell[u]\rangle \leq 2\delta\cdot\mathcal{E}_k(v,P). \qquad (8)$$

Observing that $\mathcal{S}^k[u] = \mathcal{S}^k[v]$ and using (8), we obtain

$$\begin{aligned}\langle v, P^\ell[u]\rangle &\geq \langle v, \mathcal{S}^k[u]\rangle - 2\delta \cdot \mathcal{E}_k(v, P) = \langle v, \mathcal{S}^k[v]\rangle - 2\delta \cdot \mathcal{E}_k(v, P)\\ &\geq \mathbf{U}_k(v, P) - 3\delta \cdot \mathcal{E}_k(v, P) \geq \mathbf{U}_k(v, P) - (\varepsilon/2) \cdot \mathcal{E}_k(v, P).\end{aligned}$$

Similarly, we can prove that for any $0 \leq \ell \leq k$, $\langle v, P_\ell[u]\rangle \leq \mathbf{L}_k(v, P) + (\varepsilon/2) \cdot \mathcal{E}_k(v, P)$.

Hence, $\mathcal{E}_k(v, \mathcal{V}_k(u, P)) \geq (1 - \varepsilon) \cdot \mathcal{E}_k(v, P)$, as claimed. □

We are now ready to bound the number of iterations of the incremental algorithm.

Theorem 3.6 *The number of iterations of the incremental algorithm for fitting the minimum-width slab with k outliers is bounded by* $O((k^2/\varepsilon^{d-1})^{d-1})$, *which is independent of n.*

Proof Let $u_i \in \mathbb{S}^{d-1}$ be the direction orthogonal to the slab computed in the ith iteration. We say that a cell $\Delta \in \mathcal{A}(\mathcal{H})$ is *visited* if $u_i \in \Delta$. Suppose a cell is visited by two iterations i and j during the execution of the algorithm. Assume $i < j$. Then in iteration j, we have $\mathcal{V}_k(u_i, P) \subseteq R$. Let σ be the slab returned by $A_{\text{opt}}(R, k)$ in iteration j. Then $|\sigma|$—the width of σ—is equal to $\mathcal{E}_{a,b}(u_j, R)$ for some appropriate $a, b \leq k$ with $a + b = k$. By Lemma 3.5, we have

$$|\sigma| = \mathcal{E}_{a,b}(u_j, R) \geq \mathcal{E}_{a,b}(u_j, \mathcal{V}_k(u_i, P)) \geq (1 - \varepsilon)\mathcal{E}_{a,b}(u_j, P),$$

or equivalently $\frac{1}{1-\varepsilon}|\sigma| \geq \mathcal{E}_{a,b}(u_j, P)$. This implies that the algorithm would satisfy the stopping criterion in iteration j. Thus the number of iterations is bounded by $|\mathcal{A}(\mathcal{H})| + 1 = O((k^2/\varepsilon^{d-1})^{d-1})$. □

3.3 Other Shapes

The incremental algorithm of Sect. 3.1 for computing an ε-approximation of the minimum-width slab with k outliers can be extended to fitting other shapes as well, such as minimum-width spherical shells or cylindrical shells, minimum-radius cylinders, etc. In this section we describe these extensions.

Spherical Shells and Cylindrical Shells A *spherical shell* is a closed region lying between two concentric spheres in $\mathbb{R}^d$. A *cylinderical shell* is a closed region lying between two co-axial cylinders in $\mathbb{R}^d$. Because fitting spherical shells or cylindrical shells can be formulated as computing the minimum extent of a family $\mathcal{F}$ of m-variate functions for some parameter m [7], we describe a general incremental algorithm for the latter problem. For $x \in \mathbb{R}^m$ and $0 \leq k < n$, we denote

$$\widehat{\mathcal{E}}_k(x, \mathcal{F}) = \min_{a+b=k} \mathcal{E}_{a,b}(x, \mathcal{F}),$$

where $\mathcal{E}_{a,b}(x, \mathcal{F})$ is as defined in Sect. 2.4. Let $A_{\text{opt}}(\mathcal{F}, k)$ be an algorithm that returns

$$x^* = \arg\min_{x \in \mathbb{R}^m} \widehat{\mathcal{E}}_k(x, \mathcal{F}).$$

The incremental algorithm starts by picking an arbitrary subset $\mathcal{R} \subseteq \mathcal{F}$ of constant size and compute $x^* = A_{\text{opt}}(\mathcal{R}, k)$. If $\widehat{\mathcal{E}}_k(x^*, \mathcal{R}) \geq (1-\varepsilon)\widehat{\mathcal{E}}_k(x^*, \mathcal{F})$, then $\widehat{\mathcal{E}}_k(x^*, \mathcal{R})$ is an ε-approximation of $\min_{x\in\mathbb{R}^m} \widehat{\mathcal{E}}_k(x, \mathcal{F})$ and we can stop. Otherwise, we add $\mathcal{V}_k(x^*, \mathcal{F})$—union of the $2(k+1)$ functions of $\mathcal{F}$ that attain the $k+1$ largest values and the $k+1$ smallest values in $\mathcal{F}(x^*) = \{f(x^*) \mid f \in \mathcal{F}\}$—to $\mathcal{R}$, and repeat the above step.

To analyze the above algorithm, we need the following lemma which is the dual version of Lemma 3.5.

Lemma 3.7 *Let $\mathcal{F}$ be a finite family of m-variate linear functions, and $0 < \delta \leq 1/2$ be a parameter. Then there exists a set $\mathcal{H}$ of $O(k^2/\delta^m)$ hyperplanes in $\mathbb{R}^m$ such that for any $u, v \in \mathbb{R}^m$ lying in the same cell of $\mathcal{A}(\mathcal{H})$, and any $0 \leq a, b \leq k$, we have*

$$\mathcal{E}_{a,b}(v, \mathcal{V}_k(u, \mathcal{F})) \geq (1-\delta) \cdot \mathcal{E}_{a,b}(v, \mathcal{F}).$$

Lemma 3.8 *Let $\mathcal{F}$ be a finite family of m-variate polynomials that admits a linearization of dimension ℓ, and $0 < \delta \leq 1/2$ be a parameter. Then there exists a decomposition of $\mathbb{R}^m$ into $O(k^{2m}/\delta^{m\ell})$ cells such that for any $u, v \in \mathbb{R}^m$ lying in the same cell of the decomposition, and any $0 \leq a, b \leq k$, we have*

$$\mathcal{E}_{a,b}(v, \mathcal{V}_k(u, \mathcal{F})) \geq (1-\delta) \cdot \mathcal{E}_{a,b}(v, \mathcal{F}).$$

Proof Let $\varphi : \mathbb{R}^m \to \mathbb{R}^\ell$ be the map so that each function $f \in \mathcal{F}$ maps to a linear function h_f in the sense that $f(x) = h_f(\varphi(x))$ for all $x \in \mathbb{R}^m$. Note that $\Gamma = \{\varphi(x) \mid x \in \mathbb{R}^m\}$, the image of φ, is an m-dimensional surface in $\mathbb{R}^\ell$. Let $\mathcal{F}' = \{h_f \mid f \in \mathcal{F}\}$. Applying Lemma 3.7 to $\mathcal{F}'$, we obtain a set $\mathcal{H}$ of $O(k^2/\delta^\ell)$ hyperplanes in $\mathbb{R}^\ell$. Set $\mathcal{H}^{-1} = \{h^{-1} = \varphi^{-1}(h \cap \Gamma) \mid h \in \mathcal{H}\}$, where each $h^{-1} \in \mathcal{H}^{-1}$ is an $(m-1)$-dimensional algebraic surface in $\mathbb{R}^m$. If $u, v \in \mathbb{R}^m$ lie in the same cell of $\mathcal{A}(\mathcal{H}^{-1})$, then $\varphi(u), \varphi(v)$ lie in the same cell of $\mathcal{A}(\mathcal{H})$. Since $f(x) = h_f(\varphi(x))$ for all $f \in \mathcal{F}$ and $x \in \mathbb{R}^m$, Lemma 3.7 implies that $\mathcal{E}_{a,b}(v, \mathcal{V}_k(u, \mathcal{F})) \geq (1-\delta) \cdot \mathcal{E}_{a,b}(v, \mathcal{F})$. The lemma now follows because $\mathcal{A}(\mathcal{H}^{-1})$ induces a decomposition of $\mathbb{R}^m$ into $O((k^2/\delta^\ell)^m)$ cells [1]. □

Theorem 3.9 *Let $\mathcal{F} = \{f_1, \ldots, f_n\}$ be a family of m-variate nonnegative functions, and $0 < \varepsilon \leq 1/2$ be a parameter. Suppose there exists an m-variate positive function $\psi(x)$ and an integer $r \geq 1$, so that each $g_i(x) = \psi(x) f_i^r(x)$ is a polynomial. Furthermore, suppose $\mathcal{G} = \{g_1, \ldots, g_n\}$ admits a linearization of dimension ℓ. Then there exists a decomposition $\mathcal{D}$ of $\mathbb{R}^m$ into $O(k^{2m}/\varepsilon^{rm\ell})$ cells such that for any $u, v \in \mathbb{R}^m$ lying in the same cell of $\mathcal{D}$, and any $0 \leq a, b \leq k$, we have*

$$\mathcal{E}_{a,b}(v, \mathcal{V}_k(u, \mathcal{F})) \geq (1-\varepsilon) \cdot \mathcal{E}_{a,b}(v, \mathcal{F}).$$

In addition, the incremental algorithm computes an ε-approximation of $\min_{x\in\mathbb{R}^m} \widehat{\mathcal{E}}_k(x, \mathcal{F})$ in $O(k^{2m}/\varepsilon^{rm\ell})$ iterations.

Proof We first make the following observation: for any $\delta \leq 1$, $1 \leq i, j, h, \ell \leq n$, and $x \in \mathbb{R}^m$,

$$g_i(x) - g_j(x) \geq (1-\delta)\big(g_h(x) - g_\ell(x)\big)$$
$$\Leftrightarrow \quad f_i^r(x) - f_j^r(x) \geq (1-\delta)\big(f_h^r(x) - f_\ell^r(x)\big). \tag{9}$$

An immediate consequence of (9) is that $g_i(x) \geq g_j(x)$ if and only if $f_i(x) \geq f_j(x)$.

Consider the decomposition $\mathcal{D}$ of $\mathbb{R}^m$ obtained by applying Lemma 3.8 to the family $\mathcal{G}$ with parameter $\delta = \varepsilon^r$. Note that $|\mathcal{D}| = O(k^{2m}/\varepsilon^{rm\ell})$. For any $u, v \in \mathbb{R}^m$ lying in the same cell of $\mathcal{D}$, by Lemma 3.8 we have

$$\mathcal{E}_{a,b}(v, \mathcal{V}_k(u, \mathcal{G})) \geq (1-\varepsilon^r) \cdot \mathcal{E}_{a,b}(v, \mathcal{G}).$$

Using (9) and Lemma 2.6, we obtain that $\mathcal{E}_{a,b}(v, \mathcal{V}_k(u, \mathcal{F})) \geq (1-\varepsilon) \cdot \mathcal{E}_{a,b}(v, \mathcal{F})$, as desired.

Using the proved result and the same argument as in Theorem 3.6, we immediately obtain the second half of the theorem. □

The problem of computing the minimum-width spherical shell containing all but at most k points of a point set P in $\mathbb{R}^d$ satisfies Theorem 3.9 with $m = d$, $r = 2$, and $\ell = d + 1$ [7]. Hence the incremental algorithm for this problem terminates in $k^{O(d)}/\varepsilon^{O(d^2)}$ iterations. Similarly, the incremental algorithm for computing the minimum-width cylindrical shell terminates in $k^{O(d)}/\varepsilon^{O(d^3)}$ iterations, as it satisfies Theorem 3.9 with $m = 2d - 2$, $r = 2$, and $\ell = O(d^2)$ [7]. We thus obtain the following.

Corollary 3.10 *Let P be a set of n points in $\mathbb{R}^d$, and let $0 < \varepsilon \leq 1/2$ be a parameter. The incremental algorithm computes an ε-approximation of the smallest spherical shell containing all but k points of P in $k^{O(d)}/\varepsilon^{O(d^2)}$ iterations, and the smallest cylindrical shell containing all but k points of P in $k^{O(d)}/\varepsilon^{O(d^3)}$ iterations.*

Cylinders Unlike cylindrical shells and spherical shells, the problem of fitting cylinders cannot be directly formulated as computing the minimum extent of a family of functions. Instead, it can be reduced to computing $\min_{x\in\mathbb{R}^m} \mathbf{U}_k(x, \mathcal{F})$ for a family $\mathcal{F}$ of nonnegative functions, where $\mathbf{U}_k(x, \mathcal{F})$ is defined as the $(k+1)$-th largest value in the set $\mathcal{F}(x)$. The incremental algorithm for such type of problems is as follows. Let $A_{\text{opt}}(\mathcal{F}, k)$ be an algorithm that returns

$$x^* = \arg\min_{x\in\mathbb{R}^m} \mathbf{U}_k(x, \mathcal{F}).$$

The algorithm starts by picking an arbitrary subset $\mathcal{R} \subseteq \mathcal{F}$ of constant size and compute $x^* = A_{\text{opt}}(\mathcal{R}, k)$. If $\mathbf{U}_k(x^*, \mathcal{R}) \geq (1-\varepsilon) \cdot \mathbf{U}_k(x^*, \mathcal{F})$, then we can stop because $\mathbf{U}_k(x^*, \mathcal{R})$ is an ε-approximation of $\min_{x\in\mathbb{R}^m} \mathbf{U}_k(x, \mathcal{F})$. Otherwise, we add $\mathcal{U}_k(x^*, \mathcal{F})$—the $k+1$ functions of $\mathcal{F}$ that attain the $k+1$ largest values in $\mathcal{F}(x^*) = \{f(x^*) \mid f \in \mathcal{F}\}$—to $\mathcal{R}$, and repeat the above step.

Theorem 3.11 *Let $\mathcal{F} = \{f_1, \ldots, f_n\}$ be a family of m-variate nonnegative functions, and $0 < \varepsilon \leq 1/2$ be a parameter. Suppose there exists an m-variate positive function $\psi(x)$ and an integer $r \geq 1$, so that each $g_i(x) = \psi(x) f_i^r(x)$ is a polynomial. Furthermore, suppose $\mathcal{G} = \{g_1, \ldots, g_n\}$ admits a linearization of dimension ℓ. Then there*

exists a decomposition $\mathcal{D}$ of $\mathbb{R}^m$ into $O(k^{2m}/\varepsilon^{m\ell})$ cells such that for any $u, v \in \mathbb{R}^m$ lying in the same cell of $\mathcal{D}$, and any $0 \le a \le k$, we have

$$\mathbf{U}_a(v, \mathcal{U}_k(u, \mathcal{F})) \ge (1 - \varepsilon) \cdot \mathbf{U}_a(v, \mathcal{F}).$$

In addition, the incremental algorithm computes an ε-approximation of $\min_{x \in \mathbb{R}^m} \mathbf{U}_k(x, \mathcal{F})$ in $O(k^{2m}/\varepsilon^{m\ell})$ iterations.

Proof Let $\mathcal{G}' = \{g_1, \ldots, g_n\} \cup \{-g_1, \ldots, -g_n\}$. Since $\mathcal{G}$ admits a linearization of dimension ℓ, $\mathcal{G}'$ also admits a linearization of dimension ℓ. Let $\mathcal{D}$ be the decomposition of $\mathbb{R}^m$ obtained by applying Lemma 3.8 to $\mathcal{G}'$ with parameter $\delta = \varepsilon$. Then $|\mathcal{D}| = O(k^{2m}/\varepsilon^{m\ell})$. For any $u, v \in \mathbb{R}^m$ lying in the same cell of $\mathcal{D}$, we have

$$\mathcal{E}_{a,a}(v, \mathcal{V}_k(u, \mathcal{G}')) \ge (1 - \varepsilon)\mathcal{E}_{a,a}(v, \mathcal{G}').$$

By the symmetry of $\mathcal{G}'$, we have $\mathcal{E}_{a,a}(v, \mathcal{V}_k(u, \mathcal{G}')) = 2\psi(v)(\mathbf{U}_a(v, \mathcal{U}_k(u, \mathcal{F})))^r$, and $\mathcal{E}_{a,a}(v, \mathcal{G}') = 2\psi(v)(\mathbf{U}_a(v, \mathcal{F}))^r$. Hence the above inequality implies

$$\big(\mathbf{U}_a(v, \mathcal{U}_k(u, \mathcal{F}))\big)^r \ge (1 - \varepsilon)\big(\mathbf{U}_a(v, \mathcal{F})\big)^r.$$

It follows that $\mathbf{U}_a(v, \mathcal{U}_k(u, \mathcal{F})) \ge (1 - \varepsilon) \cdot \mathbf{U}_a(v, \mathcal{F})$, as desired.

As a direct consequence, the second half of the theorem follows. □

Note that the cylinder problem has the same linearization as the cylindrical shell problem mentioned above. We then obtain the following.

Corollary 3.12 *Let P be a set of n points in $\mathbb{R}^d$, and let $0 < \varepsilon \le 1/2$ be a parameter. The incremental algorithm computes an ε-approximation of the smallest cylinder containing all but k points of P in $k^{O(d)}/\varepsilon^{O(d^3)}$ iterations.*

We have not tried to optimize our bounds on the number of iterations, but we believe that they can be improved. As shown in Sect. 4, the number of iterations in practice is usually much smaller than the proved bounds here.

4 Experiments

In this section, we demonstrate the effectiveness of our algorithms by evaluating their performances on various synthetic and real data. All our experiments were conducted on a Dell PowerEdge 650 server equipped with 3 GHz Pentium IV processor and 3 GB memory, running Linux 2.4.20.

Computing (k, ε)-Kernels We implemented a simpler version of our (k, ε)-kernel algorithm, which does not invoke Har-Peled and Wang's algorithm [19] first. We used an implementation of Yu et al. [24] for computing δ-kernels in each iteration. Although the worst-case running time of the algorithm is larger than that mentioned in Theorem 2.3, it is simple and works well in practice.

We used three types of synthetic inputs as well as a few large 3D geometric models [20]:

Table 1 Performance of the (k, ε)-kernel algorithm on various synthetic data with $k = 5$. Running time is measured in seconds

Input type	Input size	Approximation error				Running time			
		$d = 3$	$d = 4$	$d = 5$	$d = 8$	$d = 3$	$d = 4$	$d = 5$	$d = 8$
Sphere	10^4	0.022	0.052	0.091	0.165	2.7	4.8	7.7	20.6
	10^5	0.022	0.054	0.103	0.192	9.2	14.5	19.0	42.3
	10^6	0.024	0.055	0.100	0.224	101.4	155.6	194.7	337.3
Cylinder	10^4	0.005	0.027	0.086	0.179	2.7	4.5	7.2	20.6
	10^5	0.015	0.059	0.117	0.243	10.3	13.7	19.7	42.5
	10^6	0.019	0.058	0.125	0.283	129.9	157.9	192.6	335.4
Clustered	10^4	0.001	0.010	0.026	0.061	2.5	4.4	7.7	19.0
	10^5	0.012	0.020	0.031	0.078	7.7	11.5	16.8	36.4
	10^6	0.016	0.021	0.045	0.087	80.1	104.8	138.6	266.5

Table 2 Performance of the (k, ε)-kernel algorithm on various 3D geometric models with $k = 5$. Running time is measured in seconds

Input type	Input size	Kernel size	Approx error	Running time
Bunny	35,947	1001	0.012	3.9
Dragon	437,645	888	0.016	32.8
Buddha	543,652	1064	0.014	35.6

1. Points uniformly distributed on a sphere (*sphere*);
2. Points uniformly distributed on a cylindrical surface (*cylinder*);
3. Clustered point sets (*clustered*), consisting of 20 equal-sized clusters whose centers are uniformly distributed in the unit square and radii uniformly distributed between $[0, 0.2]$;
4. 3D geometric models: *bunny* (∼36 K points), *dragon* (∼438 K points), *buddha* (∼544 K points).

For each input data, we ran our (k, ε)-kernel algorithm with $k = 5$. The algorithm performs 11 iterations and chooses roughly 100 points for the kernel in each iteration. The output size of the algorithm varies between 800 and 1100. To compute the approximation error between the k-level extents of the kernel $\mathcal{S}$ and of the input P, we choose a set $\triangle$ of 1000 random directions from $\mathbb{S}^{d-1}$ and compute

$$\mathrm{err}_{\triangle}(P, \mathcal{S}) = \max_{u \in \triangle} \frac{\mathcal{E}_k(u, P) - \mathcal{E}_k(u, \mathcal{S})}{\mathcal{E}_k(u, P)}.$$

Tables 1 and 2 summarize the approximation error and the running time of the algorithm, for each input data. As can be seen, our algorithm works well in low dimensions both in terms of the approximation error and the running time. Our algorithm also performed quite well on several 3D geometric models. In high dimensions, the performance of our algorithm deteriorates because of the curse of dimensionality.

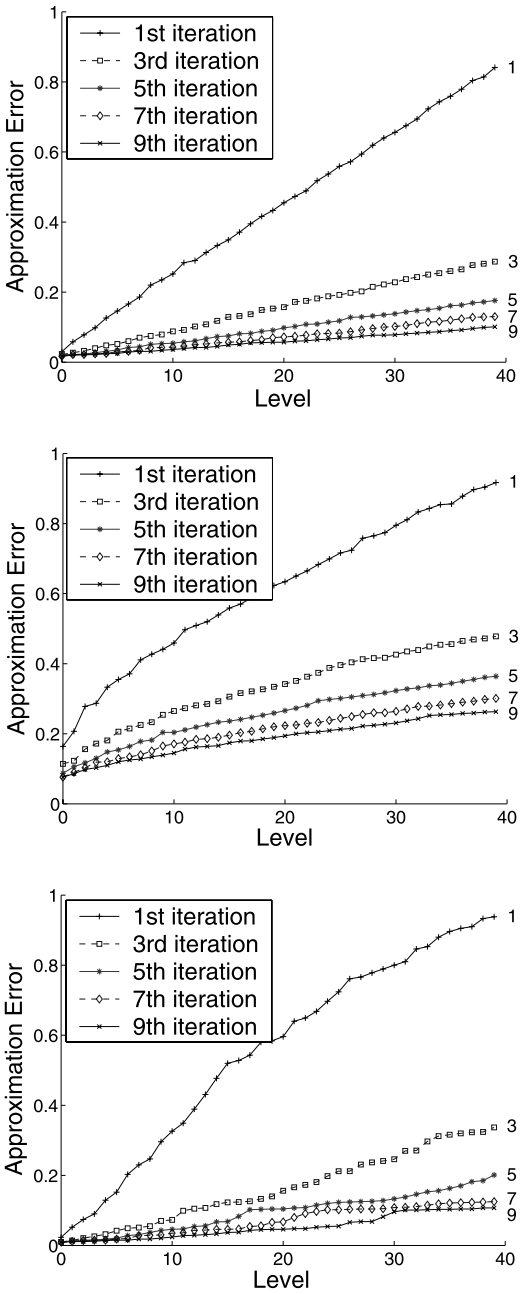


Fig. 3 Approximation errors for different levels in each iteration of the (k, ε)-kernel algorithm. **a** *Sphere* with 10^5 points in $\mathbb{R}^3$; **b** *sphere* with 10^5 points in $\mathbb{R}^5$; **c** the *budda* model. Similar results were observed for other types of inputs as well

We also recorded how the approximation error decreases for each of the first 40 levels, after each iteration of the algorithm. The results are shown in Fig. 3. Observe that the approximation error for every level monotonically decreases during the execution of the algorithm. Moreover, the error decreases rapidly in the first few it-

Table 3 Performance of the incremental algorithm for computing (a) the minimum-width annulus with $k = 10$ outliers, (b) the smallest enclosing circle with $k = 10$ outliers. The numbers of iterations performed by the algorithm are in parentheses. Running time is measured in seconds

Input width	Input size	Running time	Output width
$w = 0.05$	10^4	6.15(4)	0.0503
	10^5	14.51(5)	0.0498
	10^6	18.26(5)	0.0497
$w = 0.50$	10^4	5.74(4)	0.4987
	10^5	6.45(4)	0.4999
	10^6	22.18(5)	0.4975
$w = 5.00$	10^4	53.46(5)	4.9443
	10^5	67.26(5)	4.9996
	10^6	75.42(5)	4.9951

(a)

Input radius	Input size	Running time	Output radius
$r = 1.000$	10^4	0.05(3)	0.993
	10^5	0.14(4)	0.999
	10^6	0.41(4)	0.999

(b)

erations and then it stabilizes. For example, in our experiments for $d = 3$, the error reduces to less than 0.1 within 7 iterations even for the level $k = 40$ and then it decreases very slowly with each iteration. This phenomenon suggests that in practice it is unnecessary to run the algorithm for full $2k + 1$ iterations in order to compute (k, ε)-kernels. The larger the number of iterations is, the larger the kernel size becomes, but the approximation error does not decrease much further.

Incremental Algorithm We applied the incremental shape-fitting algorithm for computing an ε-approximate minimum-width annulus of a point set with k outliers in $\mathbb{R}^2$. We first implemented a brute-force $O(n^5)$ exact algorithm for this problem. Clearly, this algorithm is slow even on medium-sized input. Here our focus is to study the number of iterations of the incremental algorithm; a faster implementation of the exact algorithm would naturally result in a faster implementation of the incremental algorithm. We used the slow exact algorithm as a subroutine to solve the small subproblems in each iteration of the incremental algorithm. We tested this algorithm on a set of synthetic data, generated by uniformly sampling from annuli with fixed inner radius $r = 1.00$ and widths w varying from 0.05 to 5.00, and then artificially introducing $k = 10$ extra outlier points. The experimental results are summarized in Table 3a; see also Fig. 4 for a few snapshots of the running algorithm. As can be seen, the number of iterations of the incremental algorithm is never more than 5. In other words, the algorithm is able to converge to an approximate solution very quickly.

We also applied the incremental algorithm for computing an ε-approximate smallest enclosing circle of a point set with k outliers in $\mathbb{R}^2$. Again, we implemented a brute-force $O(n^4)$ exact algorithm for this problem to solve the small subproblems in each iteration; implementing a faster algorithm (such as an algorithm by Matoušek [22] or by Chan [13]) would result in a faster incremental algorithm. We tested our algorithm on a set of synthetic data, generated by uniformly sampling from a circle of radius $r = 1.00$, and then artificially introducing $k = 10$ extra outlier points. The

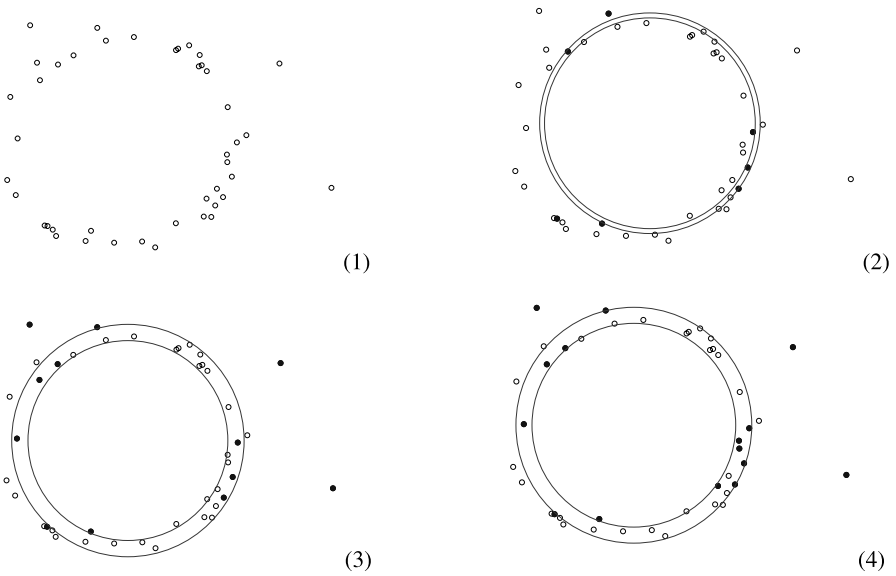


Fig. 4 Snapshots of the incremental algorithm for computing minimum-width annulus with $k = 3$ outliers on an input of size 40. *Black points* represent points in R at the beginning of each iteration

experimental results are shown in Table 3b. Similar to the annulus case, the number of iterations of the incremental algorithm is also small.

5 Conclusions

We have presented an iterative algorithm, with $O(n + k^2/\varepsilon^{d-1})$ running time, for computing a (k, ε)-kernel of size $O(k/\varepsilon^{(d-1)/2})$ for a set P of n points in $\mathbb{R}^d$. We also presented an incremental algorithm for fitting various shapes through a set of points with outliers, and exploited the (k, ε)-kernel algorithm to prove that the number of iterations of the incremental algorithm is independent of n. Both our algorithms are simple and work well in practice.

We conclude by mentioning two open problems: Can a (k, ε)-kernel of size $O(k/\varepsilon^{(d-1)/2})$ be computed in time $O(n + k/\varepsilon^{d-1})$? Can the number of iterations in the incremental algorithm for computing the minimum-width slab be improved to $O(1/\varepsilon^{(d-1)/2})$? For the first question, an anonymous referee pointed out that one can use the dynamic algorithm for ε-kernels [7] to obtain an algorithm with running time $O\big(n + k/\varepsilon^{3(d-1)/2} \cdot \text{polylog}(k, 1/\varepsilon)\big)$. This bound provides an improvement over the current running time for a sufficiently large k.

Acknowledgements The authors thank Yusu Wang for helpful discussions and two anonymous referees for constructive comments that greatly improved the presentation of the paper.

References

1. Agarwal, P.K., Sharir, M.: Arrangements and their applications. In: Sack, J.-R., Urrutia, J. (eds.) Handbook of Computational Geometry, pp. 49–119. Elsevier, Amsterdam (2000)
2. Agarwal, P.K., Yu, H.: A space-optimal data-stream algorithm for coresets in the plane. In: Proc. 23rd Annu. Sympos. Comput. Geom., pp. 1–10, 2007
3. Agarwal, P.K., Arge, L., Erickson, J., Franciosa, P., Vitter, J.S.: Efficient searching with linear constraints. J. Comput. Sys. Sci. **61**, 192–216 (2000)
4. Agarwal, P.K., Aronov, B., Har-Peled, S., Sharir, M.: Approximation and exact algorithms for minimum-width annuli and shells. Discrete Comput. Geom. **24**, 687–705 (2000)
5. Agarwal, P.K., Aronov, B., Sharir, M.: Exact and approximation algorithms for minimum-width cylindrical shells. Discrete Comput. Geom. **26**, 307–320 (2001)
6. Agarwal, P.K., Guibas, L.J., Hershberger, J., Veach, E.: Maintaining the extent of a moving point set. Discrete Comput. Geom. **26**, 353–374 (2001)
7. Agarwal, P.K., Har-Peled, S., Varadarajan, K.R.: Approximating extent measures of points. J. Assoc. Comput. Mach. **51**, 606–635 (2004)
8. Agarwal, P.K., Har-Peled, S., Varadarajan, K.: Geometric approximation via coresets. In: Goodman, J.E., Pach, J., Welzl, E. (eds.) Combinatorial and Computational Geometry. Math. Sci. Research Inst. Pub., Cambridge (2005)
9. Aronov, B., Har-Peled, S.: On approximating the depth and related problems. In: Proc. 16th ACM-SIAM Sympos. Discrete Algorithms, pp. 886–894, 2005
10. Barequet, G., Har-Peled, S.: Efficiently approximating the minimum-volume bounding box of a point set in three dimensions. J. Algorithms **38**, 91–109 (2001)
11. Basch, J., Guibas, L.J., Hershberger, J.: Data structures for mobile data. J. Algorithms **31**, 1–28 (1999)
12. Chan, T.M.: Approximating the diameter, width, smallest enclosing cylinder and minimum-width annulus. Int. J. Comput. Geom. Appl. **12**, 67–85 (2002)
13. Chan, T.M.: Low-dimensional linear programming with violations. SIAM J. Comput. 879–893 (2005)
14. Chan, T.M.: Faster core-set constructions and data-stream algorithms in fixed dimensions. Comput. Geom. Theory Appl. **35**, 20–35 (2006)
15. Chazelle, B.: Cutting hyperplanes for divide-and-conquer. Discrete Comput. Geom. **9**, 145–158 (1993)
16. Chazelle, B., Preparata, F.P.: Halfspace range search: an algorithmic application of k-sets. Discrete Comput. Geom. **1**, 83–93 (1986)
17. Chazelle, B., Guibas, L.J., Lee, D.T.: The power of geometric duality. BIT **25**, 76–90 (1985)
18. Cole, R., Sharir, M., Yap, C.K.: On k-hulls and related problems. SIAM J. Comput. **16**, 61–77 (1987)
19. Har-Peled, S., Wang, Y.: Shape fitting with outliers. SIAM J. Comput. **33**, 269–285 (2004)
20. http://www.cc.gatech.edu/projects/large_models/. Large geometric models archive
21. Matoušek, J.: Approximate levels in line arrangements. SIAM J. Comput. **20**, 222–227 (1991)
22. Matoušek, J.: On geometric optimization with few violated constraints. Discrete Comput. Geom. **14**, 365–384 (1995)
23. Matoušek, J.: Lectures on Discrete Geometry. Springer, Heidelberg (2002)
24. Yu, H., Agarwal, P.K., Poreddy, R., Varadarajan, K.R.: Practical methods for shape fitting and kinetic data structures using coresets. Algorithmica (to appear)

Siegel's Lemma and Sum-Distinct Sets

Iskander Aliev

Abstract Let $L(\mathbf{x}) = a_1x_1 + a_2x_2 + \cdots + a_nx_n$, $n \geq 2$, be a linear form with integer coefficients $a_1, a_2, \ldots, a_n$ which are not all zero. A basic problem is to determine nonzero integer vectors $\mathbf{x}$ such that $L(\mathbf{x}) = 0$, and the maximum norm $\|\mathbf{x}\|$ is relatively small compared with the size of the coefficients $a_1, a_2, \ldots, a_n$. The main result of this paper asserts that there exist linearly independent vectors $\mathbf{x}_1, \ldots, \mathbf{x}_{n-1} \in \mathbb{Z}^n$ such that $L(\mathbf{x}_i) = 0$, $i = 1, \ldots, n-1$, and

$$\|\mathbf{x}_1\| \cdots \|\mathbf{x}_{n-1}\| < \frac{\|\mathbf{a}\|}{\sigma_n},$$

where $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ and

$$\sigma_n = \frac{2}{\pi} \int_0^\infty \left(\frac{\sin t}{t} \right)^n dt.$$

This result also implies a new lower bound on the greatest element of a sum-distinct set of positive integers (Erdös–Moser problem). The main tools are the Minkowski theorem on successive minima and the Busemann theorem from convex geometry.

1 Introduction

Let $\mathbf{a} = (a_1, \ldots, a_n)$, $n \geq 2$, be a nonzero integral vector. Consider the linear form $L(\mathbf{x}) = a_1x_1 + a_2x_2 + \cdots + a_nx_n$. Siegel's lemma with respect to the maximum norm

The work was partially supported by FWF Austrian Science Fund, Project M821-N12.

I. Aliev (✉)
School of Mathematics, University of Edinburgh, James Clerk Maxwell Building, King's Buildings, Mayfield Road, Edinburgh EH9 3JZ, Scotland
e-mail: I.Aliev@ed.ac.uk

$\|\cdot\|$ asks for an optimal constant $c_n > 0$ such that the equation

$$L(\mathbf{x}) = 0$$

has an integral solution $\mathbf{x} = (x_1, \ldots, x_n)$ with

$$0 < \|\mathbf{x}\|^{n-1} \le c_n \|\mathbf{a}\|. \tag{1}$$

The only known exact values of c_n are $c_2 = 1$, $c_3 = \frac{4}{3}$ and $c_4 = \frac{27}{19}$ (see [1] and [15]). Note that for $n = 3, 4$ the equality in (1) is not attained. Schinzel [15] showed that, for $n \ge 3$,

$$c_n = \sup \Delta\big(\mathcal{H}^{n-1}_{\alpha_1,\ldots,\alpha_{n-3}}\big)^{-1} \ge 1,$$

where $\Delta(\cdot)$ denotes the critical determinant, $\mathcal{H}^{n-1}_{\alpha_1,\ldots,\alpha_{n-3}}$ is a generalized hexagon in $\mathbb{R}^{n-1}$ given by

$$|x_i| \le 1, \quad i = 1, \ldots, n-1, \quad \left|\sum_{i=1}^{n-3} \alpha_i x_i + x_{n-2} + x_{n-1}\right| \le 1,$$

and α_i range over all rational numbers in the interval $(0,\ 1]$. The values of c_n for $n \le 4$ indicate that, most likely, $c_n = \Delta(\mathcal{H}^{n-1}_{1,\ldots,1})^{-1}$. However, a proof of this conjecture does not seem within reach at present. The best known upper bound

$$c_n \le \sqrt{n} \tag{2}$$

follows from the classical result of Bombieri and Vaaler [3, Theorem 1].

In this paper we estimate c_n via values of the sinc integrals

$$\sigma_n = \frac{2}{\pi} \int_0^\infty \left(\frac{\sin t}{t}\right)^n dt.$$

The main result is as follows:

Theorem *For any nonzero vector* $\mathbf{a} \in \mathbb{Z}^n$, $n \ge 5$, *there exist linearly independent vectors* $\mathbf{x}_1, \ldots, \mathbf{x}_{n-1} \in \mathbb{Z}^n$ *such that* $L(\mathbf{x}_i) = 0$, $i = 1, \ldots, n-1$, *and*

$$\|\mathbf{x}_1\| \cdots \|\mathbf{x}_{n-1}\| < \frac{\|\mathbf{a}\|}{\sigma_n}. \tag{3}$$

From (3) we immediately get the bound

$$c_n \le \sigma_n^{-1}, \tag{4}$$

and since

$$\sigma_n^{-1} \sim \sqrt{\frac{\pi n}{6}}, \quad \text{as} \quad n \to \infty \tag{5}$$

(see Sect. 2), the theorem asymptotically improves the estimate (2). It is also known (see, e.g., [13]) that

$$\sigma_n = \frac{n}{2^{n-1}} \sum_{0 \le r < n/2,\, r \in \mathbb{Z}} \frac{(-1)^r (n-2r)^{n-1}}{r!\,(n-r)!}.$$

The sequences of numerators and denominators of $\sigma_n/2$ can be found in [16].

Remark 1

(i) Calculation shows that for all $5 \le n \le 1000$ the bound (4) is slightly better than (2).
(ii) For $n \le 4$ the constant σ_n^{-1} in (3) can be replaced by c_n. This follows from the observation that any origin-symmetric convex body in $\mathbb{R}^n$, $n \le 3$, has anomaly 1 (see [17]).

A. Schinzel (personal communication) observed that, with respect to maximum norm, Siegel's lemma can be applied to the following well-known problem from additive number theory. A finite set $\{a_1, \ldots, a_n\}$ of integers is called a *sum-distinct set* if any two of its 2^n subsums differ by at least 1. We shall assume, without loss of generality, that $0 < a_1 < a_2 < \cdots < a_n$. In 1955 Erdös and Moser [8, Problem 6] asked for an estimate on the least possible a_n of such a set. They proved that

$$a_n > \max\left\{\frac{2^n}{n}, \frac{2^n}{4\sqrt{n}}\right\} \tag{6}$$

and Erdös conjectured that $a_n > C_0 2^n$, $C_0 > 0$. In 1986 Elkies [7] showed that

$$a_n > 2^{-n}\binom{2n}{n} \tag{7}$$

and this result is still cited by Guy [11, Problem C8] as the best known lower bound for large n. Following [7], note that references [8] and [11] stated the problem equivalently in terms of an "inverse function". They asked one to maximize the size m of a sum-distinct subset of $\{1, 2, \ldots, x\}$, given x. Clearly, the bound $a_n > C_1 n^{-s} 2^n$ corresponds to

$$m < \log_2 x + s \log_2 \log_2 x + \log_2 \frac{1}{C_1} - o(1).$$

Corollary 1 *For any sum-distinct set* $\{a_1, \ldots, a_n\}$ *with* $0 < a_1 < \cdots < a_n$, *the inequality*

$$a_n > \sigma_n 2^{n-1} \tag{8}$$

holds.

Since

$$2^{-n}\binom{2n}{n} \sim \frac{2^n}{\sqrt{\pi n}} \quad \text{and} \quad \sigma_n 2^{n-1} \sim \frac{2^n}{\sqrt{2\pi n/3}}, \quad \text{as} \quad n \to \infty,$$

Corollary 1 asymptotically improves the result of Elkies with factor $\sqrt{3/2}$.

Remark 2

(i) Sum-distinct sets with a minimal largest element are known up to $n = 9$ (see [5]). In the latter case the estimate (8) predicts $a_9 \geq 116$ and the optimal bound is $a_9 \geq 161$. Calculation shows that for all $10 \leq n \leq 1000$ the bound (8) is slightly better than (7).

(ii) Professor Noam Elkies kindly informed the author about the existence of an unpublished result by him and Andrew Gleason which asymptotically improves (7) with factor $\sqrt{2}$.

2 Sections of the Cube and Sinc Integrals

Let $C = [-1, 1]^n \subset \mathbb{R}^n$ and let $\mathbf{s} = (s_1, \ldots, s_n) \in \mathbb{R}^n$ be a unit vector. It is a well-known fact (see, e.g., [2]) that

$$\operatorname{vol}_{n-1}\bigl(\mathbf{s}^{\perp} \cap C\bigr) = \frac{2^n}{\pi} \int_0^{\infty} \prod_{i=1}^{n} \frac{\sin s_i t}{s_i t}\, dt, \tag{9}$$

where $\mathbf{s}^{\perp}$ is the $(n-1)$-dimensional subspace orthogonal to $\mathbf{s}$. In particular, the volume of the section orthogonal to the vertex $\mathbf{v} = (1, \ldots, 1)$ of C is given by

$$\operatorname{vol}_{n-1}\bigl(\mathbf{v}^{\perp} \cap C\bigr) = \frac{2^n}{\pi} \int_0^{\infty} \left(\frac{\sin(t/\sqrt{n})}{t/\sqrt{n}} \right)^n dt = 2^{n-1} \sqrt{n}\, \sigma_n.$$

Laplace and Pólya (see [12, 14] and, e.g., [6]) both gave proofs that

$$\lim_{n \to \infty} \frac{\operatorname{vol}_{n-1}(\mathbf{v}^{\perp} \cap C)}{2^{n-1}} = \sqrt{\frac{6}{\pi}}.$$

Thus, (5) is justified.

Lemma 1 *For $n \geq 2$,*

$$0 < \sigma_{n+1} < \sigma_n \leq 1.$$

Proof This result is implicit in [4]. Indeed, Theorem 1(ii) of [4] applied with $a_0 = a_1 = \cdots = a_n = 1$ gives the inequalities

$$0 < \sigma_{n+1} \leq \sigma_n \leq 1.$$

The strict inequality $\sigma_{n+1} < \sigma_n$ follows from the observation that in this case the inequality in (3) of [4] is strict with $a_{n+1} = a_0 = y = 1$. □

3 An Application of the Busemann Theorem

Let $|\cdot|$ denote the euclidean norm. Recall that we can associate with each star body L the *distance function* $f_L(\mathbf{x}) = \inf\{\lambda > 0 : \mathbf{x} \in \lambda L\}$. The *intersection body IL* of a star body $L \subset \mathbb{R}^n$, $n \geq 2$, is defined as the $\mathbf{o}$-symmetric star body whose distance function f_{IL} is given by

$$f_{IL}(\mathbf{x}) = \frac{|\mathbf{x}|}{\mathrm{vol}_{n-1}(\mathbf{x}^{\perp} \cap L)}.$$

Intersection bodies played an important role in the solution to the famous Busemann–Petty problem. The Busemann theorem (see, e.g., Chap. 8 of [9]) states that if L is $\mathbf{o}$-symmetric and convex, then IL is the convex set. This result allows us to prove the following useful inequality. Let $f = f_{IC}$ denote the distance function of IC.

Lemma 2 *For any nonzero* $\mathbf{x} \in \mathbb{R}^n$,

$$f\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right) \leq f(\mathbf{v}) = \frac{1}{\sigma_n 2^{n-1}}, \tag{10}$$

with equality only if $n = 2$ *or* $\mathbf{x}/\|\mathbf{x}\|$ *is a vertex of the cube* C.

We proceed by induction on n. When $n = 2$ the result is obvious. Suppose now (10) is true for $n - 1 \geq 2$. Since, if some $x_i = 0$, the problem reduced to that in $\mathbb{R}^{n-1}$, we may assume inductively that $x_i > 0$ for all i. Clearly, we may also assume that $\mathbf{w} = \mathbf{x}/\|\mathbf{x}\|$ is not a vertex of C, in particular, $\mathbf{w} \neq \mathbf{v}$.

Let $Q = [0, 1]^n \subset \mathbb{R}^n$ and let L be the two-dimensional subspace spanned by vectors $\mathbf{v}$ and $\mathbf{x}$. Then $P = L \cap Q$ is a parallelogram on the plane L. To see this, observe that the cube Q is the intersection of two cones $\{\mathbf{y} \in \mathbb{R}^n : y_i \geq 0\}$ and $\{\mathbf{y} \in \mathbb{R}^n : y_i \leq 1\}$ with apexes at the points $\mathbf{o}$ and $\mathbf{v}$, respectively.

Suppose that P has vertices $\mathbf{o}$, $\mathbf{u}$, $\mathbf{v}$, $\mathbf{v} - \mathbf{u}$. Then the edges $\mathbf{ou}$, $\mathbf{ov} - \mathbf{u}$ of P belong to coordinate hyperplanes and the edges $\mathbf{uv}$, $\mathbf{vv} - \mathbf{u}$ lie on the boundary of C. Without loss of generality, we may assume that the point $\mathbf{w}$ lies on the edge $\mathbf{uv}$. Let

$$\mathbf{v}' = \sigma_n \mathbf{v} = \frac{\mathrm{vol}_{n-1}(\mathbf{v}^{\perp} \cap C)}{2^{n-1}} \frac{\mathbf{v}}{|\mathbf{v}|} \in \frac{1}{2^{n-1}} IC,$$

$$\mathbf{u}' = \sigma_{n-1} \mathbf{u}.$$

Since the point $\mathbf{u}$ lies in one of the coordinate hyperplanes, by the induction hypothesis

$$f(\mathbf{u}') = f(\sigma_{n-1}\mathbf{u}) \leq \frac{1}{2^{n-1}}.$$

Thus, $\mathbf{u}' \in (1/2^{n-1})IC$. Consider the triangle with vertices $\mathbf{o}$, $\mathbf{u}$, $\mathbf{v}$. Let $\mathbf{w}'$ be the point of intersection of segments $\mathbf{ow}$ and $\mathbf{u}'\mathbf{v}'$. Observing that by Lemma 10

$$|\sigma_n \mathbf{w}| < |\mathbf{w}'| < |\sigma_{n-1}\mathbf{w}|,$$

we get

$$\frac{1}{\sigma_{n-1}} < \frac{|\mathbf{w}|}{|\mathbf{w}'|} < \frac{1}{\sigma_n}. \tag{11}$$

By the Busemann theorem IC is convex. Therefore $\mathbf{w}' \in (1/2^{n-1})IC$ and thus

$$|\mathbf{w}'| \le \frac{\mathrm{vol}_{n-1}(\mathbf{w}^\perp \cap C)}{2^{n-1}}.$$

By (11) we obtain

$$f\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right) = f(\mathbf{w}) = \frac{|\mathbf{w}|}{\mathrm{vol}_{n-1}(\mathbf{w}^\perp \cap C)} \le \frac{|\mathbf{w}|}{2^{n-1}|\mathbf{w}'|} < \frac{1}{\sigma_n 2^{n-1}}.$$

Applying Lemma 2 to a unit vector $\mathbf{s}$ and using (9) we get the following inequality for sinc integrals.

Corollary 2 *For any unit vector* $\mathbf{s} = (s_1, \dots, s_n) \in \mathbb{R}^n$,

$$\|\mathbf{s}\| \int_0^\infty \prod_{i=1}^n \frac{\sin s_i t}{s_i t}\, dt \ge \int_0^\infty \left(\frac{\sin t}{t}\right)^n dt,$$

with equality only if $n = 2$ *or* $\mathbf{s}/\|\mathbf{s}\|$ *is a vertex of the cube* C.

Remark 3 Note that IC is symmetric with respect to any coordinate hyperplane. This observation and Busemann's theorem immediately imply (10) with nonstrict inequality in all cases.

4 Proof of the Theorem

Clearly, we may assume that $\|\mathbf{a}\| > 1$ and, in particular, that the inequality in Lemma 2 is strict for $\mathbf{x} = \mathbf{a}$. We also assume, without loss of generality, that $\gcd(a_1, \dots, a_n) = 1$.

Let $S = \mathbf{a}^\perp \cap C$ and $\Lambda = \mathbf{a}^\perp \cap \mathbb{Z}^n$. Then S is a centrally symmetric convex set and Λ is an $(n-1)$-dimensional sublattice of $\mathbb{Z}^n$ with determinant (covolume) $\det \Lambda = |\mathbf{a}|$. Let $\lambda_i = \lambda_i(S, \Lambda)$ be the ith successive minimum of S with respect to Λ, that is

$$\lambda_i = \inf\bigl\{\lambda > 0 : \dim(\lambda S \cap \Lambda) \ge i\bigr\}.$$

By the definition of S and Λ it is enough to show that

$$\lambda_1 \cdots \lambda_{n-1} < \frac{\|\mathbf{a}\|}{\sigma_n}.$$

The $(n-1)$-dimensional subspace $\mathbf{a}^\perp \subset \mathbb{R}^n$ can be considered as a usual $(n-1)$-dimensional Euclidean space. The Minkowski Theorem on Successive Minima (see,

e.g. Chap. 2 of [10]), applied to the **o**-symmetric convex set $S \subset \mathbf{a}^{\perp}$ and the lattice $\Lambda \subset \mathbf{a}^{\perp}$, implies that

$$\lambda_1 \cdots \lambda_{n-1} \leq \frac{2^{n-1} \det \Lambda}{\mathrm{vol}_{n-1}(S)} = \frac{2^{n-1}|\mathbf{a}|}{\mathrm{vol}_{n-1}(\mathbf{a}^{\perp} \cap C)} = 2^{n-1} f(\mathbf{a}),$$

and by Lemma 2 we get

$$\lambda_1 \cdots \lambda_{n-1} \leq 2^{n-1} f(\mathbf{a}) = 2^{n-1} f\left(\frac{\mathbf{a}}{\|\mathbf{a}\|}\right) \|\mathbf{a}\| < 2^{n-1} f(\mathbf{v}) \|\mathbf{a}\| = \frac{\|\mathbf{a}\|}{\sigma_n}.$$

This proves the theorem.

5 Proof of Corollary 1

For a sum-distinct set $\{a_1, \ldots, a_n\}$ consider the vector $\mathbf{a} = (a_1, \ldots, a_n)$. Observe that any nonzero integral vector $\mathbf{x}$ with $L(\mathbf{x}) = 0$ must have the maximum norm greater than 1. Therefore (3) implies the inequality

$$2^{n-1} < \frac{\|\mathbf{a}\|}{\sigma_n}.$$

Acknowledgements The author thanks Professors D. Borwein and A. Schinzel for valuable comments and Professor P. Gruber for fruitful discussions and suggestions.

References

1. Aliev, I.: On a decomposition of integer vectors. Ph.D. Dissertation, Institute of Mathematics, PAN, Warsaw (2001)
2. Ball, K.: Cube slicing in $\mathbb{R}^n$. Proc. Am. Math. Soc. **97**(3), 465–472 (1986)
3. Bombieri, E., Vaaler, J.: On Siegel's lemma. Invent. Math. **73**, 11–32 (1983). Addendum. Invent. Math. **75**, 377 (1984)
4. Borwein, D., Borwein, J.: Some remarkable properties of sinc and related integrals. Ramanujan J. **5**(1), 73–89 (2001)
5. Borwein, P., Mossinghoff, M.: Newman polynomials with prescribed vanishing and integer sets with distinct subset sums. Math. Comput. **72**(242), 787–800 (2003); (electronic)
6. Chakerian, D., Logothetti, D.: Cube slices, pictorial triangles, and probability. Math. Mag. **64**(4), 219–241 (1991)
7. Elkies, N.D.: An improved lower bound on the greatest element of a sum-distinct set of fixed order. J. Comb. Theory Ser. A **41**(1), 89–94 (1986)
8. Erdös, P.: Problems and results in additive number theory. In: Colloque sur la Théorie des Nombres, pp. 127–137, Bruxelles (1955)
9. Gardner, R.J.: Geometric Tomography. Encyclopedia of Mathematics and Its Applications, vol. 58. Cambridge University Press, Cambridge (1995)
10. Gruber, P.M., Lekkerkerker, C.G.: Geometry of Numbers. North-Holland, Amsterdam (1987)
11. Guy, R.K.: Unsolved Problems in Number Theory, 3rd edn. Problem Books in Mathematics. Unsolved Problems in Intuitive Mathematics. Springer, New York (2004)
12. Laplace, P.S.: Théorie Analytique des Probabilités. Courcier Imprimeur, Paris (1812)
13. Medhurst, R.G., Roberts, J.H.: Evaluation of the integral $I_n(b) = (2/\pi) \int_0^{\infty} ((\sin x)/x)^n \cos(bx)\, dx$. Math. Comput. **19**, 113–117 (1965)
14. Pólya, G.: Berechnung eines Bestimmten Integrals. Math. Ann. **74**, 204–212 (1913)

15. Schinzel, A.: A property of polynomials with an application to Siegel's lemma. Mon.hefte Math. **137**, 239–251 (2002)
16. Sloane, N.J.A.: Sequences A049330 and A049331. In: The On-Line Encyclopedia of Integer Sequences, http://www.research.att.com/~njas/sequences/
17. Woods, A.C.: The anomaly of convex bodies. Proc. Camb. Philos. Soc. **52**, 406–423 (1956)

Slicing Convex Sets and Measures by a Hyperplane

Imre Bárány · Alfredo Hubard · Jesús Jerónimo

Abstract Given convex bodies $K_1, \dots, K_d$ in $\mathbb{R}^d$ and numbers $\alpha_1, \dots, \alpha_d \in [0, 1]$, we give a sufficient condition for existence and uniqueness of an (oriented) halfspace H with $\mathrm{Vol}(H \cap K_i) = \alpha_i \cdot \mathrm{Vol}\, K_i$ for every i. The result is extended from convex bodies to measures.

Keywords Convex bodies · Well separated families · Sections of convex sets and measures

1 Transversal Spheres

A well known result in elementary geometry states that there is a unique sphere which contains a given set of $d + 1$ points in general position in $\mathbb{R}^d$. A similar thing happens with d-pointed sets and hyperplanes. What happens if we consider convex bodies instead of points?

I. Bárány (✉)
Rényi Institute of Mathematics, Hungarian Academy of Sciences, P.O. Box 127, 1364 Budapest, Hungary
e-mail: barany@renyi.hu

I. Bárány
Department of Mathematics, University College London, Gower Street, London WC1E 6BT, England

A. Hubard
Instituto de Matemáticas, UNAM, Ciudad Universitaria, Mexico D.F. 04510, Mexico
e-mail: hubard@cims.nyu.edu

J. Jerónimo
Centro de Investigación en Matemáticas A.C., Apdo. Postal 402, Guanajuato, Mexico
e-mail: jeronimo@cimat.mx

These questions are the main motivation for the present paper. The first result in this direction is due to Kramer and Németh [7]. They used the following, very natural definition.

A family $\mathcal{F}$ of connected sets in $\mathbb{R}^d$ is said to be *well separated*, if for any $k \leq d+1$ distinct elements, $K_1, \ldots, K_k$, of $\mathcal{F}$ and for any choice of points $x_i \in K_i$, the set $\operatorname{aff}\{x_1, \ldots, x_k\}$ is a $(k-1)$-dimensional flat. Here $[k]$ stands for the set $\{1, 2, \ldots, k\}$. It is well known (cf. [1, 4]), and also easy to check the following.

Proposition 1 *Assume $\mathcal{F} = \{K_1, \ldots, K_n\}$ is a family of connected sets in $\mathbb{R}^d$. The following conditions are equivalent*:

1. *The family $\mathcal{F}$ is well separated.*
2. *The family $\mathcal{F}' = \{\operatorname{conv} K_1, \ldots, \operatorname{conv} K_n\}$ is well separated.*
3. *For every pair of disjoint sets $I, J \subset [n]$ with $|I| + |J| \leq d+1$, there is a hyperplane separating the sets K_i, $i \in I$ from the sets K_j, $j \in J$.*

By an elegant application of Brouwer's fixed point theorem, Kramer and Németh proved the following:

Theorem KN *Let $\mathcal{F}$ be a well separated family of $d+1$ compact convex sets in $\mathbb{R}^d$. Then there exists a unique Euclidean ball which touches each set and whose interior is disjoint from each member of $\mathcal{F}$.*

Denote by $B(x, r)$, resp. $S(x, r)$, the Euclidean ball and sphere of radius r and center x. We say that the sphere $S(x, r)$ *supports* a compact set K if $S(x, r) \cap K \neq \emptyset$ and either $K \subset B(x, r)$ or $K \cap \operatorname{int} B(x, r) = \emptyset$. This definition is due to Klee et al. [6]. They proved the following:

Theorem KLH *Let $\mathcal{F} = \{K_1, K_2, \ldots, K_{d+1}\}$ be a well separated family of compact convex sets in $\mathbb{R}^d$, and let I, J be a partition of $[d+1]$. Then there is a unique Euclidean sphere $S(x, r)$ that supports each element of $\mathcal{F}$ in such a way that $K_i \subset B(x, r)$ for each $i \in I$ and $K_j \cap \operatorname{int} B(x, r) = \emptyset$ for each $j \in J$.*

The case $I = \emptyset$ corresponds to Theorem KN. We are going to generalize these results. Let $Q^d = [0, 1]^d$ denote the unit cube of $\mathbb{R}^d$. Given a well separated family $\mathcal{F}$ of convex sets in $\mathbb{R}^d$, a sphere $S(x, r)$ is said to be *transversal* to $\mathcal{F}$ if it intersects every element of $\mathcal{F}$. Finally, a *convex body* in $\mathbb{R}^d$ is a convex compact set with nonempty interior.

Theorem 1 *Let $\mathcal{F} = \{K_1, \ldots, K_{d+1}\}$ be a well separated family of convex bodies in $\mathbb{R}^d$, and let $\alpha = (\alpha_1, \ldots, \alpha_{d+1}) \in Q^{d+1}$. Then there exists a unique transversal Euclidean sphere $S(x, r)$ such that $\operatorname{Vol}(B(x, r) \cap K_i) = \alpha_i \cdot \operatorname{Vol}(K_i)$ for every $i \in [d+1]$.*

Remark 1 The transversality of $S(x, r)$ only matters when α_i is equal to 0 or 1; otherwise the condition $\operatorname{Vol}(B(x, r) \cap K_i) = \alpha_i \cdot \operatorname{Vol}(K_i)$ plus convexity guarantees that $S(x, r)$ intersects K_i.

2 Transversal Hyperplanes and Halfspaces

In a similar direction, Cappell et al. [3] proved an analogous theorem for the case of supporting hyperplanes, which can be seen as spheres of infinite radius. Given a family $\mathcal{F}$ of sets in $\mathbb{R}^d$, a hyperplane will be called *transversal* to $\mathcal{F}$ if it intersects each member of $\mathcal{F}$. The following result is a special case of Theorem 3 of Cappell et al. [3] (cf. [2] as well):

Theorem C *Let $\mathcal{F} = \{K_1, \ldots, K_d\}$ be a well separated family of compact convex sets in $\mathbb{R}^d$ with a partition I, J of the index set $[d]$. Then there are exactly two hyperplanes, H_1 and H_2, transversal to $\mathcal{F}$ such that both H_1 and H_2 have all K_i $(i \in I)$ on one side and all K_j $(j \in J)$ on the other side.*

Theorem C was also proved by Klee et al. [5] using Kakutani's extension of Brouwer's fixed point theorem. We are going to formulate this theorem in a slightly different way, more suitable for our purposes. So, we need to introduce new notation and terminology.

A halfspace H in $\mathbb{R}^d$ can be specified by its outer unit normal vector, v, and by the signed distance, $t \in \mathbb{R}$, of its bounding hyperplane from the origin. Thus, there is a one-to-one correspondence between halfspaces of $\mathbb{R}^d$ and pairs $(v, t) \in S^{d-1} \times \mathbb{R}$. We denote the halfspace $\{x \in \mathbb{R}^d : \langle x, v\rangle \leq t\}$ by $H(v \leq t)$. Analogously we write $H(v = t) = \{x \in \mathbb{R}^d : \langle x, v\rangle = t\}$, which is the bounding hyperplane of $H(v \leq t)$. Furthermore, given a set $K \subset \mathbb{R}^d$, a unit vector v and a scalar t, we denote the set $H(v = t) \cap K$ by $K(v = t)$, analogously $K(v \leq t) = H(v \leq t) \cap K$.

Suppose next that $\mathcal{F} = \{K_1, \ldots, K_d\}$ is a well separated family of convex sets in $\mathbb{R}^d$. Assume $a_1 \in K_1, \ldots, a_d \in K_d$. The unit normal vectors to the unique transversal hyperplane containing these points are v and $-v$. We want to make the choice between v and $-v$ unique and depend only on $\mathcal{F}$. We first make it depend on $a_1, \ldots, a_d$. Define $v = v(a_1, \ldots, a_d)$ as the (unique) unit normal vector to $\operatorname{aff}\{a_1, \ldots, a_d\}$ satisfying

$$\det \begin{vmatrix} v & a_1 & a_2 & \cdots & a_d \\ 0 & 1 & 1 & \cdots & 1 \end{vmatrix} > 0,$$

in other words, the points $v + a_1, a_1, a_2, \ldots, a_d$, in this order, are the vertices of a positively oriented d-dimensional simplex. Clearly, with $-v$ in place of v the determinant would be negative. This gives rise to the map $v : K \longrightarrow S^{d-1}$ where $K = K_1 \times \cdots \times K_d$. This definition seems to depend on the choice of the a_i, but in fact, it does not. Write $H(v = t) = \operatorname{aff}\{a_1, \ldots, a_d\}$.

Proposition 2 *Under the previous assumption, let $b_i \in K_i(v = t)$ for each i. Then $v(a_1, \ldots, a_d) = v(b_1, \ldots, b_d)$.*

Proof This is simple. The homotopy $(1 - \lambda)a_i + \lambda b_i$ $(\lambda \in [0, 1])$ moves the a_i to the b_i continuously, and keeps $(1-\lambda)a_i + \lambda b_i$ in $K_i(v = t)$. The affine hull of the moving points remains unchanged, and does not degenerate because $\mathcal{F}$ is well separated. So their outer unit normal remains v throughout the homotopy. □

The previous proposition is also mentioned by Klee et al. [5]. With this definition, a transversal hyperplane to $\mathcal{F}$ determines v and t uniquely. We call $H(v = t)$ *a positive transversal hyperplane to* $\mathcal{F}$, and similarly, $H(v \leq t)$ is *a positive transversal halfspace to* $\mathcal{F}$.

Theorem 2 *Let $\mathcal{F} = \{K_1, \ldots, K_d\}$ be a family of well separated convex bodies in $\mathbb{R}^d$, and let $\alpha = (\alpha_1, \ldots, \alpha_d) \in Q^d$. Then there is a unique positive transversal halfspace, H, such that $\mathrm{Vol}(K_i \cap H) = \alpha_i \cdot \mathrm{Vol}(K_i)$ for every $i \in [d]$.*

Theorem C follows since the partition I, J gives rise to $\alpha, \beta \in Q^d$ via $\alpha_k = 1$ if $k \in I$, otherwise $\alpha_k = 0$, and $\beta_k = 1$ if $k \in J$, otherwise $\beta_k = 0$. By Theorem 2, there are unique positive transversal halfspaces $H(\alpha)$ and $H(\beta)$ with the stated properties. Their bounding hyperplanes satisfy the statement of Theorem C and they are obviously distinct. We mention, however, that Theorem C will be used in the proof of the unicity part of Theorem 2.

Remark 2 When all $\alpha_i = 1/2$, the existence of such a halfspace is guaranteed by Borsuk's theorem, even without the condition of convexity or $\mathcal{F}$ being well separated. (Connectivity of the sets implies that the halving hyperplane is a transversal to $\mathcal{F}$.) The case of general α_i, however, needs some extra condition as the following two examples show. If all K_i are equal, then each oriented hyperplane section cuts off the same amount from each K_i, so $\alpha_1 = \cdots = \alpha_d$ must hold. The second example consists of d concentric balls with different radii, and if the radius of the first ball is very large compared to those of the others and α_1 is too small, then a hyperplane cutting off α_1 fraction of the first ball is disjoint from all other balls. Thus no hyperplane transversal exists that cuts off an α_1 fraction of the first set.

Remark 3 Cappell et al. prove, in fact, a much more general theorem [3]. Namely, assume that $\mathcal{F}$ is well separated and consists of k strictly convex sets, $k \in \{2, \ldots, d\}$ and let I, J be a partition of $[k]$. Then the set of all supporting hyperplanes separating the K_i $(i \in I)$ from the K_j $(j \in J)$ is homeomorphic to the $(d - k)$-dimensional sphere.

3 Extension to Measures

Borsuk's theorem holds not only for volumes but more generally for measures. Similarly, our Theorem 2 can and will be extended to *nice measures* that we are to define soon. We need a small piece of notation.

Let μ be a finite measure on the Borel subsets of $\mathbb{R}^d$ and let $v \in S^{d-1}$ be a unit vector. Define

$$t_0 = t_0(v) = \inf\{t \in \mathbb{R} : \ \mu(H(v \leq t)) > 0\},$$
$$t_1 = t_1(v) = \sup\{t \in \mathbb{R} : \ \mu(H(v \leq t)) < \mu(\mathbb{R}^d)\}.$$

Note that $t_0 = -\infty$ and $t_1 = \infty$ are possible.

Let $H(s_0 \le v \le s_1)$ denote the closed slab between the hyperplanes $H(v = s_0)$ and $H(v = s_1)$. Define the set K by

$$K = \bigcap_{v \in S^{d-1}} H(t_0(v) \le v \le t_1(v)).$$

K is called the *support* of μ. Note that K is convex (obviously) and $\mu(\mathbb{R}^d \setminus K) = 0$.

Definition 1 The measure μ is called *nice* if the following conditions are satisfied:

(i) $t_0(v)$ and $t_1(v)$ are finite for every $v \in S^{d-1}$,
(ii) $\mu(H(v = t)) = 0$ for every $v \in S^{d-1}$ and $t \in \mathbb{R}$,
(iii) $\mu(H(s_0 \le v \le s_1)) > 0$ for every $v \in S^{d-1}$ and for every s_0, s_1 satisfying $t_0(v) \le s_0 < s_1 \le t_1(v)$.

If μ is a nice measure, then its support is full-dimensional since, by (ii), it is not contained in any hyperplane.

The function $t \mapsto \mu(K(v \le t))$ is zero on the interval $(-\infty, t_0]$, is equal to $\mu(K)$ on $[t_1, \infty)$, strictly increases on $[t_0, t_1]$, and, in view of (iii), is continuous. Assume $\alpha \in [0, 1]$. Then there is a unique $t \in [t_0, t_1]$ with

$$\mu(K(v \le t)) = \alpha \cdot \mu(K).$$

Denote this unique t by $g(v)$; this way we defined a map $g : S^{d-1} \longrightarrow \mathbb{R}$. The following simple lemma is important and probably well known.

Lemma 1 *For fixed $\alpha \in [0, 1]$ the function g is continuous.*

Proof When $\alpha = 1$, $g(v)$ is the support functional of K, which is not only continuous but convex (when extended to all $v \in \mathbb{R}^d$). Similarly, g is continuous when $\alpha = 0$.

Assume now that $0 < \alpha < 1$. Let $v_0 \in S^{d-1}$ be an arbitrary point. In order to prove the continuity of g at v_0 we show first that $K(v = g(v))$ and $K(v_0 = g(v_0))$ have a point in common whenever v it is close enough to v_0.

Obviously, $K(v_0 = g(v_0))$ is a $(d-1)$-dimensional convex set lying in the hyperplane $H(v_0 = g(v_0))$. Then, for every small enough neighbourhood of v_0, and for each v in such a neighbourhood, the supporting hyperplane of K with unit normal v (and $-v$) is also a supporting hyperplane of $K(v_0 \ge g(v_0))$ (and $K(v_0 \le g(v_0))$).

Assume $s_v \le S_v$ and let $H(v = s_v)$ and $H(v = S_v)$ be the two supporting hyperplanes (with normal v) to $K(v_0 = g(v_0))$ which is a $(d-1)$-dimensional convex set. Since $K(v_0 = g(v_0))$ is a $(d-1)$-dimensional convex set, condition (iii) implies that $s_v < S_v$. It follows that

$$K(v \le s_v) \subset K(v_0 \le g(v_0)) \subset K(v \le S_v),$$

and so

$$\mu(K(v \le s_v)) \le \mu(K(v_0 \le g(v_0)) \le \mu(K(v \le S_v)).$$

As $\mu(K(v_0 \le g(v_0)) = \alpha \cdot \mu(K)$, we have $s_v \le g(v) \le S_v$. Consequently, $K(v = g(v))$ and $K(v_0 = g(v_0))$ have a point, say $z = z(v)$, in common. This $z(v)$ is not uniquely determined but that does not matter.

It is easy to finish the proof now. Clearly $g(v) = \langle v, z(v)\rangle$ and $g(v_0) = \langle v_0, z(v)\rangle$ for all v in a small neighbourhood of v_0. Assume the sequence v_n tends to v_0. We claim that every subsequence, $v_{n'}$, of v_n contains a subsequence $v_{n''}$ such that $\lim g(v_{n''}) = g(v_0)$, which evidently implies the continuity of g at v_0.

For the proof of this claim observe first that, since $K(v_0 = g(v_0))$ is compact, $z(v_{n'})$ contains a convergent subsequence $z(v_{n''})$ tending to z_0, say. Taking limits gives $z_o \in K(v_0 = g(v_0))$. Then $g(v_{n''}) = \langle v_{n''}, z(v_{n''})\rangle \to \langle v_0, z_0\rangle = g(v_0)$. □

Theorem 2 is extended to measures in the following way.

Theorem 3 *Suppose μ_i is a nice measure on $\mathbb{R}^d$ with support K_i for all $i \in [d]$. Assume the family $\mathcal{F} = \{K_1, \dots, K_d\}$ is well separated and let $\alpha = (\alpha_1, \dots, \alpha_d) \in Q^d$. Then there is a unique positive transversal halfspace, H, such that $\mu_i(K_i \cap H) = \alpha_i \cdot \mu_i(K_i)$, for every $i \in [d]$.*

Corollary 1 *Assume μ_i are finite measures on $\mathbb{R}^d$ satisfying conditions (i) and (ii) of Definition 1. Let K_i be the support of μ_i for all $i \in [d]$. Suppose the family $\mathcal{F} = \{K_1, \dots, K_d\}$ is well separated and let $\alpha = (\alpha_1, \dots, \alpha_d) \in Q^d$. Then there is a positive transversal halfspace, H, such that $\mu_i(K_i \cap H) = \alpha_i \cdot \mu_i(K_i)$, for every $i \in [d]$.*

The corollary easily follows from Theorem 3; we omit the simple details.

Theorem 2 is a special case of Theorem 3: when μ_i is the Lebesgue measure (or volume) restricted to the convex body K_i for all $i \in [d]$ and the family $\mathcal{F}$ is well separated. Also, Theorem C is a special case of Theorem 3: when μ_i and K_i are the same as above, and, for a given partition I, J of $[d]$, we set $\alpha_i = 1$ for $i \in I$, and $\alpha_j = 0$ for $j \in J$. Theorem 1 follows from Theorem 3 via "lifting to the paraboloid". This is explained in the last section.

4 Proof of Theorem 3

In the proof we will use Brouwer's fixed point theorem. We will define a continuous mapping from a topological ball to itself, such that a fixed point of this map yields a halfspace with the desired properties. Set $K = K_1 \times \cdots \times K_d$. Given a point $x = (x_1, \dots, x_d) \in K$ we consider the hyperplane $\operatorname{aff}\{x_1, \dots, x_d\}$. Since the family $\mathcal{F}$ is well separated, this hyperplane is well defined for each $x \in K$. Let $H(v \le t)$ be the (unique) positive transversal halfspace whose bounding hyperplane is $\operatorname{aff}\{x_1, \dots, x_d\}$.

In Sect. 2 we defined the map $v : K \longrightarrow S^{d-1}$ which is the properly chosen unit normal to $\operatorname{aff}\{x_1, \dots, x_d\}$. Clearly, this function is continuous.

We prove existence first. We start with the case when $\alpha_i \in (0, 1)$ for every $i \in [d]$. We turn to the remaining case later by constructing a suitable sequence of halfspaces.

Let $g_i : S^{d-1} \longrightarrow \mathbb{R}$ be the function such that for each $v \in S^{d-1}$, $g_i(v)$ is the real number for which $\mu_i(K_i(v \leq g_i(v))) = \alpha \cdot \mu_i(K_i)$ for each $i \in [d]$. Each g_i is a continuous function by Lemma 1. Let $h : S^{d-1} \longrightarrow K$ be the function sending $v \mapsto (s_1, \ldots, s_d)$ where s_i is the Steiner point of the $(d-1)$-dimensional section, $K_i(v = g_i(v))$ for each $i \in [d]$. As is well known, the family of sections $K_i(v = t)$ depend continuously (according to the Hausdorff metric) on the corresponding family of hyperplanes, $\{H(v = t)\}$ whenever every section is $(d-1)$-dimensional, which is obviously the case because $\alpha_i \in (0, 1)$. It is also well known that the function that assigns to a compact convex set its Steiner point is continuous. Hence, h is a continuous function.

It follows that

$$f := h \circ v : K \longrightarrow K$$

is a continuous function. As K is a compact convex set in $\mathbb{R}^d \times \cdots \times \mathbb{R}^d$ Brouwer's fixed point theorem implies the existence of a point $x \in K$ such that $f(x) = x$. Consider a fixed point, $x = (x_1, \ldots, x_d)$, of f. Then the halfspace $H(v \leq t)$ whose bounding hyperplane is aff$\{x_1, \ldots, x_d\}$ is a positive transversal halfspace to $\mathcal{F}$ and it has the required properties.

Next we prove existence for vectors $\alpha = (\alpha_1, \ldots, \alpha_d) \in Q^d$ that may have 0, 1 components as well. Consider the sequence $\{\alpha^n\} \subset Q^d$ $\alpha^n = (\alpha_1^n, \ldots, \alpha_d^n)$ (defined for every $n \geq 2$), such that for every entry $\alpha_i = 0$ we define $\alpha_i^n = \frac{1}{n}$, for every entry $\alpha_i = 1$ we define $\alpha_i^n = 1 - \frac{1}{n}$, and for every entry $\alpha_i \notin \{0, 1\}$ we define $\alpha_i^n = \alpha_i$. Also, for every $n \geq 2$ we consider the unique positive transversal halfspace $H(v_n \leq t_n)$ with $\mu_i(K_i(v_n \leq t_n)) = \alpha_i^n \cdot \mu_i(K_i)$, for each i. The compactness of K implies that the set of all possible $(v, t) \in S^{d-1} \times \mathbb{R}$ such that the hyperplane $H(v = t)$ is transversal to $\mathcal{F}$ is compact. Thus there exists a convergent subsequence $\{(v_{n'}, t_{n'})\}$ which converges to a point $(v, t) \in S^{d-1} \times \mathbb{R}$. Clearly, $H(v \leq t)$ is a positive transversal halfspace to $\mathcal{F}$ which satisfies $\mu_i(K_i(v \leq t)) = \alpha_i \cdot \mu_i(K_i)$ for every i.

Next comes uniqueness. We start with the 0, 1 case, that is, when $\alpha = (\alpha_1, \ldots, \alpha_d)$ with all $\alpha_i \in \{0, 1\}$. Such an α defines a $\beta \in Q^d$ via $\beta_i = 1 - \alpha_i$ for every i. By the previous existence proof there is a unique positive transversal halfspace $H(v \leq t)$ for α and another one $H(u \leq s)$ for β. These two halfspaces are distinct, first because $u = v$ is impossible, and second because of the following fact which implies that $u \neq -v$.

Proposition 3 *For every pair of points $(a_1, \ldots, a_d)$ and $(b_1, \ldots, b_d)$ in K, $v(a_1, \ldots, a_d)$ and $-v(b_1, \ldots, b_d)$ are distinct.*

Proof Assume $v(a_1, \ldots, a_d) = -v(b_1, \ldots, b_d)$. Then the affine hulls of the a_i and the b_i are parallel hyperplanes. We use the same homotopy as in the proof of Proposition 2. As λ moves from 0 to 1, the moving points $(1 - \lambda)a_i + \lambda b_i$ stay in K_i, and their affine hull remains parallel with aff$\{a_1, \ldots, a_d\}$. So the outer normal remains unchanged throughout the homotopy. A contradiction. □

The condition $\mu_i(K_i(v \leq t)) = \alpha_i \mu_i(K_i)$ implies, in the given case, that all K_i $(i \in I)$ are in $H(v \leq t)$ and all K_j $(j \in J)$ are in $H(v \geq t)$. Thus $H(v = t)$ is a

transversal hyperplane satisfying the conditions of Theorem C with partition I, J where $I = \{i \in [d] : \alpha_i = 0\}$ and $J = \{j \in [d] : \alpha_j = 1\}$. The same way, $H(u = s)$ is a transversal hyperplane satisfying the conditions of Theorem C with the same partition J, I.

The uniqueness of $H(v \le t)$ follows now easily. If we had two distinct positive transversal halfspaces $H(v_1 \le t_1)$ and $H(v_2 \le t_2)$ for α, then we would have four distinct transversal hyperplanes with K_i $(i \in I)$ on one side and K_j $(j \in J)$ on the other side, contradicting Theorem C.

Now we turn to uniqueness for general α. Assume that there are two distinct positive transversal halfspaces $H(v_1 \le t_1)$ and $H(v_2 \le t_2)$ for α. Their bounding hyperplanes cannot be parallel. Define $M = H(v_1 \le t_1) \cap H(v_2 \le t_2)$ and $N = H(v_1 \ge t_1) \cap H(v_2 \ge t_2)$. The partition I, J of the index set $[d]$ is defined as follows: $i \in I$ if $M \cap \operatorname{int} K_i \ne \emptyset$ and $j \in J$ if $M \cap \operatorname{int} K_j = \emptyset$. Set $K_i' = M \cap K_i$ for every $i \in I$ and $K_j' = N \cap K_j$ for every $j \in J$. Let $\mathcal{F}'$ be the family consisting of all the convex bodies K_i' $(i \in I)$ and K_j' $(j \in J)$. It is quite easy to see that no member of $\mathcal{F}'$ is empty. Moreover, $\mathcal{F}'$ is evidently well separated. Given the partition I, J, define γ by $\gamma_i = 1$ for $i \in I$ and $\gamma_j = 0$ for $j \in J$. Then there are two transversal halfspaces (with respect to $\mathcal{F}'$), namely $H(v_k \le t_k)$ $k = 1, 2$ satisfying $\mu_i(K_i(v_k \le t_k)) = \gamma_i \mu_i(K_i)$ for every i. But every $\gamma_i \in \{0, 1\}$ and we just established uniqueness in the 0, 1 case. □

5 Proof of Theorem 2

We will use the well-known technique of lifting the problem from $\mathbb{R}^d$ to a paraboloid in $\mathbb{R}^{d+1}$, and then apply Theorem 3.

In this section we change notation a little. A point in $\mathbb{R}^d$ is denoted by $x = (x_1, \ldots, x_d)$, a point in $\mathbb{R}^{d+1}$ is denoted by $\overline{x} = (x_1, \ldots, x_d, x_{d+1})$. The projection of $\overline{x}$ is $\pi(\overline{x}) = (x_1, \ldots, x_d)$, and the lifting of x is $\ell(x) = (x_1, \ldots, x_d, |x|^2)$ where $|x|^2 = x_1^2 + \cdots + x_d^2$. Clearly, $\ell(x)$ is contained in the paraboloid

$$P = \{\overline{x} \in \mathbb{R}^{d+1} : \overline{x} = (x_1, x_2, \ldots, x_d, |x|^2)\}.$$

A set $K \subset \mathbb{R}^d$ lifts to $\ell(K) = \{\ell(x) \in P : x \in K\}$. Also, $\pi(\ell(K)) = K$.

A hyperplane is called *non-vertical* if $\pi(H) = \mathbb{R}^d$. The lifting gives a bijective relation between non-vertical hyperplanes in $\mathbb{R}^{d+1}$ (intersecting P) and $(d-1)$-dimensional spheres in $\mathbb{R}^d$ in the following way. Assume $S = S(u, r)$ is the sphere centered at u, with radius r in $\mathbb{R}^d$. Of course, $\ell(S) \subset P$, but more importantly,

$$\ell(S) = P \cap H,$$

where H is the hyperplane with equation $x_{d+1} = 2\langle u, x\rangle + r^2 - |u|^2$. Conversely, given a non-vertical hyperplane H with equation $x_{d+1} = 2\langle u, x\rangle + s$ where $s = r^2 - |u|^2$ with some $r > 0$,

$$\pi(H \cap P) = S(u, r).$$

As a first application of this lifting, here is a simple proof of a slightly stronger version of Theorem KLH (we can replace the convexity assumption by connectedness).

Consider a family of $d+1$ well separated connected compact sets in $\mathbb{R}^d$ and a partition of the sets into two classes. Lift the family into the paraboloid, and for each lifted set, consider its convex hull. This gives a $(d+1)$-element family of convex bodies in $\mathbb{R}^{d+1}$. The lifted family is well separated. This can be seen using Proposition 1: the lifting of the separating $(d-1)$-dimensional planes of the original family yield (vertical) separating hyperplanes of the corresponding lifting. Thus Theorem 3 applies to the lifted family (with the obviously induced partition) and gives a hyperplane H such $H \cap P$ projects onto a sphere S in $\mathbb{R}^d$ satisfying the requirements of Theorem 1. We omit the straightforward detail.

We apply Theorem 3 to the paraboloid lifting to obtain Theorem 1, in the same way. The family $\mathcal{F} = \{K_1, \ldots, K_{d+1}\}$ lifts to the family $\ell(\mathcal{F}) = \{\ell(K_1), \ldots, \ell(K_{d+1})\}$, and we define the measures μ_i via

$$\mu_i(C) = \operatorname{Vol} \pi(C \cap \ell(K_i)),$$

where C is a Borel subset of $\mathbb{R}^{d+1}$. Clearly, μ_i is finite and $\ell(\mathcal{F})$ is well separated. Its support is $\operatorname{conv} \ell(K_i)$. It is easy to see that μ_i is a nice measure by checking that it satisfies all three conditions.

Thus Theorem 3 applies and guarantees the existence of a unique positive transversal (to $\ell(\mathcal{F})$) halfspace $H \subset \mathbb{R}^{d+1}$ with $\mu_i(H \cap \ell(K_i)) = \alpha_i \cdot \mu_i(K_i)$ for each i. This translates to the ball $B = \pi(H \cap P)$ and sphere $S = \pi(H^0 \cap P)$ (where H^0 is the bounding hyperplane of H) as follows: S is a transversal sphere of the family $\mathcal{F}$ and $\operatorname{Vol}(B \cap K_i) = \alpha_i \cdot \operatorname{Vol} K_i$. Unicity of S follows readily. □

Acknowledgements The first author was partially supported by Hungarian National Foundation Grants T 60427 and NK 62321. The second and third authors are grateful, for support and hospitality, to the Department of Mathematics at University College London where this paper was written. We also thank two anonymous referees for careful reading and useful remarks and corrections.

References

1. Bárány, I., Valtr, P.: A positive fraction Erdős-Szekeres theorem. Discrete Comput. Geom. **19**, 335–342 (1997)
2. Bisztriczky, T.: On separated families of convex bodies. Arch. Math. **54**, 193–199 (1990)
3. Cappell, S.E., Goodman, J.E., Pach, J., Pollack, R., Sharir, M., Wenger, R.: Common tangents and common transversals. Adv. Math. **106**, 198–215 (1994)
4. Goodman, J.E., Pollack, R., Wenger, R.: Bounding the number of geometric permutations induced by k-transversals. J. Comb. Theory Ser. A **75**, 187–197 (1996)
5. Klee, V., Lewis, T., Von Hohenbalken, B.: Common supports as fixed points. Geom. Dedicata **60**, 277–281 (1996)
6. Klee, V., Lewis, T., Von Hohenbalken, B.: Appollonius revisited: supporting spheres for sundered systems. Discrete Comput. Geom. **18**, 385–395 (1997)
7. Kramer, H., Németh, A.B.: Supporting spheres for families of independent convex sets. Arch. Math. **24**, 91–96 (1973)

A Centrally Symmetric Version of the Cyclic Polytope

Alexander Barvinok · Isabella Novik

Abstract We define a centrally symmetric analogue of the cyclic polytope and study its facial structure. We conjecture that our polytopes provide asymptotically the largest number of faces in all dimensions among all centrally symmetric polytopes with n vertices of a given even dimension $d = 2k$ when d is fixed and n grows. For a fixed even dimension $d = 2k$ and an integer $1 \leq j < k$ we prove that the maximum possible number of j-dimensional faces of a centrally symmetric d-dimensional polytope with n vertices is at least $(c_j(d) + o(1))\binom{n}{j+1}$ for some $c_j(d) > 0$ and at most $(1 - 2^{-d} + o(1))\binom{n}{j+1}$ as n grows. We show that $c_1(d) \geq 1 - (d-1)^{-1}$ and conjecture that the bound is best possible.

1 Introduction and Main Results

To characterize the numbers that arise as the face numbers of simplicial complexes of various types is a problem that has intrigued many researchers over the last half century and has been solved for quite a few classes of complexes, among them the class of all simplicial complexes [13, 14] as well as the class of all simplicial polytopes [4, 22]. One of the precursors of the latter result was the Upper Bound Theorem (UBT, for short) [16] that provided sharp upper bounds on the face numbers of *all*

Research of A. Barvinok partially supported by NSF grant DMS 0400617.

Research of I. Novik partially supported by Alfred P. Sloan Research Fellowship and NSF grant DMS-0500748.

A. Barvinok (✉)
Department of Mathematics, University of Michigan, Ann Arbor, MI 48109-1043, USA
e-mail: barvinok@umich.edu

I. Novik
Department of Mathematics, University of Washington, Box 354350, Seattle, WA 98195-4350, USA
e-mail: novik@math.washington.edu

d-dimensional polytopes with n vertices. While the UBT is a classic by now, the situation for centrally symmetric polytopes is wide open. For instance, the largest number of edges, $\mathrm{fmax}(d, n; 1)$, that a d-dimensional centrally symmetric polytope on n vertices can have is unknown even for $d = 4$. Furthermore, no plausible conjecture about the value of $\mathrm{fmax}(d, n; 1)$ exists.

In this paper, we establish certain bounds on $\mathrm{fmax}(d, n; 1)$ and, more generally, on $\mathrm{fmax}(d, n; j)$, the maximum number of j-dimensional faces of a centrally symmetric d-dimensional polytope with n vertices. For every even dimension d we construct a centrally symmetric polytope with n vertices, which, we conjecture, provides asymptotically the largest number of faces in every dimension as n grows and d is fixed among all d-dimensional centrally symmetric polytopes with n vertices, see the discussion after Theorem 1.4 for the precise statement of the conjecture.

Let us recall the basic definitions. A polytope will always mean a convex polytope (that is, the convex hull of finitely many points), and a d-polytope—a d-dimensional polytope. A polytope $P \subset \mathbb{R}^d$ is *centrally symmetric* (cs, for short) if for every $x \in P$, $-x$ belongs to P as well, that is, $P = -P$. The number of i-dimensional faces (i-faces, for short) of P is denoted $f_i = f_i(P)$ and is called the *ith face number* of P.

The UBT proposed by Motzkin in 1957 [17] and proved by McMullen [16] asserts that among all d-polytopes with n vertices, the cyclic polytope, $C_d(n)$, maximizes the number of i-faces for every i. Here the *cyclic polytope*, $C_d(n)$, is the convex hull of n distinct points on the *moment curve* $(t, t^2, \ldots, t^d) \in \mathbb{R}^d$ or on the *trigonometric moment curve* $(\cos t, \sin t, \cos 2t, \sin 2t, \ldots, \cos kt, \sin kt) \in \mathbb{R}^{2k}$ (assuming $d = 2k$). Both types of cyclic polytopes were investigated by Carathéodory [5] and later by Gale [10] who, in particular, showed that the two types are combinatorially equivalent (for even d) and independent of the choice of points. Cyclic polytopes were also rediscovered by Motzkin [12, 17] and many others. We refer the readers to [2] and [23] for more information on these amazing polytopes.

Here we define and study a natural centrally symmetric analogue of cyclic polytopes—bicyclic polytopes.

1.1 The Symmetric Moment Curve and Bicyclic Polytopes

Consider the curve

$$\mathrm{SM}_{2k}(t) = (\cos t, \sin t, \cos 3t, \sin 3t, \ldots, \cos(2k-1)t, \sin(2k-1)t) \quad \text{for } t \in \mathbb{R},$$

which we call the *symmetric moment curve*, $\mathrm{SM}_{2k}(t) \in \mathbb{R}^{2k}$. The difference between SM_{2k} and the trigonometric moment curve is that we employ only odd multiples of t in the former. Clearly, $\mathrm{SM}_{2k}(t + 2\pi) = \mathrm{SM}_{2k}(t)$, so SM_{2k} defines a map $\mathrm{SM}_{2k} : \mathbb{R}/2\pi\mathbb{Z} \to \mathbb{R}^{2k}$. It is convenient to identify the quotient $\mathbb{R}/2\pi\mathbb{Z}$ with the unit circle $\mathbb{S}^1 \subset \mathbb{R}^2$ via the map $t \mapsto (\cos t, \sin t)$. In particular, $\{t, t + \pi\}$ is a pair of antipodal points in $\mathbb{S}^1$. We observe that

$$\mathrm{SM}_{2k}(t + \pi) = -\mathrm{SM}_{2k}(t),$$

so the symmetric moment curve $\mathrm{SM}_{2k}(\mathbb{S}^1)$ is centrally symmetric about the origin.

Let $X \subset \mathbb{S}^1$ be a finite set. A *bicyclic* $2k$-dimensional polytope, $\mathcal{B}_{2k}(X)$, is the convex hull of the points $\mathrm{SM}_{2k}(x)$, $x \in X$:

$$\mathcal{B}_{2k}(X) = \mathrm{conv}(\mathrm{SM}_{2k}(X)).$$

We note that $\mathcal{B}_{2k}(X)$ is a centrally symmetric polytope as long as one chooses X to be a centrally symmetric subset of $\mathbb{S}^1$. In the case of $k = 2$ these polytopes were introduced and studied (among certain more general 4-dimensional polytopes) by Smilansky [20, 21], but to the best of our knowledge the higher-dimensional bicyclic polytopes have not yet been investigated. Also, in [20] Smilansky studied the convex hull of $\mathrm{SM}_4(\mathbb{S}^1)$ (among convex hulls of certain more general 4-dimensional curves) but the convex hull of higher dimensional symmetric moment curves has not been studied either.

We recall that a *face* of a convex body is the intersection of the body with a supporting hyperplane. Faces of dimension 0 are called *vertices* and faces of dimension 1 are called *edges*. Our first main result concerns the edges of the convex hull

$$\mathcal{B}_{2k} = \mathrm{conv}(\mathrm{SM}_{2k}(\mathbb{S}^1))$$

of the symmetric moment curve. Note that $\mathcal{B}_{2k}$ is centrally symmetric about the origin.

Let $\alpha \neq \beta \in \mathbb{S}^1$ be a pair of non-antipodal points. By the *arc with the endpoints* α *and* β we always mean the shorter of the two arcs defined by α and β.

Theorem 1.1 *For every positive integer k there exists a number*

$$\frac{2k-2}{2k-1}\pi \leq \psi_k < \pi$$

with the following property: if the length of the arc with the endpoints $\alpha \neq \beta \in \mathbb{S}^1$ is less than ψ_k, then the interval $[\mathrm{SM}_{2k}(\alpha), \mathrm{SM}_{2k}(\beta)]$ is an edge of $\mathcal{B}_{2k}$; and if the length of the arc with the endpoints $\alpha \neq \beta \in \mathbb{S}^1$ is greater than ψ_k, then the interval $[\mathrm{SM}_{2k}(\alpha), \mathrm{SM}_{2k}(\beta)]$ is not an edge of $\mathcal{B}_{2k}$.

It looks quite plausible that

$$\psi_k = \frac{2k-2}{2k-1}\pi$$

and, indeed, this is the case for $k = 2$, cf. Sect. 4.

One remarkable property of the convex hull of the trigonometric moment curve in $\mathbb{R}^{2k}$ is that it is *k-neighborly*, that is, the convex hull of any set of k distinct points on the curve is a $(k-1)$-dimensional face of the convex hull. The convex hull of the symmetric moment curve turns out to be *locally* k-neighborly.

Theorem 1.2 *For every positive integer k there exists a number $\phi_k > 0$ such that if $t_1, \ldots, t_k \in \mathbb{S}^1$ are distinct points that lie on an arc of length at most ϕ_k, then*

$$\mathrm{conv}(\mathrm{SM}_{2k}(t_1), \ldots, \mathrm{SM}_{2k}(t_k))$$

is a $(k-1)$-dimensional face of $\mathcal{B}_{2k}$.

From Theorems 1.1 and 1.2 on one hand and using a volume trick similar to that used in [15] on the other hand, we prove the following results on $\mathrm{fmax}(d, n; j)$—*the maximum* number of j-faces that a cs d-polytope on n vertices can have.

Theorem 1.3 *If d is a fixed even number and $n \to \infty$, then*

$$1 - \frac{1}{d-1} + o(1) \leq \frac{\mathrm{fmax}(d, n; 1)}{\binom{n}{2}} \leq 1 - \frac{1}{2^d} + o(1).$$

Theorem 1.4 *If $d = 2k$ is a fixed even number, $j \leq k-1$, and $n \to \infty$, then*

$$c_j(d) + o(1) \leq \frac{\mathrm{fmax}(d, n; j)}{\binom{n}{j+1}} \leq 1 - \frac{1}{2^d} + o(1),$$

where $c_j(d)$ is a positive constant.

Some discussion is in order. Recall that the cyclic polytope is $\lfloor d/2 \rfloor$-*neighborly*, that is, for all $j \leq \lfloor d/2 \rfloor$, every j vertices of $C_d(n)$ form the vertex set of a face. Since $C_d(n)$ is a simplicial polytope, its neighborliness implies that $f_j(C(d, n)) = \binom{n}{j+1}$ for $j < \lfloor d/2 \rfloor$. Now if P is a centrally symmetric polytope on n vertices then no two of its antipodal vertices are connected by an edge, and so $f_1(P) \leq \binom{n}{2} - \frac{n}{2}$. In fact, as was recently shown by Linial and the second author [15], this inequality is strict as long as $n > 2^d$. This leads one to wonder how big the gap between $\mathrm{fmax}(d, n; 1)$ and $\binom{n}{2}$ is. Theorems 1.3 and 1.4 (see also Propositions 2.1 and 2.2 below) provide (partial) answers to those questions. In Sect. 7.3, we discuss several available lower bounds for $c_j(d)$.

Let us fix an even dimension $d = 2k$ and let $X \subset \mathbb{S}^1$ be a set of n equally spaced points, where n is an even number. We conjecture that for every integer $j \leq k-1$

$$\limsup_{n \to +\infty} \frac{f_j(\mathcal{B}_{2k}(X))}{\binom{n}{j+1}} = \limsup_{n \to +\infty} \frac{\mathrm{fmax}(d, n; j)}{\binom{n}{j+1}}.$$

It is also worth mentioning that recently there has been a lot of interest in the problems surrounding neighborliness and face numbers of cs polytopes in connection to statistics and error-correcting codes, see [7–9, 18]. In particular, it was proved in [9] that for large n and d, if j is bigger than a certain threshold value, then the ratio between the expected number of j-faces of a random cs d-polytope with n vertices and $\binom{n}{j+1}$ is smaller than $1 - \epsilon$ for some positive constant ϵ. The upper bound part of Theorem 1.4 provides a real reason for this phenomenon: the expected number of j-faces is "small" because the jth face number of *every* cs polytope is "small".

In 1980s, in an attempt to come up with a centrally symmetric variation of cyclic polytopes, Björner [3] considered convex hulls of certain symmetric sets of points chosen on the odd-moment curve

$$\mathrm{OM}_m(t) = (t, t^3, t^5, \ldots, t^{2m-1}), \qquad \mathrm{OM}_m(t) \in \mathbb{R}^m \quad \text{for } t \in \mathbb{R}.$$

The authors are not aware of any results similar to Theorems 1.1 and 1.2 for such polytopes. The curves $\mathrm{OM}_{2k} \subset \mathbb{R}^{2k}$ and $\mathrm{SM}_{2k} \subset \mathbb{R}^{2k}$ behave differently with respect

to the affine structure on $\mathbb{R}^{2k}$: there are affine hyperplanes in $\mathbb{R}^{2k}$ intersecting OM_{2k} in as many as $4k-1$ points while any affine hyperplane in $\mathbb{R}^{2k}$ intersects SM_{2k} in at most $4k-2$ points, cf. Sect. 3.3. In contrast, the ordinary and trigonometric moment curves in even dimensions behave quite similarly to each other.

As the anonymous referee pointed out to the authors, SM_{2k} is an algebraic curve of degree $4k-2$ (it is rationalized by the substitution $s=\tan(t/2)$) and $4k-2$ is the minimum degree that an irreducible centrally symmetric algebraic curve in $\mathbb{R}^{2k}$ not lying in an affine hyperplane can have. (Note that the degree of OM_{2k} is $4k-1$.)

The structure of the paper is as follows. In Sect. 2 we prove the upper bound parts of Theorems 1.3 and 1.4. In Sect. 3 we discuss bicyclic polytopes and their relationship to non-negative trigonometric polynomials and self-inversive polynomials. Section 4 contains new short proofs of results originally due to Smilansky on the faces of 4-dimensional bicyclic polytopes. It serves as a warm-up for Sects. 5 and 6 in which we prove Theorems 1.1 and 1.2 as well as the lower bound parts of Theorems 1.3 and 1.4. We discuss 2-faces of $\mathcal{B}_6$, values of $\mathrm{fmax}(2k,n;j)$ for $j\geq k$, and lower bounds for constants $c_j(d)$ of Theorem 1.4 in Sect. 7, where we also state several open questions.

2 Upper Bounds on the Face Numbers

The goal of this section is to prove the upper bound parts of Theorems 1.3 and 1.4. The proof uses a volume trick similar to the one utilized in the proof of the Danzer–Grünbaum theorem on the number of vertices of antipodal polytopes [6] and more recently in [15, Theorem 1], where it was used to estimate maximal possible neighborliness of cs polytopes.

The upper bound part of Theorem 1.3 is an immediate consequence of the following more precise result on the number of edges of a centrally symmetric polytope.

Proposition 2.1 *Let $P\subset\mathbb{R}^d$ be a cs d-polytope on n vertices. Then*

$$f_1(P)\leq\frac{n^2}{2}(1-2^{-d}).$$

Proof Let V be the set of vertices of P. For every vertex u of P we define

$$P_u:=P+u\subset 2P$$

to be a translate of P, where "+" denotes the Minkowski addition. We claim that if the polytopes P_u and P_v have intersecting interiors then the vertices u and $-v$ are not connected by an edge. (Note that this includes the case of $u=v$, since clearly $\mathrm{int}(P_v)\cap\mathrm{int}(P_v)\neq\emptyset$ and $(v,-v)$ is not an edge of P.) Indeed, the assumption $\mathrm{int}(P_u)\cap\mathrm{int}(P_v)\neq\emptyset$ implies that there exist $x,y\in\mathrm{int}(P)$ such that $x+u=y+v$, or equivalently, that $(y-x)/2=(u-v)/2$. Since P is centrally symmetric, and $x,y\in\mathrm{int}(P)$, the point $q:=(y-x)/2$ is an interior point of P. As q is also the barycenter of the line segment connecting u and $-v$, this line segment is not an edge of P.

Now normalize the Lebesgue measure dx in $\mathbb{R}^d$ in such a way that $\mathrm{vol}(2P) = 1$ and hence

$$\mathrm{vol}(P) = \mathrm{vol}(P_u) = 2^{-d} \quad \text{for all } u \in V.$$

For a set $A \subset \mathbb{R}^d$, let $[A] : \mathbb{R}^d \to \mathbb{R}$ be the indicator of A, that is, $[A](x) = 1$ for $x \in A$ and $[A](x) = 0$. Define

$$h = \sum_{u \in V} [\mathrm{int}\, P_u].$$

Then

$$\int_{2P} h dx = n2^{-d},$$

and hence by the Hölder inequality

$$\int_{2P} h^2 dx \geq n^2 2^{-2d}.$$

On the other hand, the first paragraph of the proof implies that

$$\int_{2P} h^2(x) dx = \sum_{u,v \in V} \mathrm{vol}(P_u \cap P_v) \leq n2^{-d} + 2\left(\binom{n}{2} - f_1(P)\right)2^{-d},$$

and the statement follows. □

As a corollary, we obtain the following upper bound on the number $f_j(P)$ of j-dimensional faces for any $1 \leq j \leq (d-2)/2$. This upper bound implies the upper bound part of Theorem 1.4.

Proposition 2.2 *Let $P \subset \mathbb{R}^d$ be a cs d-polytope with n vertices, and let $j \leq (d-2)/2$. Then*

$$f_j(P) \leq \frac{n}{n-1}(1 - 2^{-d})\binom{n}{j+1}.$$

Proof We rely on Proposition 2.1 and two additional results.

The first result is an adaptation of the well-known perturbation argument, see, for example, [11, Sect. 5.2], to the centrally symmetric situation. Namely, we claim that for every cs d-polytope P there exists a simplicial cs d-polytope Q such that $f_0(P) = f_0(Q)$ and $f_j(P) \leq f_j(Q)$ for all $1 \leq j \leq d-1$. The polytope Q is obtained from P by pulling the vertices of P in a generic way but so as to preserve the symmetry. The proof is completely similar to that of [11, Sect. 5.2] and hence is omitted.

The second result states that for every $(d-1)$-dimensional simplicial complex K with n vertices, we have

$$f_j(K) \leq f_1(K)\binom{n}{j+1}\Big/\binom{n}{2} \quad \text{for } 1 \leq j \leq d-1.$$

The standard double-counting argument goes as follows: every j-dimensional simplex of K contains exactly $\binom{j+1}{2}$ edges and every edge of K is contained in at most $\binom{n-2}{j-1}$ of the j-dimensional simplices of K. Hence

$$f_j(K)/f_1(K) \leq \binom{n-2}{j-1}\Big/\binom{j+1}{2} = \binom{n}{j+1}\Big/\binom{n}{2}.$$

The statement now follows by Proposition 2.1. □

3 Faces and Polynomials

In this section, we relate the facial structure of the convex hull $\mathcal{B}_{2k}$ of the symmetric moment curve (Sect. 1.1) to properties of trigonometric and complex polynomials from particular families.

3.1 Preliminaries

A proper face of a convex body $B \subset \mathbb{R}^{2k}$ is the intersection of B with its supporting hyperplane, that is, the intersection of B with the zero-set of an affine function

$$A(x) = \alpha_0 + \alpha_1\xi_1 + \cdots + \alpha_{2k}\xi_{2k} \quad \text{for } x = (\xi_1, \ldots, \xi_{2k})$$

that satisfies $A(x) \geq 0$ for all $x \in B$.

A useful observation is that $\mathcal{B}_{2k}$ remains invariant under a one-parametric group of rotations that acts transitively on $\mathrm{SM}_{2k}(\mathbb{S}^1)$. Such a rotation is represented by a $2k \times 2k$ block-diagonal matrix with the jth block being

$$\begin{pmatrix} \cos(2j-1)\tau & \sin(2j-1)\tau \\ -\sin(2j-1)\tau & \cos(2j-1)\tau \end{pmatrix}$$

for $\tau \in \mathbb{R}$. If $\{t_1, \ldots, t_s\} \subset \mathbb{S}^1$ are distinct points such that

$$\mathrm{conv}(\mathrm{SM}_{2k}(t_1), \ldots, \mathrm{SM}_{2k}(t_s))$$

is a face of $\mathcal{B}_{2k}$ and points $t_1', \ldots, t_s' \in \mathbb{S}^1$ are obtained from $t_1, \ldots, t_s$ by the rotation $t_i' = t_i + \tau$ for $i = 1, \ldots, s$ of $\mathbb{S}^1$, then

$$\mathrm{conv}(\mathrm{SM}_{2k}(t_1'), \ldots, \mathrm{SM}_{2k}(t_s'))$$

is a face of $\mathcal{B}_{2k}$ as well.

Finally, we note that the natural projection $\mathbb{R}^{2k} \to \mathbb{R}^{2k'}$ for $k' < k$ that erases the last $2k - 2k'$ coordinates maps $\mathcal{B}_{2k}$ onto $\mathcal{B}_{2k'}$ and $\mathcal{B}_{2k}(X)$ onto $\mathcal{B}_{2k'}(X)$. Hence, if for some sets $Y \subset X \subset \mathbb{S}^1$ the set $\mathrm{conv}(\mathrm{SM}_{2k'}(Y))$ is a face of $\mathcal{B}_{2k'}(X)$, then $\mathrm{conv}(\mathrm{SM}_{2k}(Y))$ is a face of $\mathcal{B}_{2k}(X)$. We call a face of $\mathcal{B}_{2k}$ an *old face* if it is an inverse image of a face of $\mathcal{B}_{2k'}$ for some $k' < k$, and call it a *new face* otherwise.

3.2 Raked Trigonometric Polynomials

The value of an affine function $A(x)$ on the symmetric moment curve SM_{2k} is represented by a trigonometric polynomial

$$A(t) = c + \sum_{j=1}^{k} a_j \cos(2j-1)t + \sum_{j=1}^{k} b_j \sin(2j-1)t. \tag{1}$$

Note that all summands involving the even terms $\sin 2jt$ and $\cos 2jt$ except for the constant term vanish from $A(t)$. We refer to such trigonometric polynomials as *raked trigonometric polynomials* of degree at most $2k-1$. As before, it is convenient to think of $A(t)$ as defined on $\mathbb{S}^1 = \mathbb{R}/2\pi\mathbb{Z}$.

Admitting, for convenience, the whole body $\mathcal{B}_{2k}$ and the empty set as faces of $\mathcal{B}_{2k}$, we obtain the following result.

Lemma 3.1 *The faces of $\mathcal{B}_{2k}$ are defined by raked trigonometric polynomials of degree at most $2k-1$ that are non-negative on $\mathbb{S}^1$. If $A(t)$ is such a polynomial and $\{t_1, \ldots, t_s\} \subset \mathbb{S}^1$ is the set of its zeroes, then the face of $\mathcal{B}_{2k}$ defined by $A(t)$ is the convex hull of $\{\mathrm{SM}_{2k}(t_1), \ldots, \mathrm{SM}_{2k}(t_s)\}$.*

It is worth noticing that if a polynomial $A(t)$ of Lemma 3.1 has degree smaller than $2k-1$, then the convex hull of $\{\mathrm{SM}_{2k-2}(t_1), \ldots, \mathrm{SM}_{2k-2}(t_s)\}$ is a face of $\mathcal{B}_{2k-2}$. Thus all new faces of $\mathcal{B}_{2k}$ are defined by raked trigonometric polynomials of degree $2k-1$.

3.3 Raked Self-Inversive Polynomials

Let us substitute $z = e^{it}$ in (1). Using that

$$\cos(2j-1)t = \frac{z^{2j-1} + z^{1-2j}}{2} \quad \text{and} \quad \sin(2j-1)t = \frac{z^{2j-1} - z^{1-2j}}{2i}$$

we can write $A(t) = z^{-2k+1} D(z)$, where

$$D(z) = cz^{2k-1} + \sum_{j=1}^{k} \frac{a_j - ib_j}{2} z^{2j+2k-2} + \sum_{j=1}^{k} \frac{a_j + ib_j}{2} z^{2k-2j}.$$

In other words, $D(z)$ is a polynomial satisfying

$$D(z) = z^{4k-2}\overline{D(1/\overline{z})} \tag{2}$$

and such that

$$D(z) = cz^{2k-1} + \sum_{j=0}^{2k-1} d_{2j} z^{2j}, \tag{3}$$

so that all odd terms with the possible exception of the middle term vanish. We note that (2) is equivalent to $\overline{d_{2j}} = d_{4k-2-2j}$ for $j = 0, \ldots, 2k-1$.

The polynomials $D(z)$ satisfying (2) are well studied and known in the literature by the name *self-inversive polynomials*, see for instance [19, Chap. 7]). (Some sources require that a self-inversive polynomial D satisfies $D(0) \neq 0$, but we do not.) In analogy with raked trigonometric polynomials, we refer to polynomials D satisfying both (2) and (3) as *raked self-inversive polynomials*. We note that any polynomial $D(z)$ satisfying (2) and (3) gives rise to a raked trigonometric polynomial $A(t)$ of degree at most $2k-1$ such that $A(t) = z^{-2k+1}D(z)$ for $z = e^{it}$. Furthermore, the multiplicity of a root t of A is equal to the multiplicity of the root $z = e^{it}$ of D. Hence we obtain the following restatement of Lemma 3.1.

Lemma 3.2 *The faces of $\mathcal{B}_{2k}$ are defined by raked self-inversive polynomials $D(z)$ that satisfy* (2) *and* (3) *and all of whose roots of modulus one have even multiplicities. If $D(z)$ is such a polynomial and $\{e^{it_1}, \ldots, e^{it_s}\}$ is the set of its roots of modulus* 1*, then the face of $\mathcal{B}_{2k}$ defined by $D(z)$ is the convex hull of $\{\mathrm{SM}_{2k}(t_1), \ldots, \mathrm{SM}_{2k}(t_s)\}$.*

Clearly, new faces of $\mathcal{B}_{2k}$ are defined by polynomials of degree $4k-2$, see Sect. 3.1.

Let $D(z)$ be a polynomial satisfying (2), and let

$$M = \{\zeta_1, \ldots, \zeta_1, \zeta_2, \ldots, \zeta_2, \ldots, \zeta_s, \ldots, \zeta_s\}$$

be the multiset of all roots of D where each root is listed the number of times equal to its multiplicity. We note that if $\deg D = 4k-2$ then $0 \notin M$ and $|M| = 4k-2$. We need a straightforward characterization of the raked self-inversive polynomials in terms of their zero multisets M.

Lemma 3.3 *A multiset $M \subset \mathbb{C} \setminus \{0\}$ of size $|M| = 4k-2$ is the multiset of roots of a raked self-inversive polynomial of degree $4k-2$ if and only if the following conditions* (*a*) *and* (*b*) *are satisfied.*

(a) *We have*

$$\overline{M} = M^{-1},$$

that is, $\zeta \in M$ if and only if $\overline{\zeta}^{-1} \in M$ and the multiplicities of ζ and $\overline{\zeta}^{-1}$ in M are equal, and

(b)

$$\sum_{\zeta \in M} \zeta^{2j-1} = 0 \quad \textit{for } j = 1, \ldots, k-1.$$

Proof It is known and not hard to see that M is the zero-multiset of a self-inversive polynomial if and only if $M^{-1} = \overline{M}$ [19, p. 149, 228]. Indeed, if $D(0) \neq 0$ then (2) implies $\overline{M} = M^{-1}$. Conversely, suppose that M is the multiset satisfying $\overline{M} = M^{-1}$ and such that $|M| = 4k-2$. Then

$$\prod_{\zeta \in M} |\zeta| = 1,$$

and hence we can choose numbers a_ζ such that

$$\prod_{\zeta\in M}\frac{\overline{a}_\zeta}{a_\zeta}=\prod_{\zeta\in M}(-\zeta).$$

Then the polynomial

$$D(z)=\prod_{\zeta\in M}a_\zeta(z-\zeta)$$

satisfies (2).

Let

$$s_m=\sum_{\zeta\in M}\zeta^m \quad \text{and let } D(z)=\sum_{m=0}^{4k-2}d_{4k-2-m}z^m.$$

Using Newton's formulas to express elementary symmetric functions in terms of power sums, we obtain

$$md_m+\sum_{j=1}^{m}s_jd_{m-j}=0 \quad \text{for } m=1,2,\ldots,4k-2.$$

Hence we conclude that

$$d_1=d_3=\cdots=d_{2k-3}=0 \quad \text{if and only if} \quad s_1=s_3=\cdots=s_{2k-3}=0,$$

which completes the proof. □

Note that even if a raked self-inversive polynomial D has 0 as its root, it must still satisfy $\sum_{\zeta\in M}\zeta^{2j-1}=0$, for all $1\le j\le k-1$, where M is the multiset of all roots of D.

We conclude this section with the description of a particular family of faces of $\mathcal{B}_{2k}$.

3.4 Some Simplicial Faces of $\mathcal{B}_{2k}$

Let

$$A(t)=1-\cos((2k-1)t).$$

Clearly, $A(t)$ is a raked trigonometric polynomial and $A(t)\ge 0$ for all $t\in\mathbb{S}^1$. Moreover, $A(t)=0$ at the $2k-1$ points

$$\tau_j=\frac{2\pi j}{2k-1} \quad \text{for } j=1,\ldots,2k-1$$

on the circle $\mathbb{S}^1$, which form the vertex set of a regular $(2k-1)$-gon. By Lemma 3.1 the set

$$\Delta_0=\operatorname{conv}(\mathrm{SM}_{2k}(\tau_1),\ldots,\mathrm{SM}_{2k}(\tau_{2k-1}))$$

is a face of $\mathcal{B}_{2k}$.

One can observe that Δ_0 is a $(2k-2)$-dimensional simplex. To prove that, we have to show that the points $\mathrm{SM}_{2k}(\tau_1), \ldots, \mathrm{SM}_{2k}(\tau_{2k-1})$ are affinely independent, or, equivalently, that the affine hyperplanes in $\mathbb{R}^{2k}$ passing through these points form a one-parametric family (topologically, this set of hyperplanes is a circle). As in Sect. 3.3, those hyperplanes correspond to non-zero complex polynomials $D(z)$ that satisfy (2) and (3) and for which $D(e^{i\tau_j}) = 0$ for $j = 1, \ldots, 2k-1$. Hence for each such D we must have

$$D(z) = (z^{2k-1} - 1)D_1(z),$$

for some polynomial $D_1(z)$ with $\deg D_1 = s \leq 2k-1$. Moreover, it follows from (2) that

$$D_1(z) = -z^{2k-1}\overline{D_1(1/\overline{z})}. \tag{4}$$

Let M be the multiset of all roots of D and let M_1 be the multiset of all roots of D_1. Applying Lemma 3.3, we deduce that

$$\sum_{\zeta \in M_1} \zeta^{2j-1} = \sum_{\zeta \in M} \zeta^{2j-1} = 0 \quad \text{for } j = 1, \ldots, k-1.$$

Now, as in the proof of Lemma 3.3 we conclude that the odd-power coefficients of z in D_1 are zeros except, possibly, for that of z^{2k-1}. In view of (4), all other coefficients of D_1 except, possibly, the constant term, must be zeros as well. Summarizing, $D_1(z) = \alpha z^{2k-1} - \overline{\alpha}$ for some $\alpha \in \mathbb{C} \setminus \{0\}$. Normalizing $|D(0)| = 1$ we get a one-parametric family of polynomials $D(z) = (z^{2k-1} - 1)(\alpha z^{2k-1} - \overline{\alpha})$, $|\alpha| = 1$, that corresponds to the set of affine hyperplanes in $\mathbb{R}^{2k}$ passing through the vertices of Δ_0.

Therefore Δ_0 is indeed a $(2k-2)$-dimensional simplex. Furthermore, we have a one-parametric family of simplicial faces

$$\Delta_\tau = \mathrm{conv}(\mathrm{SM}_{2k}(\tau_1 + \tau), \ldots, \mathrm{SM}_{2k}(\tau_{2k-1} + \tau)) \quad \text{for } 0 \leq \tau < 2\pi$$

of $\mathcal{B}_{2k}$. The dimension of the boundary of $\mathcal{B}_{2k}$ is $(2k-1)$, so this one-parametric family of simplices covers a "chunk" of the boundary of $\mathcal{B}_{2k}$. Borrowing a term from the polytope theory, we may call Δ_τ a "ridge" of $\mathcal{B}_{2k}$. As follows from Sect. 4, for $k = 2$ the simplices Δ_τ are the only ridges of $\mathcal{B}_{2k}$.

4 The Faces of $\mathcal{B}_4$

In this section we provide a complete characterization of the faces of $\mathcal{B}_4$. This result is not new, it was proved by Smilansky [20] who also described the facial structure of the convex hull of the more general curve $(\cos pt, \sin pt, \cos qt, \sin qt)$, where p and q are any positive integers. Our proof serves as a warm-up for the following section where we discuss the edges of $\mathcal{B}_{2k}$ for $k > 2$.

Theorem 4.1 [20] *The proper faces of $\mathcal{B}_4$ are*

(0) *the* 0*-dimensional faces* (*vertices*)

$$\mathrm{SM}_4(t), \quad t \in \mathbb{S}^1;$$

(1) *the* 1*-dimensional faces* (*edges*)

$$[\mathrm{SM}_4(t_1), \mathrm{SM}_4(t_2)],$$

where $t_1 \neq t_2$ are the endpoints of an arc of $\mathbb{S}^1$ of length less than $2\pi/3$; and

(2) *the* 2*-dimensional faces* (*equilateral triangles*)

$$\Delta_t = \mathrm{conv}(\mathrm{SM}_4(t), \mathrm{SM}_4(t + 2\pi/3), \mathrm{SM}_4(t + 4\pi/3)), \quad t \in \mathbb{S}^1.$$

Proof We use Lemma 3.2. A face of $\mathcal{B}_4$ is determined by a raked self-inversive polynomial D of degree at most 6. Such a polynomial D has at most 3 roots on the circle $\mathbb{S}^1$, each having an even multiplicity. Furthermore, by Lemma 3.3, the sum of all the roots of D is 0. Therefore, we have the following three cases.

Polynomial D has 3 double roots $\zeta_1 = e^{it_1}, \zeta_2 = e^{it_2}, \zeta_3 = e^{it_3}$. Since $\zeta_1 + \zeta_2 + \zeta_3 = 0$, the points $t_1, t_2, t_3 \in \mathbb{S}^1$ form the vertex set of an equilateral triangle, and we obtain the 2-dimensional face defined in Part (2).

Polynomial D has two double roots $\zeta_1 = e^{it_1}$, $\zeta_2 = e^{it_2}$, and a pair of simple roots ζ and $\overline{\zeta}^{-1}$ with $|\zeta| \neq 1$. Applying a rotation, if necessary, we may assume without loss of generality that $t_1 = -t_2 = t$. Since we must have

$$\zeta + \overline{\zeta}^{-1} + 2e^{it} + 2e^{-it} = 0,$$

we conclude that $\zeta \in \mathbb{R}$. Hence the equation reads

$$\zeta + \zeta^{-1} = -4\cos t \quad \text{for some } \zeta \in \mathbb{R},\ |\zeta| \neq 1.$$

If $|\cos t| > 1/2$ then the solutions ζ, ζ^{-1} of this equation are indeed real and satisfy $|\zeta|, |\zeta^{-1}| \neq 1$. If $|\cos t| \leq 1/2$ then $\{\zeta, \zeta^{-1}\}$ is a pair of complex conjugate numbers satisfying $|\zeta| = |\zeta^{-1}| = 1$. Therefore, the interval $[\mathrm{SM}_4(-t), \mathrm{SM}_4(t)]$ is a face of $\mathcal{B}_4$ if and only if $-\pi/3 < t < \pi/3$ or $2\pi/3 < t < 4\pi/3$, so we obtain the 1-dimensional faces as in Part (1).

Finally, Lemma 3.1 applied to $A(t) = 1 - \cos(\tau - t)$ yields that $\mathrm{SM}_4(\tau)$ is a 0-dimensional face (vertex) of $\mathcal{B}_4$ for every $\tau \in \mathbb{S}^1$, which concludes the proof. □

5 Edges of $\mathcal{B}_{2k}$

In this section we prove Theorems 1.1 and 1.3. Our main tool is a certain deformation of simplicial faces of $\mathcal{B}_{2k}$, cf. Sect. 3.4.

5.1 Deformation

Let M be a finite multiset of non-zero complex numbers such that $M = M^{-1}$. In other words, for every $\zeta \in M$ we have $\zeta^{-1} \in M$ and the multiplicities of ζ and ζ^{-1} in M are equal. In addition, we assume that the multiplicities of 1 and -1 in M are even, possibly 0. For every $\lambda \in \mathbb{R} \setminus \{0\}$ we define a multiset M_λ, which we call a *deformation* of M, as follows.

We think of M as a multiset of unordered pairs $\{\zeta, \zeta^{-1}\}$. For each such pair, we consider the equation

$$z + z^{-1} = \lambda(\zeta + \zeta^{-1}). \tag{5}$$

We let M_λ to be the multiset consisting of the pairs $\{z, z^{-1}\}$ of solutions of (5) as $\{\zeta, \zeta^{-1}\}$ range over M. Clearly, $|M_\lambda| = |M|$ and $M_\lambda^{-1} = M_\lambda$. In addition, if $\overline{M} = M$ then $\overline{M_\lambda} = M_\lambda$, since λ in (5) is real.

Our interest in the deformation $M \mapsto M_\lambda$ is explained by the following lemma.

Lemma 5.1 *Let $D(z)$ be a raked self-inversive polynomial of degree $4k-2$ with real coefficients and such that $D(0) \neq 0$. Let M be the multiset of all roots of D and suppose that both 1 and -1 have an even, possibly 0, multiplicity in M. Then, for every real $\lambda \neq 0$, the deformation M_λ of M is the multiset of all roots of a raked self-inversive polynomial $D_\lambda(z)$ of degree $4k-2$ with real coefficients.*

Proof We use Lemma 3.3. Since D has real coefficients, we have $M = \overline{M}$, so by Lemma 3.3, we have $M = M^{-1}$ as well. Then $M_\lambda = M_\lambda^{-1}$ and $M_\lambda = \overline{M_\lambda}$, so M_λ is the multiset of the roots of a self-inversive real polynomial D_λ of degree $4k-2$. It remains to check that

$$\sum_{\zeta \in M_\lambda} \zeta^{2j-1} = 0 \quad \text{for } j = 1, \ldots, k-1.$$

We have

$$(x + x^{-1})^{2n-1} = \sum_{m=1}^{n} \binom{2n-1}{n+m-1} (x^{2m-1} + x^{-2m+1}). \tag{6}$$

Since by Lemma 3.3

$$\sum_{\zeta \in M} \zeta^{2j-1} = \sum_{\zeta \in M} \zeta^{1-2j} = 0 \quad \text{for } j = 1, \ldots, k-1,$$

it follows by (6) that

$$\sum_{\zeta \in M} (\zeta + \zeta^{-1})^{2j-1} = 0 \quad \text{for } j = 1, \ldots, k-1.$$

Therefore, by (5), we have

$$\sum_{\zeta \in M_\lambda} (\zeta + \zeta^{-1})^{2j-1} = 0 \quad \text{for } j = 1, \ldots, k-1,$$

from which by (6) we obtain

$$\sum_{\zeta \in M_\lambda} \zeta^{2j-1} = \frac{1}{2} \sum_{\zeta \in M_\lambda} (\zeta^{2j-1} + \zeta^{-2j+1}) = 0 \quad \text{for } j = 1, \ldots, k-1,$$

as claimed. Hence $D_\lambda(z)$ is a raked self-inversive polynomial. □

To prove Theorem 1.1 we need another auxiliary result.

Lemma 5.2 *Let $\alpha, \beta \in \mathbb{S}^1$ be such that the interval $[\mathrm{SM}_{2k}(\alpha), \mathrm{SM}_{2k}(\beta)]$ is an edge of $\mathcal{B}_{2k}$ and let $\alpha' \neq \beta' \in \mathbb{S}^1$ be some other points such that the arc with the endpoints α', β' is shorter than the arc with the endpoints $\alpha, \beta \in \mathbb{S}^1$. Then the interval $[\mathrm{SM}_{2k}(\alpha'), \mathrm{SM}_{2k}(\beta')]$ is an edge of $\mathcal{B}_{2k}$.*

Proof Because of rotational invariance, we assume, without loss of generality, that $\alpha = \tau$ and $\beta = -\tau$ for some $0 < \tau < \pi/2$. Let $A(t)$ be a raked trigonometric polynomial that defines the edge $[\mathrm{SM}_{2k}(\alpha), \mathrm{SM}_{2k}(\beta)]$, see Lemma 3.1. Hence $A(t) \geq 0$ for all $t \in \mathbb{S}^1$ and $A(t) = 0$ if and only $t = \pm\tau$. Let $A_1(t) = A(t) + A(-t)$. Then $A_1(t)$ is a raked trigonometric polynomial such that $A_1(t) \geq 0$ for all $t \in \mathbb{S}^1$ and $A_1(t) = 0$ if and only if $t = \pm\tau$. Furthermore, we can write

$$A_1(t) = c + \sum_{j=1}^{k} a_j \cos(2j-1)t$$

for some real a_j and c. Moreover, we assume, without loss of generality, that $a_k \neq 0$. (Otherwise choose k' to be the largest index j with $a_j \neq 0$ and project $\mathcal{B}_{2k}$ onto $\mathcal{B}_{2k'}$, cf. Sect. 3.1.) Hence the polynomial $D(z)$ defined by $A_1(t) = z^{-2k+1} D(z)$ for $z = e^{it}$, see Sect. 3.3, is a raked self-inversive polynomial of degree $4k-2$ with real coefficients satisfying $D(0) \neq 0$. Moreover, the only roots of $D(z)$ that lie on the circle $|z| = 1$ are $e^{i\tau}$ and $e^{-i\tau}$ and those roots have equal even multiplicities.

Choose an arbitrary $0 < \tau' < \tau$ and let

$$\lambda = \frac{\cos\tau'}{\cos\tau} > 1.$$

Let D_λ be the raked self-inversive polynomial of degree $4k-2$ whose existence is established by Lemma 5.1. Since

$$e^{i\tau'} + e^{-i\tau'} = \lambda(e^{i\tau} + e^{-i\tau}),$$

the numbers $e^{i\tau'}$ and $e^{-i\tau'}$ are roots of D_λ of even multiplicity. Moreover, suppose that z is a root of D_λ such that $|z| = 1$. Then $z + z^{-1} \in \mathbb{R}$ and $-2 \leq z + z^{-1} \leq 2$. By (5) and from the fact that $\lambda > 1$, it follows that there is a pair ζ, ζ^{-1} of roots of D such that $\zeta + \zeta^{-1} \in \mathbb{R}$ and $-2 < |\zeta + \zeta^{-1}| < 2$. It follows then that $|\zeta| = |\zeta^{-1}| = 1$, which necessarily yields that $\{\zeta, \zeta^{-1}\} = \{e^{i\tau}, e^{-i\tau}\}$, and hence that $\{z, z^{-1}\} = \{e^{i\tau'}, e^{-i\tau'}\}$. Therefore, by Lemma 3.2, $[\mathrm{SM}_{2k}(-\tau'), \mathrm{SM}_{2k}(\tau')]$ is an edge of $\mathcal{B}_{2k}$. Using rotational invariance, we infer that $[\mathrm{SM}_{2k}(\alpha'), \mathrm{SM}_{2k}(\beta')]$ is an edge of $\mathcal{B}_{2k}$, where points α', β' are obtained from $\tau', -\tau'$ by an appropriate rotation of $\mathbb{S}^1$. □

We are now ready to complete the proof of Theorem 1.1.

Proof of Theorem 1.1 In view of Lemma 5.2, it remains to show that one can find an arbitrarily small $\delta > 0$ and two points $\alpha, \beta \in \mathbb{S}^1$ such that the interval $[\mathrm{SM}_{2k}(\alpha), \mathrm{SM}_{2k}(\beta)]$ is an edge of $\mathcal{B}_{2k}$ and the length of the arc with the endpoints α and β is at least $\frac{2\pi(k-1)}{2k-1} - \delta$.

Consider the polynomial

$$D(z) = (z^{2k-1} - 1)^2 = z^{4k-2} - 2z^{2k-1} + 1.$$

Clearly, $D(z)$ is a raked self-inversive polynomial of degree $4k-2$ and the multiset M of the roots of D consists of all roots of unity of degree $2k-1$, each with multiplicity 2. In fact, $D(z)$ defines a simplicial face of $\mathcal{B}_{2k}$, cf. Sect. 3.4. Note that since $\mathcal{B}_{2k}$ is not polyhedral, a face of a face of $\mathcal{B}_{2k}$ does not have to be a face of $\mathcal{B}_{2k}$.

For $\epsilon > 0$ we consider the deformation $D_{1+\epsilon}(z)$ of $D(z)$ and its roots, see Lemma 5.1. In view of equation (5), for all sufficiently small $\epsilon > 0$, the multiset $M_{1+\epsilon}$ of the roots of $D_{1+\epsilon}$ consists of two positive simple real roots defined by the equation

$$z + z^{-1} = 2(1+\epsilon)$$

and $2k-2$ double roots on the unit circle defined by the equation

$$z + z^{-1} = 2(1+\epsilon)\cos\frac{2\pi j}{2k-1} \quad \text{for } j = 1, \ldots, 2k-2.$$

The first two of these roots are deformations of the double root at 1 (one of them is strictly larger and another one is strictly smaller than number 1), while the other roots are deformations of the remaining $2k-2$ roots of unity.

For $\epsilon > 0$ small enough let $\zeta_j(\epsilon)$ denote the deformation of the root

$$\zeta_j = \cos\frac{2\pi j}{2k-1} + i\sin\frac{2\pi j}{2k-1} \quad \text{for } j = 1, \ldots, 2k-2$$

that lies close to ζ_j, see Fig. 1. Thus we have

$$\zeta_j^{-1}(\epsilon) = \overline{\zeta_j}(\epsilon) = \zeta_{2k-1-j}(\epsilon). \tag{7}$$

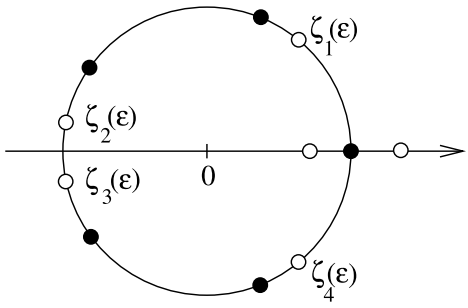


Fig. 1 The roots of unity (black dots) and their deformations (white dots) for $k = 3$

Write

$$\zeta_j(\epsilon) = e^{i\alpha_j} \quad \text{where } 0 < \alpha_j < 2\pi \text{ for } j = 1, \ldots, 2k-2.$$

Then

$$\cos\alpha_j = (1+\epsilon)\cos\frac{2\pi j}{2k-1},$$

and hence

$$\alpha_j = \frac{2\pi j}{2k-1} - \epsilon \operatorname{ctg}\frac{2\pi j}{2k-1} + O(\epsilon^2). \tag{8}$$

We now prove that the interval

$$[\mathrm{SM}_{2k}(\alpha_1), \mathrm{SM}_{2k}(\alpha_k)]$$

is an edge of $\mathcal{B}_{2k}$. We obtain this edge as the intersection of two faces of $\mathcal{B}_{2k}$. The first face is

$$\operatorname{conv}(\mathrm{SM}_{2k}(\alpha_1), \ldots, \mathrm{SM}_{2k}(\alpha_{2k-2})), \tag{9}$$

which by Lemmas 3.2 and 5.1 is indeed a face of $\mathcal{B}_{2k}$. The second face is obtained by a rotation of (9). Namely, consider the clockwise rotation of the circle $|z| = 1$ that maps $\zeta_{k-1}(\epsilon)$ onto $\zeta_1(\epsilon)$. Because of (7) this rotation also maps $\zeta_{2k-2}(\epsilon)$ onto $\zeta_k(\epsilon)$. Furthermore, for $j = 1, \ldots, 2k-2$ define

$$\zeta'_j(\epsilon) = e^{i\alpha'_j}, \quad \text{where } 0 < \alpha'_j < 2\pi,$$

as the image under this rotation of $\zeta_{j+k-2}(\epsilon)$ if $j \le k$, of $\zeta_{j-k-1}(\epsilon)$ if $j > k+1$, and of $\zeta_{k-2}(\epsilon)$ if $j = k+1$. By rotational invariance, see Sect. 3.1,

$$\operatorname{conv}(\mathrm{SM}_{2k}(\alpha'_1), \ldots, \mathrm{SM}_{2k}(\alpha'_{2k-2})) \tag{10}$$

is a face of $\mathcal{B}_{2k}$ as well.

Using (8), we conclude that

$$\alpha'_j - \alpha_j = \epsilon\left(\operatorname{ctg}\frac{2\pi j}{2k-1} - \operatorname{ctg}\frac{2\pi j - 3\pi}{2k-1} - \operatorname{ctg}\frac{\pi}{2k-1} - \operatorname{ctg}\frac{2\pi}{2k-1}\right) + O(\epsilon^2)$$
$$\text{for } 1 \le j \le 2k-2,\ j \ne k+1$$

and

$$\alpha'_{k+1} = o(1) \quad \text{as } \epsilon \to 0+.$$

Therefore for a sufficiently small $\epsilon > 0$ and $j \ne k+1$, the value of α'_j is close to and strictly smaller than α_j unless $j = 1$ or $j = k$, in which case the two values are equal. Furthermore, $\alpha'_{k+1} \ne \alpha_j$ for all j. Thus faces (9) and (10) intersect along the interval

$$[\mathrm{SM}_{2k}(\alpha_1), \mathrm{SM}_{2k}(\alpha_k)],$$

and this interval is an edge of $\mathcal{B}_{2k}$. Since α_1 and α_k are the endpoints of an arc of length

$$\pi\frac{2k-2}{2k-1} - O(\epsilon),$$

the statement follows. □

Proof of Theorem 1.3 The upper bound follows by Proposition 2.1. To prove the lower bound, consider the family of polytopes $\mathcal{B}_{2k}(X_n)$, where $X_n \subset \mathbb{S}^1$ is the set of n equally spaced points (n is even). The lower bound then follows by Theorem 1.1. □

6 Faces of $\mathcal{B}_{2k}$

In this section, we prove Theorems 1.2 and 1.4. Theorem 1.2 is deduced from the following proposition.

Proposition 6.1 *For every positive integer k there exists a number $\phi_k > 0$ such that every set of $2k$ distinct points $t_1, \ldots, t_{2k} \in \mathbb{S}^1$ lying on an arc of length at most ϕ_k is the set of the roots of some raked trigonometric polynomial $A : \mathbb{S}^1 \to \mathbb{R}$,*

$$A(t) = c_0 + \sum_{j=1}^{k} a_j \sin(2j-1)t + \sum_{j=1}^{k} b_j \cos(2j-1)t.$$

To prove Proposition 6.1, we establish first that the curve $\mathrm{SM}_{2k}(t)$ is nowhere locally flat.

Lemma 6.2 *Let*

$$\mathrm{SM}_{2k}(t) = (\cos t, \sin t, \cos 3t, \sin 3t, \ldots, \cos(2k-1)t, \sin(2k-1)t)$$

be the symmetric moment curve. Then, for every $t \in \mathbb{R}^1$, the vectors

$$\mathrm{SM}_{2k}(t), \qquad \frac{d}{dt}\mathrm{SM}_{2k}(t), \qquad \frac{d^2}{dt^2}\mathrm{SM}_{2k}(t), \qquad \ldots, \qquad \frac{d^{2k-1}}{dt^{2k-1}}\mathrm{SM}_{2k}(t)$$

are linearly independent.

Proof Because of rotational invariance, it suffices to prove the result for $t = 0$. Consider the $2k$ vectors

$$\mathrm{SM}_{2k}(0), \qquad \frac{d}{dt}\mathrm{SM}_{2k}(0), \qquad \ldots, \qquad \frac{d^{2k-1}}{dt^{2k-1}}\mathrm{SM}_{2k}(0),$$

that is, the vectors

$$a_j = (-1)^j(1, 0, 3^{2j}, 0, \ldots, 0, (2k-1)^{2j}) \quad \text{and}$$
$$b_j = (-1)^j(0, 1, 0, 3^{2j+1}, \ldots, (2k-1)^{2j+1}, 0)$$

for $j = 0, \ldots, k-1$. It is easy to see that the set of vectors $\{a_j, b_j : j = 0, \ldots, k-1\}$ is linearly independent if and only if both sets of vectors $\{a_j, j = 0, \ldots, k-1\}$ and $\{b_j : j = 0, \ldots, k-1\}$ are linearly independent. And indeed, the odd-numbered coordinates of $(-1)^j a_j$ form the $k \times k$ Vandermonde matrix

$$\begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & 3^2 & 5^2 & \ldots & (2k-1)^2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 3^{2k-2} & 5^{2k-2} & \ldots & (2k-1)^{2k-2} \end{pmatrix}$$

while the even-numbered coordinates of $(-1)^j b_j$ form the $k \times k$ Vandermonde matrix

$$\begin{pmatrix} 1 & 3 & 5 & \ldots & (2k-1) \\ 1 & 3^3 & 5^3 & \ldots & (2k-1)^3 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 3^{2k-1} & 5^{2k-1} & \ldots & (2k-1)^{2k-1} \end{pmatrix}.$$

Hence the statement follows. □

Next, we establish a curious property of zeros of raked trigonometric polynomials.

Lemma 6.3 *Let*

$$A(t) = c_0 + \sum_{j=1}^{k} a_j \sin(2j-1)t + \sum_{j=1}^{k} b_j \cos(2j-1)t$$

be a raked trigonometric polynomial $A : \mathbb{S}^1 \to \mathbb{R}$ that is not identically 0. Suppose that A has $2k$ distinct roots in an arc $\Omega \subset \mathbb{S}^1$ of length less than π. Then, if A has yet another root on $\mathbb{S}^1$, that root must lie in the arc $\Omega + \pi$.

Proof Consider the derivative of $A(t)$,

$$A'(t) = \sum_{j=1}^{k} a_j (2j-1) \cos(2j-1)t - \sum_{j=1}^{k} b_j (2j-1) \sin(2j-1)t,$$

as a map from $\mathbb{S}^1$ to $\mathbb{R}$. Substituting $z = e^{it}$, we can write

$$A'(t) = \frac{1}{z^{2k-1}} P(z),$$

where $P(z)$ is a polynomial of degree at most $4k-2$, cf. Sect. 3.3. Hence the total number of the roots of A' in $\mathbb{S}^1$, counting multiplicities, does not exceed $4k-2$.

Let $t_0, t_1 \in \Omega$ be the roots of A closest to the endpoints of Ω. By Rolle's theorem, A' has at least $2k-1$ distinct roots between t_0 and t_1 in Ω. Since $A'(t+\pi) = -A'(t)$, we must have another $2k-1$ distinct roots of A' in the arc $\Omega + \pi$ between $t_0 + \pi$ and $t_1 + \pi$, see Fig. 2.

Suppose that A has a root $z \in \mathbb{S}^1$ outside of $\Omega \cup (\Omega + \pi)$. Then either z lies in the open arc with the endpoints t_0 and $t_1 + \pi$ or z lies in the open arc with the endpoints

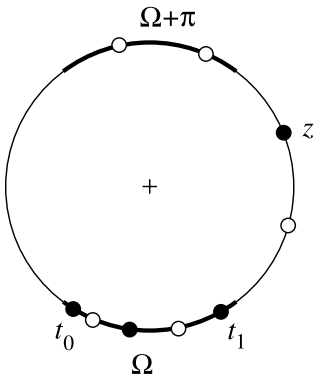


Fig. 2 The roots of A (black dots) and roots of A' (white dots)

t_1 and $t_0 + \pi$. By Rolle's theorem, A' has yet another root in $\mathbb{S}^1$ between t_0 on z in the first case, and between z and t_1 in the second case, which is a contradiction. □

We are now ready to prove Proposition 6.1.

Proof of Proposition 6.1 First, we observe that for any $2k$ points $t_1, \ldots, t_{2k} \in \mathbb{S}^1$ there is an affine hyperplane passing through the points $\mathrm{SM}_{2k}(t_1), \ldots, \mathrm{SM}_{2k}(t_{2k})$ in $\mathbb{R}^{2k}$ and hence there is a non-zero raked trigonometric polynomial A such that $A(t_1) = \cdots = A(t_{2k}) = 0$. Moreover, if $t_1, \ldots, t_{2k}$ are distinct points that lie in an arc Ω of length less than π then the hyperplane is unique. Indeed, if the hyperplane is not unique then the points $\mathrm{SM}_{2k}(t_1), \ldots, \mathrm{SM}_{2k}(t_{2k})$ lie in an affine subspace of codimension at least 2. Therefore, for any point $t_{2k+1} \in \Omega \setminus \{t_1, \ldots, t_{2k}\}$ there is an affine hyperplane passing through $\mathrm{SM}_{2k}(t_1), \ldots, \mathrm{SM}_{2k}(t_{2k+1})$ and hence there is a raked trigonometric polynomial that has $2k + 1$ roots in Ω and is not identically 0, contradicting Lemma 6.3.

Suppose now that no matter how small $\phi_k > 0$ is, there is always an arc $\Omega \subset \mathbb{S}^1$ of length at most ϕ_k and a non-zero raked trigonometric polynomial A of degree $2k - 1$ that has $2k$ distinct roots in Ω and at least one more root elsewhere in $\mathbb{S}^1$. By Lemma 6.3, that remaining root must lie in the arc $\Omega + \pi$. In other words, for any positive integer n there exists an arc $\Omega_n \subset \mathbb{S}^1$ of length at most $1/n$ and an affine hyperplane H_n which intersects $\mathrm{SM}_{2k}(\Omega_n)$ in $2k$ distinct points and also intersects the set $\mathrm{SM}_{2k}(\Omega_n + \pi)$. The set of all affine hyperplanes intersecting the compact set $\mathrm{SM}_{2k}(\mathbb{S}^1)$ is compact in the natural topology; for example if we view the set of affine hyperplanes in $\mathbb{R}^{2k}$ as a subset of the Grassmannian of all (linear) hyperplanes in $\mathbb{R}^{2k+1}$. Therefore, the sequence of hyperplanes H_n has a limit hyperplane H. By Lemma 6.2, the affine hyperplane H is the $(2k - 1)$th order tangent hyperplane to $\mathrm{SM}_{2k}(\mathbb{S}^1)$ at some point $\mathrm{SM}_{2k}(t_0)$ where t_0 is a limit point of the arcs Ω_n. Also, H passes through the point $-\mathrm{SM}_{2k}(t_0)$. The corresponding trigonometric polynomial $A(t)$ is a raked polynomial of degree at most $2k - 1$ that is not identically 0 and has two roots t_0 and $t_0 + \pi$ with the multiplicity of t_0 being at least $2k$.

Let

$$A(t) = c_0 + \sum_{j=1}^{k} a_j \sin(2j-1)t + \sum_{j=1}^{k} b_j \cos(2j-1)t.$$

Since $A(t_0) = A(t_0 + \pi) = 0$ we conclude that $c_0 = 0$. This, however, contradicts Lemma 6.2 since the non-zero $2k$-vector

$$(b_1, a_1, b_2, a_2, \ldots, b_k, a_k)$$

turns out to be orthogonal to vectors

$$\mathrm{SM}_{2k}(t_0), \qquad \frac{d}{dt}\mathrm{SM}_{2k}(t_0), \qquad \ldots, \qquad \frac{d^{2k-1}}{dt^{2k-1}}\mathrm{SM}_{2k}(t_0).$$

□

Proof of Theorem 1.2 Let $\phi_k > 0$ be the number whose existence is established in Proposition 6.1. Given k distinct points $t_1, \ldots, t_k$ lying on an arc of length at most ϕ_k, we must present a raked trigonometric polynomial A that has roots of multiplicity two at $t_1, \ldots, t_k$ and no other roots on the circle. In geometric terms, we must present an affine hyperplane that is the first order tangent to the points $\mathrm{SM}_{2k}(t_1), \ldots, \mathrm{SM}_{2k}(t_k)$ and does not intersect $\mathrm{SM}_{2k}(\mathbb{S}^1)$ anywhere else. As in the proof of Proposition 6.1, such a hyperplane is obtained as a limit of the affine hyperplanes that for every $j = 1, \ldots, k$ intersect $\mathrm{SM}_{2k}(\mathbb{S}^1)$ at two distinct points converging to t_j. □

Proof of Theorem 1.4 The upper bound follows by Proposition 2.2. To prove the lower bound, consider the family of polytopes $\mathcal{B}_{2k}(X_n)$, where $X_n \subset \mathbb{S}^1$ is the set of n equally spaced points (n is even). The lower bound then follows by Theorem 1.2. In fact, one can show that $c_j(2k) \geq 2^{-j}$: to obtain this inequality consider the polytope $\mathcal{B}_{2k}(Z)$ where $Z = Y \cup (Y + \pi)$ and Y lies in an arc of length at most ϕ_k as defined in Theorem 1.2. □

It was observed by G. Ziegler [24] that the bound $c_j(d) \geq 2^{-j}$ can also be obtained by considering the family of cs polytopes $P_n = \mathrm{conv}(Q_n \cup (-Q_n))$, where Q_n is the d-dimensional cyclic polytope whose vertices are $(i, i^2, \ldots, i^d)$ for $1 \leq i \leq n$.

7 Concluding Remarks

We close the paper with three additional remarks on the face numbers of centrally symmetric polytopes and several open questions.

7.1 The Upper Half of the Face Vector

Theorems 1.3 and 1.4 provide estimates on $\mathrm{fmax}(2k, n; j)$—the maximal possible number of j-faces that a cs $2k$-polytope on n vertices can have—for $j \leq k-1$. What can be said about $\mathrm{fmax}(2k, n; j)$ for $j \geq k$? Here we prove that for every $k \leq j < 2k$, the value of $\mathrm{fmax}(2k, n; j)$ is of the order of n^k.

Theorem 7.1 *For every positive even integer $d = 2k$ and an integer $k \leq j < 2k$, there exist positive constants $\gamma_j(d)$ and $\Gamma_j(d)$ such that*

$$\gamma_j(d) + o(1) \leq \frac{\mathrm{fmax}(d, n; j)}{\binom{n}{k}} \leq \Gamma_j(d) + o(1) \quad \text{as } n \to +\infty.$$

Proof The upper bound estimate follows from the Upper Bound Theorem [16] which holds for all polytopes. To verify the lower bound, consider a cs $2k$-polytope P_n on n vertices that satisfies $f_{k-1}(P_n) = \mathrm{fmax}(2k, n; k-1)$. As in the proof of Proposition 2.2 we can assume that P_n is a simplicial polytope. Let

$$h(P_n) = (h_0(P_n), h_1(P_n), \ldots, h_{2k}(P_n))$$

be the h-vector of P_n (see for instance [23, Chap. 8]), that is, the vector whose entries are defined by the polynomial identity

$$\sum_{i=0}^{d} h_i(P_n) x^{2k-i} = \sum_{i=0}^{d} f_{i-1}(P_n)(x-1)^{2k-i}.$$

Equivalently,

$$f_{j-1}(P_n) = \sum_{i=0}^{j} \binom{2k-i}{2k-j} h_i(P_n), \quad j = 0, 1, \ldots, 2k. \tag{11}$$

The h-numbers of a simplicial polytope are well-known to be nonnegative and symmetric [23, Chap. 8], that is, $h_j(P_n) = h_{2k-j}(P_n)$ for $j = 0, 1, \ldots, 2k$. Moreover, McMullen's proof of the UBT implies that the h-numbers of any simplicial $2k$-polytope with n vertices satisfy

$$h_j \leq \binom{n-2k+j-1}{j} = O(n^j), \quad \text{for } 0 \leq j \leq k.$$

Substituting these inequalities into (11) for $j = k-1$ and using that

$$f_{k-1}(P_n) = \mathrm{fmax}(2k, n; k-1) = \Omega(n^k)$$

by Theorem 1.4, we obtain

$$h_k(P_n) = \Omega(n^k).$$

Together with nonnegativity of h-numbers and (11), this implies that

$$\mathrm{fmax}(2k, n; j) \geq f_j(P_n) = \Omega(n^k) \quad \text{for all } k \leq j < 2k,$$

as required. □

7.2 2-Faces of $\mathcal{B}_6$

We provide some additional estimates on the extent to which $\mathcal{B}_6$ is 3-neighborly.

Theorem 7.2 *Let $t_1, t_2, t_3 \in \mathbb{R}$ be such that the points $z_1 = e^{it_1}$, $z_2 = e^{it_2}$, and $z_3 = e^{it_3}$ are distinct and lie on an arc of the unit circle of length at most* $\arccos(1/8)$. *Then the convex hull of the set $\{\mathrm{SM}_6(t_1), \mathrm{SM}_6(t_2), \mathrm{SM}_6(t_3)\}$ is a 2-dimensional face of $\mathcal{B}_6$.*

Proof As in Proposition 6.1 and Theorem 1.2, the proof reduces to verifying the following statement:

Let $z_1, \dots, z_6 \in \mathbb{C}$ be distinct points that lie on an arc of the unit circle $|z| = 1$ of length at most $\arccos(1/8)$. Let $D(z)$ be a raked self-inversive polynomial of degree 10 such that $D(z_j) = 0$ for $j = 1, \dots, 6$. Then none of the remaining roots of D has the absolute value of 1.

Let z_7, z_8, z_9, and z_{10} be the remaining roots of D (some of the roots may coincide). Let Φ be an arc of the unit circle $|z| = 1$ of length $l \le \arccos(1/8)$ that contains $z_1, \dots, z_6$. Consider the line L through the origin that bisects Φ. Since D is a raked polynomial, we must have

$$\sum_{j=1}^{10} z_j = \sum_{j=1}^{10} z_j^3 = 0, \tag{12}$$

cf. Lemma 3.3. Let Σ_1 be the sum of the orthogonal projections of $z_1, \dots, z_6$ onto L and let Σ_2 be the sum of the orthogonal projections of $z_7, \dots, z_{10}$ onto L, so

$$\Sigma_1 + \Sigma_2 = 0.$$

As $\cos l \ge 1/8$, we have $\cos(l/2) \ge 3/4$, and hence

$$|\Sigma_2| = |\Sigma_1| \ge 6 \cdot \frac{3}{4} = \frac{9}{2}. \tag{13}$$

Therefore, for at least one of the roots of D, say, z_9 we have $|z_9| > 1$. Then, for another root of D, say, z_{10} we have $|z_{10}| = 1/|z_9| < 1$, cf. Lemma 3.3. If $|z_7| > 1$ then $|z_8| < 1$ and we are done. Hence the only remaining case to consider is $|z_7| = |z_8| = 1$. In this case, by (13), we should have $|z_9| \ge 2$. Using that $z_{10} = 1/\overline{z_9}$ we obtain

$$|z_9^3 + z_{10}^3| = |z_9^3| + |z_{10}^3| > 8 = \sum_{j=1}^{8} |z_j|^3 \ge |\sum_{j=1}^{8} z_j^3|,$$

which contradicts (12). □

7.3 Lower Bounds for $c_j(d)$

I. Bárány [1] suggested to the authors to look at the following family of polytopes as a source of cs polytopes with many faces. Consider $\mathbb{R}^{2k}$ as a direct sum of k copies of $\mathbb{R}^2$:

$$\mathbb{R}^{2k} = \mathbb{R}^2 \oplus \cdots \oplus \mathbb{R}^2,$$

and let C_i denote the unit circle $\mathbb{S}^1 \subset \mathbb{R}^2$ in the ith copy. Let $n = km$ be the multiple of an even integer $m \ge 4$ and let $X_i \subset C_i$ be a set of m equally spaced points. Define

$$P_{n,k} := \operatorname{conv}\left(\bigcup_{i=1}^{k} X_i\right).$$

In other words, $P_{n,k}$ is the join of k cs m-gons. Thus $P_{n,k}$ is a cs $2k$-dimensional polytope with the property that for every subset of indices $I \subset \{1, \dots, k\}$ and a choice of points $x_i \in X_i$, one for each $i \in I$, the set $\mathrm{conv}(x_i : i \in I)$ is a face of $P_{n,k}$. Hence

$$f_j(P_{n,k}) \geq \binom{k}{j+1}\left(\frac{n}{k}\right)^{j+1} \quad \text{for } 0 \leq j \leq k-1,$$

which gives the bound

$$c_j(2k) \geq \frac{k(k-1)\cdots(k-j)}{k^{j+1}}.$$

We note that for $j = 1$ the obtained bound $c_1(2k) \geq 1 - k^{-1}$ is weaker than the bound $c_1(2k) \geq 1 - (2k-1)^{-1}$ of Theorem 1.3. Also for $j = k-1$, the obtained bound $c_{k-1}(2k) = k!/k^k \approx e^{-k}$ is weaker than the bound $c_{k-1}(2k) \geq 2^{-k}$ following from the proof of Theorem 1.4. Still, we can conclude that for any fixed j,

$$\lim_{d \to +\infty} c_j(2k) = 1.$$

7.4 Open Questions

There are several natural questions that we have not been able to answer so far.

- It seems plausible that ψ_k in Theorem 1.1 satisfies

$$\psi_k = \frac{2k-2}{2k-1}\pi,$$

  but we are unable to prove that.
- We do not know what is the best value of ϕ_k in Theorem 1.2 for $k > 2$ nor the values of $c_j(d)$ in Theorem 1.4.
- The most intriguing question is, of course, whether the class of polytopes $\mathcal{B}_{2k}(X)$ indeed provides (asymptotically or even exactly) polytopes with the largest number of faces among all centrally symmetric polytopes with a given number of vertices.

Acknowledgements The authors are grateful to J.E. Goodman, R. Pollack, and J. Pach, the organizers of the AMS–IMS–SIAM Summer Research Conference "Discrete and Computational Geometry—twenty years later" (Snowbird, June 2006), where this project started, to L. Billera for encouragement, to I. Bárány for suggesting the example of section 7.3, and to the anonymous referee for helpful comments.

References

1. Bárány, I.: personal communication (2007)
2. Barvinok, A.: A Course in Convexity. Graduate Studies in Mathematics, vol. 54. American Mathematical Society, Providence (2002)
3. Björner, A.: personal communication (2006)
4. Billera, L.J., Lee, C.W.: A proof of the sufficiency of McMullen's conditions for f-vectors of simplicial convex polytopes. J. Comb. Theory Ser. A **31**, 237–255 (1981)
5. Caratheodory, C.: Über den Variabilitatsbereich det Fourierschen Konstanten von Positiven harmonischen Furktionen. Rend. Circ. Mat. Palermo **32**, 193–217 (1911)

6. Danzer, L., Grünbaum, B.: Über zwei Probleme bezüglich konvexer Körper von P. Erdös und von V.L. Klee. Math. Z. **79**, 95–99 (1962) (in German)
7. Donoho, D.L.: High-dimensional centrosymmetric polytopes with neighborliness proportional to dimension. Discrete Comput. Geom. **35**, 617–652 (2006)
8. Donoho, D.L.: Neighborly polytopes and sparse solutions of underdetermined linear equations. Preprint (2004)
9. Donoho, D.L., Tanner, J.: Counting faces of randomly-projected polytopes when the projection radically lowers dimension. Preprint, math.MG/0607364
10. Gale, D.: Neighborly and cyclic polytopes. In: Proc. Sympos. Pure Math., vol. VII, pp. 225–232. American Mathematical Society, Providence (1963)
11. Grünbaum, B.: Convex polytopes, 2nd edn. (prepared and with a preface by V. Kaibel, V. Klee and G.M. Ziegler). Graduate Texts in Mathematics, vol. 221. Springer, New York (2003)
12. Grünbaum, B., Motzkin, T.S.: On polyhedral gaps. In: Proc. Sympos. Pure Math., vol. VII, pp. 285–290. American Mathematical Society, Providence (1963)
13. Katona, G.O.H.: A theorem of finite sets. In: Theory of Graphs, Proc. Colloq., Tihany, 1966, pp. 187–207. Academic Press, New York (1968)
14. Kruskal, J.B.: The number of simplices in a complex. In: Mathematical Optimization Techniques, pp. 251–278. University of California Press, Berkeley (1963)
15. Linial, N., Novik, I.: How neighborly can a centrally symmetric polytope be? Discrete Comput. Geom. **36**, 273–281 (2006)
16. McMullen, P.: The maximum numbers of faces of a convex polytope. Mathematika **17**, 179–184 (1970)
17. Motzkin, T.S.: Comonotone curves and polyhedra. Bull. Am. Math. Soc. **63**, 35 (1957)
18. Rudelson, M., Vershynin, R.: Geometric approach to error correcting codes and reconstruction of signals. Int. Math. Res. Not. **64**, 4019–4041 (2005)
19. Sheil-Small, T.: Complex Polynomials. Cambridge Studies in Advanced Mathematics, vol. 75. Cambridge University Press, Cambridge (2002)
20. Smilansky, Z.: Convex hulls of generalized moment curves. Isr. J. Math. **52**, 115–128 (1985)
21. Smilansky, Z.: Bi-cyclic 4-polytopes. Isr. J. Math. **70**, 82–92 (1990)
22. Stanley, R.: The number of faces of simplicial convex polytopes. Adv. Math. **35**, 236–238 (1980)
23. Ziegler, G.M.: Lectures on Polytopes. Graduate Texts in Mathematics, vol. 152. Springer, New York (1995)
24. Ziegler, G.M.: personal communication (2007)

On Projections of Semi-Algebraic Sets Defined by Few Quadratic Inequalities

Saugata Basu · Thierry Zell

Abstract Let $S \subset \mathbb{R}^{k+m}$ be a compact semi-algebraic set defined by $P_1 \geq 0, \ldots, P_\ell \geq 0$, where $P_i \in \mathbb{R}[X_1, \ldots, X_k, Y_1, \ldots, Y_m]$, and $\deg(P_i) \leq 2$, $1 \leq i \leq \ell$. Let π denote the standard projection from $\mathbb{R}^{k+m}$ onto $\mathbb{R}^m$. We prove that for any $q > 0$, the sum of the first q Betti numbers of $\pi(S)$ is bounded by $(k+m)^{O(q\ell)}$. We also present an algorithm for computing the first q Betti numbers of $\pi(S)$, whose complexity is $(k+m)^{2^{O(q\ell)}}$. For fixed q and ℓ, both the bounds are polynomial in $k+m$.

Keywords Betti numbers · Quadratic inequalities · Semi-algebraic sets · Spectral sequences · Cohomological descent

1 Introduction

Designing efficient algorithms for computing the Betti numbers of semi-algebraic sets is one of the outstanding open problems in algorithmic semi-algebraic geometry. There has been some recent progress in this area. It has been known for a while that the zero-th Betti number (which is also the number of connected components) of semi-algebraic sets can be computed in single exponential time. Very recently, it has been shown that even the first Betti number, and more generally the first q Betti numbers for any fixed constant q, can be computed in single exponential time [7, 9]. Since the problem of deciding whether a given semi-algebraic set in $\mathbb{R}^k$ is empty or

The author was supported in part by an NSF Career Award 0133597 and a Sloan Foundation Fellowship.

S. Basu (✉) · T. Zell
School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, USA
e-mail: saugata@math.gatech.edu

T. Zell
e-mail: zell@math.gatech.edu

not is NP-hard, and that of computing its zero-th Betti number is #P-hard, the existence of polynomial time algorithms for computing the Betti numbers is considered unlikely.

One particularly interesting case is that of semi-algebraic sets defined by quadratic inequalities. The class of semi-algebraic sets defined by quadratic inequalities is the first interesting class of semi-algebraic sets after sets defined by linear inequalities, in which case the problem of computing topological information reduces to linear programming for which (weakly) polynomial time algorithms are known. From the point of view of computational complexity, it is easy to see that the Boolean satisfiability problem can be posed as the problem of deciding whether a certain semi-algebraic set defined by quadratic inequalities is empty or not. Thus, deciding whether such a set is empty is clearly NP-hard and counting its number of connected components is #P-hard. However, semi-algebraic sets defined by quadratic inequalities are distinguished from arbitrary semi-algebraic sets in the sense that, *if the number of inequalities is fixed*, then the sum of their Betti numbers is bounded polynomially in the dimension. The following bound was proved by Barvinok [2].

Theorem 1.1 *Let $S \subset \mathbb{R}^k$ be a semi-algebraic set defined by the inequalities, $P_1 \geq 0, \ldots, P_\ell \geq 0$, $\deg(P_i) \leq 2, 1 \leq i \leq \ell$. Then, $\sum_{i=0}^{k} b_i(S) \leq k^{O(\ell)}$, where $b_i(S)$ denotes the i-th Betti number, which is the dimension of the i-th singular cohomology group of S, $H^i(S; \mathbb{Q})$, with coefficients in $\mathbb{Q}$.*

In view of Theorem 1.1, it is natural to consider the *class of semi-algebraic sets defined by a fixed number of quadratic inequalities* from a computational point of view. Algorithms for computing various topological properties of this class of semi-algebraic sets have been developed, starting from the work of Barvinok [1], who described an algorithm for testing whether a system of homogeneous quadratic equations has a projective solution. Barvinok's algorithm runs in polynomial time when the number of equations is constant. This was later generalized and made constructive by Grigoriev and Pasechnik in [16], where an algorithm is described for computing sample points in every connected component of a semi-algebraic set defined over a quadratic map. More recently, polynomial time algorithms have been designed for computing the Euler–Poincaré characteristic [8] as well as all the Betti numbers [6] of sets defined by a fixed number of quadratic inequalities (with different dependence on the number of inequalities in the complexity bound). Note also that the problem of deciding the emptiness of a set defined by a *single quartic equation* is already NP-hard and hence it is unlikely that there exists polynomial time algorithms for any of the above problems if the degree is allowed to be greater than two.

A case of intermediate complexity between semi-algebraic sets defined by polynomials of higher degree and sets defined by a fixed number of quadratic sign conditions is obtained by considering projections of such sets. The operation of linear projection of semi-algebraic sets plays a very significant role in algorithmic semi-algebraic geometry. It is a consequence of the Tarski–Seidenberg principle (see for example [10], p. 61) that the image of a semi-algebraic set under a linear projection is semi-algebraic, and designing efficient algorithms for computing properties of projections of semi-algebraic sets (such as its description by a quantifier-free formula) is a central problem of the area and is a very well-studied topic (see for example [22]

or [10], Chap. 14). However, the complexities of the best algorithms for computing descriptions of projections of general semi-algebraic sets is single exponential in the dimension and do not significantly improve when restricted to the class of semi-algebraic sets defined by a constant number of quadratic inequalities. Indeed, any semi-algebraic set can be realized as the projection of a set defined by quadratic inequalities, and it is not known whether quantifier elimination can be performed efficiently when the number of quadratic inequalities is kept constant. However, we show in this paper that, with a fixed number of inequalities, the projections of such sets are topologically simpler than projections of general semi-algebraic sets. This suggests, from the point of view of designing efficient (polynomial time) algorithms in semi-algebraic geometry, that projections of semi-algebraic sets defined by a constant number of quadratic inequalities is the next natural class of sets to consider, after sets defined by linear and (constant number of) quadratic inequalities, and this is what we proceed to do in this paper.

In this paper, we describe a polynomial time algorithm (Algorithm 2) for computing certain Betti numbers (including the zero-th Betti number which is the number of connected components) of projections of sets defined by a constant number of quadratic inequalities, without having to compute a semi-algebraic description of the projection. More precisely, let $S \subset \mathbb{R}^{k+m}$ be a compact semi-algebraic set defined by $P_1 \geq 0, \ldots, P_\ell \geq 0$, with $P_i \in \mathbb{R}[X_1, \ldots, X_k, Y_1, \ldots, Y_m]$, $\deg(P_i) \leq 2$, $1 \leq i \leq \ell$. Let $\pi : \mathbb{R}^{k+m} \to \mathbb{R}^m$ be the projection onto the last m coordinates. In what follows, the number of inequalities, ℓ, used in the definition of S will be considered as some fixed constant. Since, $\pi(S)$ is not necessarily describable using only quadratic inequalities, the bound in Theorem 1.1 does not hold for $\pi(S)$ and $\pi(S)$ can in principle be quite complicated. Using the best known complexity estimates for quantifier elimination algorithms over the reals (see [10]), we get single exponential (in k and m) bounds on the degrees and the number of polynomials necessary to obtain a semi-algebraic description of $\pi(S)$. In fact, there is no known algorithm for computing a semi-algebraic description of $\pi(S)$ in time polynomial in k and m. Nevertheless, we are able to prove that for any fixed constant $q > 0$, the sum of the first q Betti numbers of $\pi(S)$ are bounded by a polynomial in k and m. More precisely, we obtain the following complexity bound (see Sect. 4).

Theorem 1.2 *Let $S \subset \mathbb{R}^{k+m}$ be a compact semi-algebraic set defined by*

$$P_1 \geq 0, \ldots, P_\ell \geq 0, P_i \in \mathbb{R}[X_1, \ldots, X_k, Y_1, \ldots, Y_m], \deg(P_i) \leq 2, \ 1 \leq i \leq \ell.$$

Let $\pi : \mathbb{R}^{k+m} \to \mathbb{R}^m$ be the projection onto the last m coordinates. For any $q > 0$, $0 \leq q \leq k$,

$$\sum_{i=0}^{q} b_i(\pi(S)) \leq (k+m)^{O(q\ell)}.$$

We also consider the problem of computing the Betti numbers of $\pi(S)$. Previously, there was no polynomial time algorithm for computing any non-trivial topological property of projections of sets defined by few quadratic inequalities. We describe a polynomial time algorithm for computing the first few Betti numbers of $\pi(S)$. The

algorithm (Algorithm 2 in Sect. 7) computes $b_0(\pi(S)), \ldots, b_q(\pi(S))$. The complexity of the algorithm is $(k+m)^{2^{O(q\ell)}}$. If the coefficients of the input polynomials are integers of bit-size bounded by τ, then the bit-size of the integers appearing in the intermediate computations and the output are bounded by $\tau(k+m)^{2^{O(q\ell)}}$. Note that the output of the algorithm includes $b_0(\pi(S))$, which is the number of connected components of $\pi(S)$. Alternatively, one could obtain $b_0(\pi(S)), \ldots, b_q(\pi(S))$ by computing a semi-algebraic description of $\pi(S)$ using an efficient quantifier elimination algorithm (such as the one described in [4]) and then using the algorithm described in [7] to compute the first few Betti numbers. However, the complexity of this method would be worse: single exponential in k and m. Thus, our algorithm is able to compute efficiently non-trivial topological information about the projection, even though it does not compute a semi-algebraic description of that projection (it is not even known whether such a description could be computed in polynomial time).

In order to obtain Algorithm 2, we rely heavily on a certain spectral sequence, namely the *cohomological descent spectral sequence*. Even though variants of this spectral sequence have been known for some time [12, 13, 18, 21, 23, 24], to our knowledge this is the first time it has been used in designing efficient algorithms. As most constructions of the descent spectral sequence tend to use procedures which are infinitary in nature, it was more convenient for our algorithmic purposes to take a more constructive approach to building the sequence. This new construction is formally analogous to that of the Mayer–Vietoris spectral sequence, which has been used several times recently in designing algorithms for computing Betti numbers of semi-algebraic sets (see [5–7, 9]), and thus this construction (see Proposition 5.3 below) might be of independent interest.

2 Main Ideas

There are two main ingredients behind the results in this paper. The first is the use of cohomological descent, a spectral sequence first introduced by Deligne [12, 23] in the context of sheaf cohomology. This descent spectral sequence is used to compute the cohomology of the target of a continuous surjection (under certain hypotheses only, the limit of this spectral sequence is not, in general, the homology of the target). The first terms of the sequence are cohomology groups of certain fibered products over the surjection, and this allows to bound the Betti numbers of the target space in terms of the Betti numbers of those fibered products. This estimate was first used by Gabrielov, Vorobjov and Zell in [14] to give estimates on the Betti numbers of projections of semi-algebraic sets (and more generally, of semi-algebraic sets defined by arbitrary quantified formulas) without resorting to quantifier elimination. Another use of this sequence to establish upper-bounds can be found in [25] which contains effective estimates for the Betti numbers of semi-algebraic Hausdorff limits.

The most striking feature of this spectral sequence argument is that it enables one to deduce properties (for instance, bounds on the Betti numbers) of the projection of a set without having to explicitly describe the projection. For instance, consider a semi-algebraic subset of $\mathbb{R}^k$ defined by a polynomial having a constant number (say m) of monomials (often referred to as a fewnomial). It is known due to classical results of Khovansky [19] (see also [3]) that the Betti numbers of such sets can be

bounded in terms of m and k independent of the degree of the polynomial. Using the spectral sequence argument mentioned above, it was proved in [14] that even the Betti numbers of the projection of such a set can be bounded in terms of the number of monomials, even though it is known (see [15]) that the projection itself might not admit a description in terms of fewnomials.

The construction of the descent spectral sequence given in [14] involves consideration of join spaces and their filtrations and is not directly amenable for algorithmic applications. In Sect. 5, we give an alternate construction of a descent spectral sequence. When applied to surjections between open sets this spectral sequence converges to the cohomology of the image. The proof of this fact is formally analogous to the proof of convergence of the spectral sequence arising from the generalized Mayer–Vietoris sequence. This new proof allows us to identify a certain double complex, whose individual terms correspond to the chain groups of the fibered products of the original set. The fibered product (taken a constant number of times) of a set defined by few quadratic inequalities is again a set of the same type.

However, since there is no known algorithm for efficiently triangulating semi-algebraic sets (even those defined by few quadratic inequalities) we cannot directly use the spectral sequence to actually compute the Betti numbers of the projections. In order to do that we need an additional ingredient. This second main ingredient is a polynomial time algorithm described in [6] for computing a complex whose cohomology groups are isomorphic to those of a given semi-algebraic set defined by a constant number of quadratic inequalities. Using this algorithm we are able to construct a certain double complex, whose associated total complex is quasi-isomorphic to (implying having isomorphic homology groups) a suitable truncation of the one obtained from the cohomological descent spectral sequence mentioned above. This complex is of much smaller size and can be computed in polynomial time and is enough for computing the first q Betti numbers of the projection in polynomial time for any fixed constant q.

The rest of the paper is organized as follows. In Sect. 3 we recall certain basic facts from algebraic topology including the notions of complexes, and double complexes of vector spaces and spectral sequences. We do not prove any results since all of them are quite classical and we refer the reader to appropriate references [10, 11, 20] for the proofs. In Sect. 4 we prove the estimate on the sum of Betti numbers (Theorem 1.2) of projections of semi-algebraic sets defined by quadratic inequalities. In Sect. 5, we give our new construction of the cohomological descent spectral sequence. In Sect. 6, we briefly describe Algorithm 1 which is used to compute cohomology groups of semi-algebraic sets given by quadratic inequalities. This algorithm runs in polynomial time when the number of inequalities is constant. We only describe the inputs, outputs and the complexity estimates of the algorithms, referring the reader to [6] for more details. Finally, in Sect. 7 we describe our algorithm (Algorithm 2) for computing the first few Betti numbers of projections of semi-algebraic sets defined by quadratic inequalities.

3 Topological Preliminaries

We first recall some basic facts from algebraic topology, related to double complexes, and spectral sequences associated to double complexes as well as to continuous maps

between semi-algebraic sets. We refer the reader to [11, 20] for detailed proofs. We also fix our notations for these objects. All the facts that we need are well known, and we merely give a brief overview.

3.1 Complex of Vector Spaces

A *co-chain complex* is a sequence $C^\bullet = \{C^i \mid i \in \mathbb{Z}\}$ of $\mathbb{Q}$-vector spaces together with a sequence of homomorphisms $\delta^i : C^i \to C^{i+1}$ for which $\delta^{i+1} \circ \delta^i = 0$ for all p.

The cohomology groups, $H^i(C^\bullet)$ are defined by,

$$H^i(C^\bullet) = Z^i(C^\bullet)/B^i(C^\bullet),$$

where $B^i(C^\bullet) = \mathrm{Im}(\delta^{i-1})$, and $Z^i(C^\bullet) = \mathrm{Ker}(\delta^i)$. The cohomology groups, $H^*(C^\bullet)$, are all $\mathbb{Q}$-vector spaces (finite dimensional if the vector spaces C^i are themselves finite dimensional). We will henceforth omit reference to the field of coefficients $\mathbb{Q}$ which is fixed throughout the rest of the paper.

Given two complexes, $C^\bullet = (C^i, \delta^i)$ and $D^\bullet = (D^i, \partial^i)$, a homomorphism of complexes, $\phi : C^\bullet \to D^\bullet$, is a sequence of linear maps $\phi^i : C^i \to D^i$ verifying $\partial^i \circ \phi^i = \phi^{i+1} \circ \delta^i$ for all i.

In other words, the following diagram is commutative for all i.

$$\begin{array}{ccccccc}
\cdots & \longrightarrow & C^i & \xrightarrow{\delta^i} & C^{i+1} & \longrightarrow & \cdots \\
 & & \downarrow{\scriptstyle \phi^i} & & \downarrow{\scriptstyle \phi^{i+1}} & & \\
\cdots & \longrightarrow & D^i & \xrightarrow{\partial^i} & D^{i+1} & \longrightarrow & \cdots
\end{array}$$

A homomorphism of complexes, $\phi : C^\bullet \to D^\bullet$, induces homomorphisms, $\phi^* : H^*(C^\bullet) \to H^*(D^\bullet)$. The homomorphism ϕ is called a *quasi-isomorphism* if the homomorphisms ϕ^* are isomorphisms.

3.2 Double Complexes

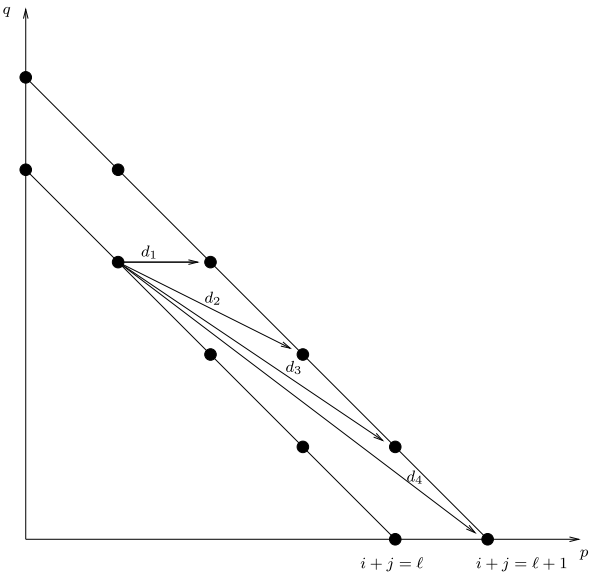
A double complex is a bi-graded vector space

$$C^{\bullet,\bullet} = \bigoplus_{i,j \in \mathbb{Z}} C^{i,j},$$

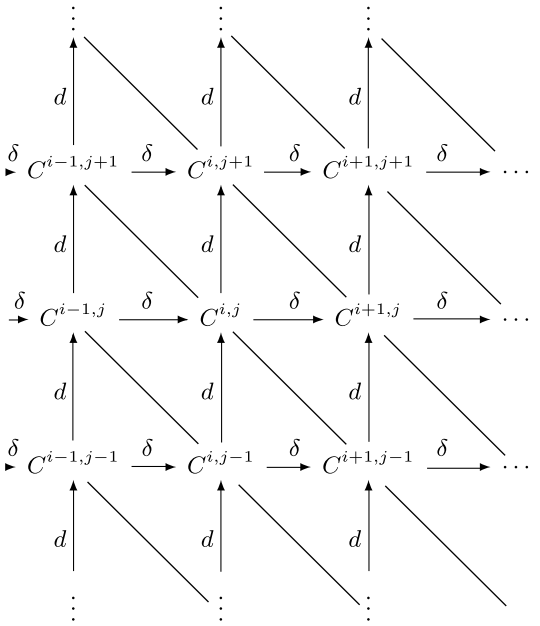
with co-boundary operators $d : C^{i,j} \to C^{i,j+1}$ and $\delta : C^{i,j} \to C^{i+1,j}$ such that $d^2 = \delta^2 = d\delta + \delta d = 0$. We say that $C^{\bullet,\bullet}$ is a first quadrant double complex if it satisfies the condition that $C^{i,j} = 0$ when $i, j < 0$.

Given a double complex $C^{\bullet,\bullet}$, we can construct a complex $\mathrm{Tot}^\bullet(C^{\bullet,\bullet})$, called the *associated total complex of* $C^{\bullet,\bullet}$ and defined by $\mathrm{Tot}^n(C^{\bullet,\bullet}) = \bigoplus_{i+j=n} C^{i,j}$, with

Fig. 1
$d_r : E_r^{i,j} \to E_r^{i+r,j-r+1}$



differential $D^n : \mathrm{Tot}^n(C^{\bullet,\bullet}) \longrightarrow \mathrm{Tot}^{n+1}(C^{\bullet,\bullet})$ given by $D^n = d + \delta$.



3.3 Spectral Sequences

A (*cohomology*) *spectral sequence* is a sequence of bi-graded complexes $\{E_r^{i,j} \mid i, j, r \in \mathbb{Z}, r \geq a\}$ endowed with differentials $d_r^{i,j} : E_r^{i,j} \to E_r^{i+r,j-r+1}$ such that

$(d_r)^2 = 0$ for all r. Moreover, we require the existence of isomorphism between the complex E_{r+1} and the homology of E_r with respect to d_r:

$$E_{r+1}^{i,j} \cong H_{d_r}(E_r^{i,j}) = \frac{\ker d_r^{i,j}}{d_r^{i+r,j-r+1}(E_r^{i+r,j-r+1})}.$$

The spectral sequence is called a *first quadrant spectral sequence* if the initial complex E_a lies in the first quadrant, i.e. $E_a^{i,j} = 0$ whenever $ij < 0$. In that case, all subsequent complexes E_r also lie in the first quadrant. Since the differential $d_r^{i,j}$ maps outside of the first quadrant for $r > i$, the homomorphisms of a first quadrant spectral sequence d_r are eventually zero, and thus the groups $E_r^{i,j}$ are all isomorphic to a fixed group $E_\infty^{i,j}$ for r large enough, and we say the spectral sequence is convergent.

Given a double complex $C^{\bullet,\bullet}$, we can associate to it two spectral sequences, ${}'E_*^{i,j}, {}''E_*^{i,j}$ (corresponding to taking row-wise or column-wise filtrations respectively).

If the double complex lies in the first quadrant, both of these spectral sequences are first quadrant spectral sequence, and both converge to $H^*(\mathrm{Tot}^\bullet(C^{\bullet,\bullet}))$, meaning that the limit groups verify

$$\bigoplus_{i+j=n} {}'E_\infty^{i,j} \cong \bigoplus_{i+j=n} {}''E_\infty^{i,j} \cong H^n(\mathrm{Tot}^\bullet(C^{\bullet,\bullet})), \tag{3.1}$$

for each $n \geq 0$.

The first terms of these are ${}'E_1 = H_\delta(C^{\bullet,\bullet})$, ${}'E_2 = H_d H_\delta(C^{\bullet,\bullet})$, and ${}''E_1 = H_d(C^{\bullet,\bullet})$, ${}''E_2 = H_\delta H_d(C^{\bullet,\bullet})$.

Given two (first quadrant) double complexes, $C^{\bullet,\bullet}$ and $\bar{C}^{\bullet,\bullet}$, a *homomorphism of double complexes* $\phi : C^{\bullet,\bullet} \to \bar{C}^{\bullet,\bullet}$ is a collection of homomorphisms, $\phi^{i,j} : C^{i,j} \to \bar{C}^{i,j}$, such that the following diagrams commute.

$$\begin{array}{ccc} C^{i,j} & \xrightarrow{\delta} & C^{i+1,j} \\ \downarrow{\scriptstyle \phi^{i,j}} & & \downarrow{\scriptstyle \phi^{i+1,j}} \\ \bar{C}^{i,j} & \xrightarrow{\delta} & \bar{C}^{i+1,j} \end{array}$$

$$\begin{array}{ccc} C^{i,j} & \xrightarrow{d} & C^{i,j+1} \\ \downarrow{\scriptstyle \phi^{i,j}} & & \downarrow{\scriptstyle \phi^{i,j+1}} \\ \bar{C}^{i,j} & \xrightarrow{d} & \bar{C}^{i,j+1} \end{array}$$

A homomorphism of double complexes, $\phi : C^{\bullet,\bullet} \to \bar{C}^{\bullet,\bullet}$ induces homomorphisms $\phi_r^{i,j} : E_r^{i,j} \to \bar{E}_r^{i,j}$ between the terms of the associated spectral sequences (corresponding either to the row-wise or column-wise filtrations).

We will need the following useful fact (see [20], p. 66, Theorem 3.4 for a proof).

Theorem 3.1 *If $\phi_s^{i,j}$ is an isomorphism for some $s \geq 1$ (and all i, j), then $E_r^{i,j}$ and $\bar{E}_r^{i,j}$ are isomorphic for all $r \geq s$. In other words, the induced homomorphism, ϕ : $\mathrm{Tot}^\bullet(C^{\bullet,\bullet}) \longrightarrow \mathrm{Tot}^\bullet(\bar{C}^{\bullet,\bullet})$ is a quasi-isomorphism.*

4 Proof of Theorem 1.2

The proof of Theorem 1.2 relies on the bounds from Theorem 1.1, and on the following theorem that appears in [14].

Theorem 4.1 *Let X and Y be two semi-algebraic sets and $f : X \to Y$ a semi-algebraic continuous surjection such that f is closed. Then for any integer n, we have*

$$b_n(Y) \leq \sum_{i+j=n} b_j(W_f^i(X)), \tag{4.1}$$

where $W_f^i(X)$ denotes the $(i+1)$-fold fibered product of X over f:

$$W_f^i(X) = \{(\bar{x}_0, \ldots, \bar{x}_i) \in X^{i+1} \mid f(\bar{x}_0) = \cdots = f(\bar{x}_i)\}.$$

This theorem follows from the existence of a spectral sequence $E_r^{i,j}$ converging to $H^*(Y)$ and such that $E_1^{i,j} \cong H^j(W_f^i(X))$. Since, in any spectral sequence, the dimensions of the terms $E_r^{i,j}$ are decreasing when i and j are fixed and r increases, we obtain using the definition (3.1) of convergence:

$$b_n(Y) = \sum_{i+j=n} \dim(E_\infty^{i,j}) \leq \sum_{i+j=n} \dim(E_1^{i,j}),$$

yielding inequality (4.1).

The spectral sequence $E_r^{i,j}$, known as *cohomological descent*, originated with the work of Deligne [12, 23], in the framework of sheaf cohomology. In [14], the sequence is obtained as the spectral sequence associated to the filtration of an infinite dimensional topological object, the *join space*, constructed from f. For the purposes of Algorithm 2, we will give a different construction of this sequence (see Sect. 5).

Proof of Theorem 1.2 Since S is compact, the semi-algebraic continuous surjection $\pi : S \to \pi(S)$ is closed: applying Theorem 4.1 to π, inequality (4.1) yields for each n with $0 \leq n \leq q$,

$$b_n(\pi(S)) \leq \sum_{i+j=n} b_j(W_\pi^i(S)). \tag{4.2}$$

Notice that $W_\pi^i(S) = \{(\bar{x}_0, \ldots, \bar{x}_i, y) \mid P_h(\bar{x}_t, y) \geq 0, 1 \leq h \leq \ell, 0 \leq t \leq i\}$. Thus, each $W_\pi^i(S) \subset \mathbb{R}^{(i+1)k+m}$ is defined by $\ell(i+1)$ quadratic inequalities. Applying the bound in Theorem 1.1 we get that,

$$b_j(W_\pi^i(S)) \leq ((i+1)k+m)^{O(\ell(i+1))}. \tag{4.3}$$

Using inequalities (4.2) and (4.3) and the fact that $q \leq k$, we get that,

$$\sum_{i=0}^{q} b_i(\pi(S)) \leq (k+m)^{O(q\ell)},$$

which proves the theorem. □

5 Cohomological Descent

This section is devoted a new construction of the cohomological descent spectral sequence (already discussed in Sect. 4). In Theorem 5.6, we obtain this sequence as the spectral sequence associated to a double complex associated to the fibered powers of X, rather than through the filtration of the join space. Convergence to the cohomology of the target space occurs when the map $f : X \to Y$ is *locally split* (see definition below). By deformation, we are able to extend the result to our case of interest: the projection of a compact semi-algebraic set (Corollary 5.8).

We will use this construction for Algorithm 2.

Definition 5.1 A continuous surjection $f : X \to Y$ is called *locally split* if there exists an open covering $\mathcal{U}$ of Y such that for all $U \in \mathcal{U}$, there exists a continuous section $\sigma : U \to X$ of f, i.e. σ is a continuous map such that $f(\sigma(y)) = y$ for all $y \in U$.

In particular, if X is an open semi-algebraic set and $f : X \to Y$ is a projection, the map f is obviously locally split. This specific case is what we will use in Algorithm 2, as we will reduce the projection of compact semi-algebraic sets to projections of open semi-algebraic sets (see Proposition 5.7) in order to apply the spectral sequence.

Recall that for any semi-algebraic surjection $f : X \to Y$, we denoted by $W_f^p(X)$ the $(p+1)$-fold fibered power of X over f,

$$W_f^p(X) = \{(\bar{x}_0, \ldots, \bar{x}_p) \in X^{p+1} \mid f(\bar{x}_0) = \cdots = f(\bar{x}_p)\}.$$

The map f induces for each $p \geq 0$, a map from $W_f^p(X)$ to Y, sending $(\bar{x}_0, \ldots, \bar{x}_p)$ to the common value $f(\bar{x}_0) = \cdots = f(\bar{x}_p)$, and abusing notations a little we will denote this map by f as well.

5.1 Singular (Co-)homology

We recall here the basic definitions related to singular (co-)homology theory directing the reader to [17] for details.

For any semi-algebraic set X, let $C_\bullet(X)$ denote the complex of singular chains of X with boundary map denoted by ∂.

Recall that $C_\bullet(X)$ is defined as follows: For $m \geq 0$, a singular m-simplex s is a continuous map, $s : \Delta_m \to X$, where Δ_m is the standard m-dimensional simplex defined by,

$$\Delta_m = \left\{ (t_0, \ldots, t_m) \mid t_i \geq 0, \sum_{i=0}^{m} t_i = 1 \right\}.$$

$C_m(X)$ is the vector space spanned by all singular m-simplices with boundary maps defined as follows. As usual we first define the face maps

$$f_{m,i} : \Delta_m \to \Delta_{m+1},$$

by $f_{m,i}((t_0, \ldots, t_m)) = (t_0, \ldots, t_{i-1}, 0, t_{i+1}, \ldots, t_{m+1})$.

For a singular m-simplex s we define

$$\partial s = \sum_{i=0}^{m} (-1)^i s \circ f_{m-1,i} \tag{5.1}$$

and extend ∂ to $C_m(X)$ by linearity. We will denote by $C^\bullet(X)$ the dual complex and by d the corresponding co-boundary map. More precisely, given $\phi \in C^m(X)$, and a singular $(m+1)$-simplex s of X, we have

$$d\phi(s) = \sum_{i=0}^{m+1} (-1)^i \phi(s \circ f_{m,i}). \tag{5.2}$$

If $f : X \to Y$ is a continuous map, then it naturally induces a homomorphism $f_* : C_\bullet(X) \to C_\bullet(Y)$ by defining, for each singular m-simplex $s : \Delta_m \to X$, $f_*(s) = s \circ f : \Delta_m \to Y$, which is a singular m-simplex of Y. We will denote by $f^* : C^\bullet(Y) \to C^\bullet(X)$ the dual homomorphism. More generally, suppose that $s = (s_0, \ldots, s_p) : \Delta_m \to W_f^p(X)$ is a singular m-simplex of $W_f^p(X)$. Notice that each component, $s_i, 0 \leq i \leq p$ are themselves singular m-simplices of X and that $f_*(s_0) = \cdots = f_*(s_p)$ are equal as singular m-simplices of Y. We will denote their common image by $f_*(s)$.

We will require the notion of small simplices subordinate to an open covering of a topological space (see [17]). Assuming that $f : X \to Y$ is locally split, let $\mathcal{U}$ be an open covering of Y on which local continuous sections exist. We denote by $\mathcal{V}$ the open covering of X given by the inverse images of elements of $\mathcal{U}$, i.e. $\mathcal{V} = \{f^{-1}(U) \mid U \in \mathcal{U}\}$. We let $C_\bullet^{\mathcal{U}}(Y)$ be the sub-complex of $C_\bullet(Y)$ spanned by those singular simplices of Y whose images are contained in some element of the cover $\mathcal{U}$. Similarly, we let $C_\bullet^{\mathcal{V}}(X)$ be the sub-complex of $C_\bullet(X)$ spanned by the simplices of X with image in $\mathcal{V}$, and more generally, for any integer p, $C_\bullet^{\mathcal{V}}(W_f^p(X))$ denotes the sub-complex of $C_\bullet(W_f^p(X))$ spanned by simplices with image contained in V^{p+1} for some $V \in \mathcal{V}$. The corresponding dual co chain complexes will be denoted by $C_{\mathcal{U}}^\bullet(Y)$ and $C_{\mathcal{V}}^\bullet(W_f^p(X))$ respectively. We will henceforth call any singular simplex of $C_\bullet^{\mathcal{U}}(Y)$ and any singular simplex of $C_\bullet^{\mathcal{V}}(W_f^p(X))$ *admissible* simplices.

The inclusion homomorphism, $\iota_\bullet : C_\bullet^{\mathcal{U}}(Y) \hookrightarrow C_\bullet(Y)$ induces a dual homomorphism, $\iota^\bullet : C^\bullet(Y) \to C_{\mathcal{U}}^\bullet(Y)$. We also have corresponding induced homomorphisms, $\iota^\bullet : C^\bullet(W_f^p(X)) \to C_{\mathcal{V}}^\bullet(W_f^p(X))$ for each $p \geq 0$.

Proposition 5.2 *The homomorphism $\iota^\bullet : C^\bullet(Y) \to C_{\mathcal{U}}^\bullet(Y)$ (resp. $C^\bullet(W_f^p(X)) \to C_{\mathcal{V}}^\bullet(W_f^p(X))$ for each $p \geq 0$) is a chain homotopy equivalence. In particular, we have $H^*(C_{\mathcal{U}}^\bullet(Y)) \cong H^*(C^\bullet(Y)) \cong H^*(Y)$ and $H^*(C_{\mathcal{V}}^\bullet(W_f^p(X))) \cong H^*(C^\bullet(W_f^p(X))) \cong H^*(W_f^p(X))$.*

Proof This follows from a similar result for homology, see Proposition 2.21 in [17]. □

5.2 A Long Exact Sequence

For each $p \geq 0$, we now define a homomorphism,

$$\delta^p : C^\bullet(W_f^p(X)) \longrightarrow C^\bullet(W_f^{p+1}(X))$$

as follows: for each i, $0 \leq i \leq p$, define $\pi_{p,i} : W_f^p(X) \to W_f^{p-1}(X)$ by,

$$\pi_{p,i}(x_0, \ldots, x_p) = (x_0, \ldots, \widehat{x_i}, \ldots, x_p)$$

($\pi_{p,i}$ drops the i-th coordinate).

We will denote by $(\pi_{p,i})_*$ the induced map on $C_\bullet(W_f^p(X)) \to C_\bullet(W_f^{p-1}(X))$ and let $\pi_{p,i}^* : C^\bullet(W_f^{p-1}(X)) \to C^\bullet(W_f^p(X))$ denote the dual map. For $\phi \in C^\bullet(W_f^p(X))$, we define $\delta^p\phi$ by,

$$\delta^p\phi = \sum_{i=0}^{p+1}(-1)^i \pi_{p+1,i}^*\, \phi. \tag{5.3}$$

Note that for any open covering $\mathcal{V}$ of X, the map δ^p induces by restriction a map $C_\mathcal{V}^\bullet(W_f^p(X)) \to C_\mathcal{V}^\bullet(W_f^{p+1}(X))$ which we will still denote by δ^p.

The following proposition is analogous to the exactness of the generalized Mayer–Vietoris sequence (cf. Lemma 1 in [5]).

Proposition 5.3 *Let $f : X \to Y$ be a continuous, locally split surjection, where X and Y are semi-algebraic subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. Let $\mathcal{U}$ denote an open covering of Y in which continuous sections of f can be defined on every $U \in \mathcal{U}$, and $\mathcal{V}$ denote the open covering of X obtained by inverse image of $\mathcal{U}$ under f. The following sequence is exact.*

$$0 \longrightarrow C_\mathcal{U}^\bullet(Y) \xrightarrow{f^*} C_\mathcal{V}^\bullet(W_f^0(X)) \xrightarrow{\delta^0} C_\mathcal{V}^\bullet(W_f^1(X)) \xrightarrow{\delta^1} \cdots$$

$$\cdots \xrightarrow{\delta^{p-1}} C_\mathcal{V}^\bullet(W_f^p(X)) \xrightarrow{\delta^p} C_\mathcal{V}^\bullet(W_f^{p+1}(X)) \xrightarrow{\delta^{p+1}} \cdots$$

Proof We will start by treating separately the first two positions in the sequence, then prove exactness for $p \geq 1$.

(A) $f^* : C_\mathcal{U}^\bullet(Y) \to C_\mathcal{V}^\bullet(X)$ is injective.

Let $U \in \mathcal{U}$ and let s be a simplex whose image is contained in U. If σ is a continuous section of f defined on U, the simplex $t = \sigma_*(s)$ is in $C_\bullet^\mathcal{V}(X)$, and verifies $f_*(t) = s$. Hence, $f_* : C_\bullet^\mathcal{V}(X) \to C_\bullet^\mathcal{U}(Y)$ is surjective, so f^* is injective.

(B) $f^*(C_\mathcal{U}^\bullet(Y)) = \ker \delta^0$.

Let $\phi \in C_\mathcal{V}^m(X)$. Any simplex $s \in C_m^\mathcal{V}(W_f^1(X))$ is a pair (s_0, s_1) of simplices in $C_m^\mathcal{V}(X)$ verifying $f_*(s_0) = f_*(s_1)$. We then have $\delta^0\phi(s) = \phi(s_1) - \phi(s_0)$. If $\phi = f^*\psi$ for some $\psi \in C_\mathcal{U}^m(Y)$, we have for any s,

$$\delta^0\phi(s) = f^*\psi(s_1) - f^*\psi(s_0) = \psi(f_*(s_1)) - \psi(f_*(s_0)) = 0,$$

since we must have $f_*(s_0) = f_*(s_1)$. Thus, we have $f^*(C^\bullet_{\mathcal{U}}(Y)) \subset \ker \delta^0$.

Conversely, if ϕ is such that $\delta^0\phi = 0$, this means that for any pair (s_0, s_1) of simplices in $C^{\mathcal{V}}_m(X)$ verifying $f_*(s_0) = f_*(s_1)$, we have $\phi(s_0) = \phi(s_1)$. Since we just proved in part (A) that $f_* : C^{\mathcal{V}}_\bullet(X) \to C^{\mathcal{U}}_\bullet(Y)$ is surjective, any element $t \in C^{\mathcal{U}}_m(Y)$ is of the form $t = f_*(s)$ for some $s \in C^{\mathcal{V}}_m(X)$. Thus, we can define $\psi \in C^m_{\mathcal{U}}(Y)$ by $\psi(t) = \phi(s)$, and the condition $\delta^0\phi = 0$ ensures that ψ is well defined since its value does not depend on the choice of s in the representation $t = f_*(s)$. This yields the reverse inclusion, and hence exactness at $p = 0$.

(C) $\delta^{p+1} \circ \delta^p = 0$.

From the definitions of the maps $\pi^*_{p+1,i}$, $\pi^*_{p+2,j}$ we have that for $0 \le i \le p+1$, $0 \le j \le p+2$,

$$\pi^*_{p+2,j} \circ \pi^*_{p+1,i}(\phi) = \pi^*_{p+2,i+1} \circ \pi^*_{p+1,j}(\phi) \text{ if } j < i. \tag{5.4}$$

Let $\phi \in C^m_{\mathcal{V}}(W^p_f(X))$. Now from the definitions of δ^p and δ^{p+1} we have that,

$$\begin{aligned}
\delta^{p+1} \circ \delta^p(\phi) &= \delta^{p+1}\left(\sum_{i=0}^{p+1}(-1)^i \pi^*_{p+1,i}(\phi)\right),\\
&= \sum_{i=0}^{p+1}(-1)^i \delta^{p+1}(\pi^*_{p+1,i}(\phi)),\\
&= \sum_{i=0}^{p+1}\sum_{j=0}^{p+2}(-1)^{i+j}\pi^*_{p+2,j} \circ \pi^*_{p+1,i}(\phi),\\
&= \sum_{i=0}^{p+1}\Bigg[\sum_{0\le j<i}(-1)^{i+j}\pi^*_{p+2,j} \circ \pi^*_{p+1,i}(\phi)\\
&\qquad + \sum_{i\le j\le p+2}(-1)^{i+j}\pi^*_{p+2,j} \circ \pi^*_{p+1,i}(\phi)\Bigg],\\
&= \sum_{i\le j}(-1)^{i+j}\pi^*_{p+2,j} \circ \pi^*_{p+1,i}(\phi)\\
&\qquad + \sum_{i>j}(-1)^{i+j}\pi^*_{p+2,j} \circ \pi^*_{p+1,i}(\phi).
\end{aligned}$$

Now using (5.4), the previous line becomes

$$= \sum_{i\le j}(-1)^{i+j}\pi^*_{p+2,j} \circ \pi^*_{p+1,i}(\phi) + \sum_{i>j}(-1)^{i+j}\pi^*_{p+2,i+1} \circ \pi^*_{p+1,j}(\phi).$$

Interchanging i and j in the second summand of the previous line, we get

$$= \sum_{i \leq j} (-1)^{i+j} \pi^*_{p+2,j} \circ \pi^*_{p+1,i}(\phi) + \sum_{i<j} (-1)^{i+j} \pi^*_{p+2,j+1} \circ \pi^*_{p+1,i}(\phi).$$

Finally, replacing $j+1$ by j in the second summand above, we obtain

$$= \sum_{i \leq j} (-1)^{i+j} \pi^*_{p+2,j} \circ \pi^*_{p+1,i}(\phi) + \sum_{i<j-1} (-1)^{i+j-1} \pi^*_{p+2,j} \circ \pi^*_{p+1,i}(\phi),$$

and isolating the terms corresponding to $j = i$ and $j = i+1$ in the first sum gives

$$\begin{aligned} &= (-1)^{2i} \pi^*_{p+2,i} \circ \pi^*_{p+1,i}(\phi) + (-1)^{2i+1} \pi^*_{p+2,i} \circ \pi^*_{p+1,i+1}(\phi) \\ &\quad + \sum_{i<j-1} (-1)^{i+j} \pi^*_{p+2,j} \circ \pi^*_{p+1,i}(\phi) + \sum_{i<j-1} (-1)^{i+j-1} \pi^*_{p+2,j} \circ \pi^*_{p+1,i}(\phi), \\ &= 0, \end{aligned}$$

(since, again, by (5.4), we have $\pi^*_{p+2,i} \circ \pi^*_{p+1,i+1} = \pi^*_{p+2,i} \circ \pi^*_{p+1,i}$).

(D) $\mathrm{Im}(\delta^p) \supset \mathrm{Ker}(\delta^{p+1})$.

Let $\phi \in \mathrm{Ker}(\delta^{p+1})$. In other words, for each admissible singular m-simplex $s = (s_0, \ldots, s_{p+1}) : \Delta_m \to W_f^{p+2}(X)$

$$\sum_{i=0}^{p+2} (-1)^i \phi((s_0, \ldots, \hat{s}_i, \ldots, s_{p+2})) = 0. \tag{5.5}$$

For each admissible singular m-simplex s of Y let s_* denote a fixed admissible singular m-simplex of X such that $f_*(s_*) = s$. Such a choice is possible since, as we proved in part (A), f_* is surjective onto $C^{\mathcal{U}}_\bullet(Y)$. Let $\psi \in C^m_{\mathcal{V}}(W_f^p(X))$ be defined as follows. For an admissible singular m-simplex $t = (t_0, \ldots, t_p)$ of $W_f^p(X)$ we define

$$\psi(t) = \phi(f_*(t)^*, t_0, \ldots, t_p).$$

Now for an admissible singular m simplex $t = (t_0, \ldots, t_{p+1})$ of W_f^{p+1}

$$\begin{aligned} \delta^p \psi(t) &= \sum_{i=0}^{p+1} (-1)^i \pi^*_{p+1,i} \psi(t) \\ &= \sum_{i=0}^{p+1} (-1)^i \psi((t_0, \ldots, \hat{t}_i, \ldots, t_{p+1})) \\ &= \sum_{i=0}^{p+1} (-1)^i \phi((f_*(t)^*, t_0, \ldots, \hat{t}_i, \ldots, t_{p+1})). \end{aligned}$$

Now let s denote the admissible singular m-simplex of $W_f^{p+2}(X)$ defined by $s = (f_*(t)^*, t_0, \ldots, \hat{t}_i, \ldots, t_{p+1})$. Now applying (5.5), we get

$$\sum_{i=0}^{p+2}(-1)^i \phi((s_0, \ldots, \widehat{s_i}, \ldots, s_{p+2})) = 0.$$

Separating the first term from the rest we obtain,

$$\phi((t_0, \ldots, t_{p+1})) = \sum_{i=0}^{p+1}(-1)^i \phi((f_*(t)^*, t_0, \ldots, \hat{t}_i, \ldots, t_{p+1})) = \delta^p \psi(t).$$

This finally proves the exactness of the sequence. □

5.3 The Descent Double Complex

Now, let $D^{\bullet,\bullet}(X)$ denote the double complex defined by, $D^{p,q}(X) = C^q(W_f^p(X))$ with vertical and horizontal homomorphisms given by $\tilde{d}^q = (-1)^p d^q$ and δ respectively, where d is the singular coboundary operator (5.2) and δ is the map defined in (5.3). Also, let $D^{p,q}(X) = 0$ if $p < 0$ or $q < 0$.

$$\begin{array}{ccccccccc}
 & & \vdots & & \vdots & & \vdots & & \\
 & & \uparrow \tilde{d} & & \uparrow \tilde{d} & & \uparrow \tilde{d} & & \\
0 & \longrightarrow & C^3(W_f^0(X)) & \xrightarrow{\delta} & C^3(W_f^1(X)) & \xrightarrow{\delta} & C^3(W_f^2(X)) & \longrightarrow \\
 & & \uparrow \tilde{d} & & \uparrow \tilde{d} & & \uparrow \tilde{d} & & \\
0 & \longrightarrow & C^2(W_f^0(X)) & \xrightarrow{\delta} & C^2(W_f^1(X)) & \xrightarrow{\delta} & C^2(W_f^2(X)) & \longrightarrow \\
 & & \uparrow \tilde{d} & & \uparrow \tilde{d} & & \uparrow \tilde{d} & & \\
0 & \longrightarrow & C^1(W_f^0(X)) & \xrightarrow{\delta} & C^1(W_f^1(X)) & \xrightarrow{\delta} & C^1(W_f^2(X)) & \longrightarrow \\
 & & \uparrow d & & \uparrow d & & \uparrow d & & \\
0 & \longrightarrow & C^0(W_f^0(X)) & \xrightarrow{\delta} & C^0(W_f^1(X)) & \xrightarrow{\delta} & C^0(W_f^2(X)) & \longrightarrow \\
 & & \uparrow d & & \uparrow d & & \uparrow d & & \\
 & & 0 & & 0 & & 0 & &
\end{array}$$

Lemma 5.4 *The families of maps $\tilde{d}$ and δ make $D^{\bullet,\bullet}$ into a double complex.*

Proof We need to check that $\tilde{d}^2 = \delta^2 = \tilde{d}\delta + \delta\tilde{d} = 0$. We know that $\tilde{d}^2 = d^2 = 0$ since $C^\bullet(W_f^p(X))$ is a cochain complex for all p, and we proved that $\delta^2 = 0$ in Proposition 5.3.

Now, suppose that $\phi \in C^q(W_f^p(X))$ and let $s = (s_0, \ldots, s_{p+1})$ be an admissible singular $(q+1)$-simplex of $W_f^{p+1}(X)$. Then,

$$\tilde{d}(\delta\phi)(s) = \tilde{d}\left(\sum_{i=0}^{p+1}(-1)^i \phi((s_0, \ldots, \widehat{s_i}, \ldots, s_{p+1}))\right),$$

$$= (-1)^p \sum_{j=0}^{q+1} \sum_{i=0}^{p+1} (-1)^{i+j} \phi(s_0 \circ f_{q,j}, \ldots, \widehat{s_i \circ f_{q,j}}, \ldots, s_{p+1} \circ f_{q,j}).$$

We also have

$$\delta(\tilde{d}\phi)(s) = \delta\left((-1)^{p+1} \sum_{j=0}^{q+1} (-1)^j \phi(s_0 \circ f_{q,j}, \ldots, s_{p+1} \circ f_{q,j}) \right),$$

$$= (-1)^{p+1} \sum_{j=0}^{q+1} \sum_{i=0}^{p+1} (-1)^{i+j} \phi(s_0 \circ f_{q,j}, \ldots, \widehat{s_i \circ f_{q,j}}, \ldots, s_{p+1} \circ f_{q,j}).$$

Thus, it follows that $\tilde{d}\delta + \delta\tilde{d} = 0$, so $D^{\bullet,\bullet}$ is indeed a double complex. □

If $f : X \to Y$ is locally split, and if $\mathcal{V}$ is the corresponding open covering of X defined in Sect. 5.1, the double complex $D^{\bullet,\bullet}$ induces by restriction a double complex $D_{\mathcal{V}}^{\bullet;\bullet}$, where $D_{\mathcal{V}}^{p,q} = C_{\mathcal{V}}^q(W_f^p(X))$ when $p \geq 0$ and $q \geq 0$ and $D_{\mathcal{V}}^{\bullet;\bullet} = 0$ otherwise.

The initial terms of the two spectral sequences associated with $D_{\mathcal{V}}^{\bullet;\bullet}$ (cf. Sect. 3.3) are as follows. The first terms of the spectral sequence ${}'E_*^{i,j}$ are ${}'E_1 = H_\delta(D_{\mathcal{V}}^{\bullet;\bullet}(X))$, ${}'E_2 = H_{\tilde{d}} H_\delta(D_{\mathcal{V}}^{\bullet;\bullet}(X))$. By the exactness of the sequence in Proposition 5.3, we have that the spectral sequence ${}'E_*^{i,j}$ degenerates at the ${}'E_2$ term as shown below.

$${}'E_1 = \left| \begin{array}{ccccc} \vdots & \vdots & \vdots & \vdots & \vdots \\ \uparrow d & \uparrow 0 & \uparrow 0 & \uparrow 0 & \uparrow 0 \\ C_{\mathcal{V}}^3(Y) & 0 & 0 & 0 & 0 \quad \cdots \\ \uparrow d & \uparrow 0 & \uparrow 0 & \uparrow 0 & \uparrow 0 \\ C_{\mathcal{V}}^2(Y) & 0 & 0 & 0 & 0 \quad \cdots \\ \uparrow d & \uparrow 0 & \uparrow 0 & \uparrow 0 & \uparrow 0 \\ C_{\mathcal{V}}^1(Y) & 0 & 0 & 0 & 0 \quad \cdots \\ \uparrow d & \uparrow 0 & \uparrow 0 & \uparrow 0 & \uparrow 0 \\ C_{\mathcal{V}}^0(Y) & 0 & 0 & 0 & 0 \quad \cdots \\ \hline \end{array} \right.$$

and, by Proposition 5.2,

$$
{}'E_2 = \begin{array}{|cccccc|}
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
H^3(Y) & 0 & 0 & 0 & 0 & 0 \cdots \\
H^2(Y) & 0 & 0 & 0 & 0 & 0 \cdots \\
H^1(Y) & 0 & 0 & 0 & 0 & 0 \cdots \\
H^0(Y) & 0 & 0 & 0 & 0 & 0 \cdots \\
\hline
\end{array}
$$

The degeneration of this sequence at ${}'E_2$ shows that

$$H^*(\mathrm{Tot}^\bullet(D^{\bullet,\bullet}_{\mathcal{V}}(X))) \cong H^*(Y).$$

The initial term ${}''E_1$ of the second spectral sequence is given by,

$$
{}''E_1 = \begin{array}{|ccccccc|}
\vdots & & \vdots & & \vdots & & \\
H^3(W^0_f(X)) & \xrightarrow{\delta} & H^3(W^1_f(X)) & \xrightarrow{\delta} & H^3(W^2_f(X)) & \longrightarrow \\
H^2(W^0_f(X)) & \xrightarrow{\delta} & H^2(W^1_f(X)) & \xrightarrow{\delta} & H^2(W^2_f(X)) & \longrightarrow \\
H^1(W^0_f(X)) & \xrightarrow{\delta} & H^1(W^1_f(X)) & \xrightarrow{\delta} & H^1(W^2_f(X)) & \longrightarrow \\
H^0(W^0_f(X)) & \xrightarrow{\delta} & H^0(W^1_f(X)) & \xrightarrow{\delta} & H^0(W^2_f(X)) & \longrightarrow \\
\hline
\end{array}
$$

Since this spectral sequence also converges to $H^*(\mathrm{Tot}^\bullet(D^{\bullet,\bullet}_{\mathcal{V}})(X))$, we have the following proposition.

Proposition 5.5

$$H^*(\mathrm{Tot}^\bullet(D^{\bullet,\bullet}_{\mathcal{V}})(X)) \cong H^*(Y).$$

Proposition 5.5 now implies,

Theorem 5.6 *For any continuous semi-algebraic surjection $f : X \to Y$, where X and Y are open semi-algebraic subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively (or, more generally, for any locally split continuous surjection f), the spectral sequence associated to the double complex $D^{\bullet,\bullet}(X)$ with $E_1 = H_d(D^{\bullet,\bullet}(X))$ converges to $H^*(C^\bullet(Y)) \cong H^*(Y)$. In particular,*

(A) $E_1^{i,j} = H^j(W^i_f(X))$, *and*

(B) $E_\infty \cong H^*(\mathrm{Tot}^\bullet(D^{\bullet,\bullet}(X))) \cong H^*(Y)$.

Proof By Proposition 5.2, we have that the component-wise homomorphisms, $\iota^\bullet$, induces a homomorphism of double complexes,

$$\iota^{\bullet,\bullet} : D^{\bullet,\bullet} \to D^{\bullet,\bullet}_{\mathcal{V}},$$

which in turn induces an isomorphism between the E_1 terms of the corresponding spectral sequences. Hence, by Theorem 3.1 we have that, $H^*(\mathrm{Tot}^\bullet(D^{\bullet,\bullet}_{\mathcal{V}})) \cong H^*(\mathrm{Tot}^\bullet(D^{\bullet,\bullet}))$. The Theorem now follows from Proposition 5.5. □

5.4 Truncation of the Double Complex

If we denote by $D^{\bullet,\bullet}_q(X)$ the truncated complex defined by,

$$\begin{aligned} D^{i,j}_q(X) &= D^{i,j}(X), \quad \text{if } 0 \le i+j \le q+1, \\ &= 0, \qquad\qquad \text{otherwise,} \end{aligned}$$

then it is clear that,

$$H^i(Y) \cong H^i(\mathrm{Tot}^\bullet(D^{\bullet,\bullet}_q(X))), \quad \text{for} \quad 0 \le i \le q. \tag{5.6}$$

Now suppose that $X \subset \mathbb{R}^{k+m}$ is a compact semi-algebraic set defined by the inequalities, $P_1 \ge 0, \ldots, P_\ell \ge 0$. Let π denote the projection map, $\pi : \mathbb{R}^{k+m} \to \mathbb{R}^m$. Let $\epsilon > 0$ and let $\tilde{X} \subset \mathbb{R}^{k+m}$ be the set defined by $P_1 + \epsilon > 0, \ldots, P_\ell + \epsilon > 0$.

Proposition 5.7

(A) *For $\epsilon > 0$ sufficiently small, we have*

$$H^*(W^p_\pi(\tilde{X})) \cong H^*(W^p_\pi(X)), \quad \textit{for all } p \ge 0,$$

$$\textit{and} \quad H^*(\pi(\tilde{X})) \cong H^*(\pi(X)).$$

(B) *The map, $\pi|_{\tilde{X}}$ is a locally split semi-algebraic surjection onto its image.*

Proof When $\epsilon > 0$ is small, the sets X and $\tilde{X}$ are homotopy equivalent and so are the sets $\pi(X)$ and $\pi(\tilde{X})$ and the fibered products $W^p_\pi(\tilde{X})$ and $W^p_\pi(X)$ for all $p \ge 0$ (see [3]). The first part of the proposition follows from the homotopy invariance property of singular cohomology groups. The second part of the proposition is clear once we note that $\tilde{X}$ is an open subset of $\mathbb{R}^{k+m}$: projections of open sets always admit local continuous sections. □

We can combine Theorem 5.6 and Proposition 5.7 to construct, from the projection of a *compact* basic semi-algebraic set, a double complex giving rise to a cohomological descent spectral sequence.

Corollary 5.8 *Let $X \subset \mathbb{R}^{k+m}$ be a compact semi-algebraic set defined by $P_1 \geq 0, \ldots, P_\ell \geq 0$ and $\pi : R^{k+m} \to \mathbb{R}^m$ the projection onto the last m co-ordinates. The spectral sequence associated to the double complex $D^{\bullet,\bullet}(X)$ with $E_1 = H_d(D^{\bullet,\bullet}(X))$ converges to $H^*(C^\bullet(\pi(X))) \cong H^*(\pi(X))$. In particular,*

(A) $E_1^{i,j} = H^j(W_f^i(X))$, *and*
(B) $E_\infty \cong H^*(\mathrm{Tot}^\bullet(D^{\bullet,\bullet}(X))) \cong H^*(\pi(X))$.

Remark 5.9 Note that it is not obvious how to prove directly an exact sequence at the level of singular (or even simplicial) cochains for the projection of a compact set, as we do in Proposition 5.3 in the locally-split setting. One difficulty is the fact that semi-algebraic maps are not, in general, triangulable.

Now let X be a compact semi-algebraic set defined by a constant number of quadratic inequalities and f a projection map. We cannot hope to compute even the truncated complex $D_q^{\bullet,\bullet}(X)$ since these are defined in terms of singular chain complexes which are infinite-dimensional. We overcome this problem by computing another double complex $\mathcal{D}_q^{\bullet,\bullet}(X)$, such that there exists a homomorphism of double complexes, $\psi : \mathcal{D}_q^{\bullet,\bullet}(X) \longrightarrow D_q^{\bullet,\bullet}(X)$, which induces an isomorphism between the $'E_1$ terms of the spectral sequences associated to the double complexes $D_q^{\bullet,\bullet}(X)$ and $\mathcal{D}_q^{\bullet,\bullet}(X)$. This implies, by virtue of Theorem 3.1, that the cohomology groups of the associated total complexes are isomorphic, that is,

$$H^*(\mathrm{Tot}^\bullet(D_q^{\bullet,\bullet}(X))) \cong H^*(\mathrm{Tot}^\bullet(\mathcal{D}_q^{\bullet,\bullet}(X))).$$

The construction of the double complex $\mathcal{D}_q^{\bullet,\bullet}(X)$ is described in Sect. 7.

6 Algorithmic Preliminaries

We now recall an algorithm described in [6], where the following theorem is proved.

Theorem 6.1 *There exists an algorithm, which takes as input a family of polynomials $\{P_1, \ldots, P_s\} \subset \mathbb{R}[X_1 \ldots, X_k]$, with $\deg(P_i) \leq 2$, and a number $\ell \leq k$, and outputs a complex $\mathcal{D}_\ell^{\bullet,\bullet}$. The complex $\mathrm{Tot}^\bullet(\mathcal{D}_\ell^{\bullet,\bullet})$ is quasi-isomorphic to $\mathrm{C}_\bullet^\ell(S)$, the truncated singular chain complex of S, where*

$$S = \bigcap_{P \in \mathcal{P}} \{x \in \mathbb{R}^k \mid P(x) \leq 0\}.$$

Moreover, given a subset $\mathcal{P}' \subset \mathcal{P}$, with

$$S' = \bigcap_{P \in \mathcal{P}'} \{x \in \mathbb{R}^k \mid P(x) \leq 0\}$$

the algorithm outputs both complexes $\mathcal{D}_\ell^{\bullet,\bullet}$ and $\mathcal{D}'^{\bullet,\bullet}_\ell$ (corresponding to the sets S and S' respectively) along with the matrices defining a homomorphism $\Phi_{\mathcal{P},\mathcal{P}'}$, such

that $\Phi^*_{\mathcal{P},\mathcal{P}'} : H^*(\mathrm{Tot}^\bullet(\mathcal{D}^{\bullet,\bullet}_\ell)) \cong H^*(S) \to H^*(S') \cong H^*(\mathrm{Tot}^\bullet(\mathcal{D}'^{\bullet,\bullet}_\ell))$ *is the homomorphism induced by the inclusion* $i : S \hookrightarrow S'$. *The complexity of the algorithm is* $\sum_{i=0}^{\ell+2} \binom{s}{i} k^{2^{O(\min(\ell,s))}}$.

For completeness, we formally state the input and output of the algorithm mentioned in Theorem 6.1.

We first introduce some notations which will be used to describe the input and output of the algorithm. Let $\mathcal{Q} = \{Q_1, \ldots, Q_s\} \subset \mathbb{R}[X_1, \ldots, X_k]$ be a family of polynomials with $\deg(Q_i) \leq 2$, $1 \leq i \leq s$. For each subset $J \subset \{1, \ldots, s\}$, we let S_J denote the semi-algebraic set defined by $\{Q_j \geq 0 \mid j \in J\}$. Notice that for each pair $I \subset J \subset \{1, \ldots, s\}$, we have an inclusion $S_J \subset S_I$.

Algorithm 1 (Build Complex)

Input: A family of polynomials $\mathcal{Q} = \{Q_1, \ldots, Q_s\} \subset \mathbb{R}[X_1, \ldots, X_k]$ with $\deg(Q_i) \leq 2$, for $1 \leq i \leq s$.

Output:

(A) For each subset $J \subset \{1, \ldots, s\}$, a description of a complex $F^\bullet_J$, consisting of a basis for each term of the complex and matrices (in this basis) for the differentials, and

(B) for each pair $I \subset J \subset \{1, \ldots, s\}$, a homomorphism, $\phi_{I,J} : F^\bullet_I \longrightarrow F^\bullet_J$.

The complexes, $F^\bullet_J$ and the homomorphisms $\phi_{I,J}$ satisfy the following.

(A) For each $J \subset \{1, \ldots, s\}$,

$$H^*(F^\bullet_J) \cong H^*(S_J). \tag{6.1}$$

(B) For each pair $I \subset J \subset \{1, \ldots, s\}$, the following diagram commutes.

$$\begin{array}{ccc} H^*(F^\bullet_I) & \xrightarrow{(\phi_{I,J})^*} & H^*(F^\bullet_J) \\ \cong \uparrow & & \uparrow \cong \\ H^*(S_I) & \xrightarrow{r^*} & H^*(S_J) \end{array}$$

Here, $(\phi_{I,J})^*$ is the homomorphism induced by $\phi_{I,J}$, the vertical homomorphisms are the isomorphisms from (6.1), and r^* is the homomorphism induced by restriction.

Complexity: The complexity of the algorithm is $k^{2^{O(s)}}$. □

For the purposes of this paper, we need to slightly modify Algorithm 1 in order to be able to handle permutations of the co-ordinates. More precisely, suppose that $\sigma \in \mathfrak{S}_k$ is a given permutation of the co-ordinates, and for any $I \subset \{1, \ldots, s\}$, let $S_{I,\sigma} = \{(x_{\sigma(1)}, \ldots, x_{\sigma(k)}) \mid (x_1, \ldots, x_k) \in S_I\}$. Let $F^\bullet_{I,\sigma}$ denote the complex computed by the algorithm corresponding to the set $S_{I,\sigma}$. It is easy to modify Algorithm 1 slightly without changing the complexity estimate, such that for any fixed σ, the algorithm outputs, complexes $F^\bullet_I, F^\bullet_{I,\sigma}$ as well as the matrices corresponding to the

induced isomorphisms, $\phi_\sigma^\bullet : F_I^\bullet \to F_{I,\sigma}^\bullet$. We assume this implicitly in the description of Algorithm 2 in the next section.

7 Algorithm for Projections

Let $S \subset \mathbb{R}^{k+m}$ be a basic semi-algebraic set defined by

$$P_1 \geq 0, \ldots, P_\ell \geq 0, P_i \in \mathbb{R}[X_1, \ldots, X_k, Y_1, \ldots, Y_m],$$

with $\deg(P_i) \leq 2,\ 1 \leq i \leq \ell$. Let $\pi : \mathbb{R}^{k+m} \to \mathbb{R}^m$ be the projection onto the last m coordinates.

The algorithm will compute a double complex, $\mathcal{D}_q^{\bullet,\bullet}(S)$, such that $\mathrm{Tot}^\bullet(\mathcal{D}_q^{\bullet,\bullet}(S))$ is quasi-isomorphic to the complex $\mathrm{Tot}^\bullet(D_q^{\bullet,\bullet}(S))$. The double complex, $\mathcal{D}_q^{\bullet,\bullet}(S)$ is defined as follows.

We introduce $k(q+2)$ variables, which we denote by $X_{i,j}$, $1 \leq i \leq k, 0 \leq j \leq q+1$. For each $j, 0 \leq j \leq q+1$, we denote by, $P_{i,j}$ the polynomial

$$P_i(X_{1,j}, \ldots, X_{k,j}, Y_1, \ldots, Y_m)$$

(substituting $X_{1,j}, \ldots, X_{k,j}$ in place of $X_1, \ldots, X_k$ in the polynomial P_i). We consider each $P_{i,j}$ to be an element of $\mathbb{R}[X_{1,0}, \ldots, X_{k,q+1}, Y_1, \ldots, Y_m]$. For each $p, 0 \leq p \leq q+1$, we denote by $S_p \subset \mathbb{R}^{k(q+2)+m}$ the semi-algebraic set defined by,

$$P_{1,0} \geq 0, \ldots, P_{\ell,0} \geq 0, \ldots, P_{1,p} \geq 0, \ldots, P_{\ell,p} \geq 0.$$

Note that, for each $p, 0 < p \leq q+1$, and each $j, 0 \leq j \leq p$ we have a natural map, $\pi_{p,j} : S_p \to S_{p-1}$ given by,

$$\pi_{p,j}(\bar{x}_0, \ldots, \bar{x}_p, \ldots, \bar{x}_{q+1}, \bar{y}) = (\bar{x}_0, \ldots, \bar{x}_p, \ldots, \bar{x}_j, \ldots, \bar{x}_{q+1}, \bar{y}).$$

Note that in the definition above, each $\bar{x}_i \in \mathbb{R}^k$ and $\pi_{p,j}$ exchanges the coordinates $\bar{x}_j$ and $\bar{x}_p$.

We are now in a position to define $\mathcal{D}_q^{\bullet,\bullet}$. We follow the notations introduced in Sect. 6. Let $\mathcal{Q} = \{Q_1, \ldots, Q_{\ell(q+2)}\} = \{P_{1,0}, \ldots, P_{\ell,q+1}\}$. For $0 \leq j \leq q+1$, we let $L_j = \{1, \ldots, (j+1)\ell\} \subset \{1, \ldots, (q+2)\ell\}$.

$$\begin{aligned} \mathcal{D}_q^{i,j}(X) &= F_{L_i}^j, \quad 0 \leq i+j \leq q+1, \\ &= 0, \qquad \text{otherwise.} \end{aligned}$$

The vertical homomorphisms, d, in the complex $\mathcal{D}_q^{\bullet,\bullet}$ are those induced from the complexes $F_{L_i}^\bullet$ or zero. The horizontal homomorphisms, $\delta^j : F_{L_i}^j \longrightarrow F_{L_{i+1}}^j$ are defined as follows.

For each $h,\ 0 \leq h \leq i+1$, Algorithm 1 produces a homomorphism, $\phi_{i+1,h} : F_{L_i}^j \longrightarrow F_{L_{i+1}}^j$, corresponding to the map $\pi_{i+1,h}$ (see remark after Algorithm 1). The homomorphism δ is then defined by, $\delta = \sum_{h=0}^{i+1}(-1)^h \phi_{i+1,h}$. We have the following proposition.

Proposition 7.1 *The complex* $\mathrm{Tot}^\bullet(\mathcal{D}_q^{\bullet,\bullet}(S))$ *is quasi-isomorphic to the complex* $\mathrm{Tot}^\bullet(D_q^{\bullet,\bullet}(S))$.

Proof It follows immediately from Theorem 6.1 that the columns of the complexes $\mathcal{D}_q^{\bullet,\bullet}(S)$ and $D_q^{\bullet,\bullet}(S)$ are quasi-isomorphic. Moreover, it is easy to see that the quasi-isomorphisms induce an isomorphism between the ${}''E_1$ term of their associated spectral sequences. Now by Theorem 3.1 this implies that $\mathrm{Tot}^\bullet(\mathcal{D}_q^{\bullet,\bullet}(S))$ is quasi-isomorphic to the complex $\mathrm{Tot}^\bullet(D_q^{\bullet,\bullet}(S))$. □

Algorithm 2 (Computing the first q Betti Numbers)

Input: A $S \subset \mathbb{R}^{k+m}$ be a basic semi-algebraic set defined by

$$P_1 \geq 0, \ldots, P_\ell \geq 0,$$

with $P_i \in \mathbb{R}[X_1, \ldots, X_k, Y_1, \ldots, Y_m]$, $\deg(P_i) \leq 2$, $1 \leq i \leq \ell$.

Output: $b_0(\pi(S)), \ldots, b_q(\pi(S))$, where $\pi : \mathbb{R}^{k+m} \to \mathbb{R}^m$ be the projection onto the last m coordinates.

Procedure:

Step 1: Using Algorithm 1 compute the truncated complex $\mathcal{D}_q^{\bullet,\bullet}(S)$.

Step 2: Compute using linear algebra, the dimensions of $H^i(\mathrm{Tot}^\bullet(\mathcal{D}_q^{\bullet,\bullet}))$, $0 \leq i \leq q$.

Step 3: For each i, $0 \leq i \leq q$, output, $b_i(\pi(S)) = \dim(H^i(\mathrm{Tot}^\bullet(\mathcal{D}_q^{\bullet,\bullet})))$.

Complexity Analysis: The calls to Algorithm 1 has input consisting of $(q+1)\ell$ polynomials in $qk+m$ variables. Using the complexity bound of Algorithm 1 we see that the complexity of Algorithm 2 is bounded by $(k+m)^{2^{O(q\ell)}}$. □

Proof of Correctness: The correctness of the algorithm is a consequence of Proposition 7.1 and Theorem 3.1. □

8 Conclusion and Open Problems

For any fixed q and ℓ, we have proved a polynomial bound on the sum of the first q Betti numbers of the projection of a bounded, basic closed semi-algebraic set defined by ℓ quadratic inequalities. We have also described a polynomial time algorithm to compute the first q Betti numbers of the image of such a projection.

Since it is not known whether quantifier elimination can be performed efficiently for sets defined by a fixed number of quadratic inequalities, many questions are left open.

Our bounds become progressively worse as q increases, becoming exponential in the dimension as q approaches k. However, we do not have any examples (of projections of semi-algebraic sets defined by quadratic inequalities) where the higher Betti numbers behave exponentially in the dimension. This leaves open the problem of either constructing such examples, or removing the dependence on q from our bounds.

Another interesting open problem is to improve the complexity of Algorithm 2, from $(k+m)^{2^{O(q\ell)}}$ to $(k+m)^{O(q\ell)}$. Note that this would imply an algorithm with

complexity $k^{O(q\ell)}$ for computing the first q Betti numbers of a semi-algebraic set defined by ℓ quadratic inequalities in $\mathbb{R}^k$. The best known algorithm for computing all the Betti numbers of such sets has complexity $k^{2^{O(\ell)}}$ [6]. The only topological invariants of such sets that we currently know how to compute in time $k^{O(\ell)}$ are testing for emptiness [1, 16] and the Euler–Poincaré characteristic [8].

References

1. Barvinok, A.I.: Feasibility testing for systems of real quadratic equations. Discret. Comput. Geom. **10**, 1–13 (1993)
2. Barvinok, A.I.: On the Betti numbers of semi-algebraic sets defined by few quadratic inequalities. Math. Z. **225**, 231–244 (1997)
3. Basu, S.: On bounding the Betti numbers and computing the Euler characteristics of semi-algebraic sets. Discret. Comput. Geom. **22**, 1–18 (1999)
4. Basu, S., Pollack, R., Roy, M.-F.: On the combinatorial and algebraic complexity of quantifier elimination. J. ACM **43**, 1002–1045 (1996)
5. Basu, S.: On different bounds on different Betti numbers. Discret. Comput. Geom. **30**(1), 65–85 (2003)
6. Basu, S.: Polynomial time algorithm for computing the top Betti numbers of semi-algebraic sets defined by quadratic inequalities, Found. Comput. Math. (2006, in press)
7. Basu, S.: Single exponential time algorithm for computing the first few Betti numbers of semi-algebraic sets. J. Symb. Comput. **41**(10), 1125–1154 (2006)
8. Basu, S.: Efficient algorithm for computing the Euler–Poincaré characteristic of semi-algebraic sets defined by few quadratic inequalities. Comput. Complex. **15**, 236–251 (2006)
9. Basu, S., Pollack, R., Roy, M.-F.: Computing the first Betti number and the connected components of semi-algebraic sets. Found. Comput. Math. (2007, to appear)
10. Basu, S., Pollack, R., Roy, M.-F.: In: Algorithms in Real Algebraic Geometry, 2nd edn. Algorithms and Computation in Mathematics, vol. 10. Springer, Berlin (2006)
11. Bredon, G.E.: Sheaf Theory. Springer, Berlin (1996)
12. Deligne, P.: Théorie de Hodge III. Publ. Math. IHES **44**, 5–77 (1974)
13. Dugger, D., Isaksen, D.: Topological hypercovers and A^1-realizations. Math. Z. **246**, 667–689 (2004)
14. Gabrielov, A., Vorobjov, N., Zell, T.: Betti numbers of semi-algebraic and sub-Pfaffian sets. J. Lond. Math. Soc. **69**(2), 27–43 (2004)
15. Gabrielov, A.: Counter-examples to quantifier elimination for fewnomial and exponential expressions. Preprint, available at http://www.math.purdue.edu/~agabriel/preprint.html
16. Grigor'ev, D., Pasechnik, D.V.: Polynomial time computing over quadratic maps I. Sampling in real algebraic sets. Comput. Complex. **14**, 20–52 (2005)
17. Hatcher, A.: Algebraic Topology. Cambridge University Press, Cambridge (2002)
18. Houston, K.: An introduction to the image computing spectral sequence. In: Singularity Theory, Liverpool, 1996. London Math. Soc. Lecture Notes Ser., vol. 263, pp. 305–324. Cambridge University Press, Cambridge (1999)
19. Khovansky, A.G.: Fewnomials. American Mathematical Society, Providence (1991)
20. McCleary, J.: A User's Guide to Spectral Sequences, 2nd edn. Cambridge Studies in Advanced Mathematics (2001)
21. Murray, M.: Bundle gerbes. J. Lond. Math. Soc. **54**, 403–416 (1996)
22. Renegar, J.: On the computational complexity and geometry of the first order theory of the reals. J. Symb. Comput. **13**, 255–352 (1992)
23. Saint-Donat, B.: Techniques de descente cohomologique. In: Théorie des Topos et Cohomologie Étale des Schémas. Tome 2 (SGA 4). Lecture Notes in Mathematics, vol. 270, pp. 83–162. Springer, Berlin (1972)
24. Vassiliev, V.: In: Complements of Discriminants of Smooth Maps: Topology and Applications. Translations of Mathematical Monographs, vol. 98. American Mathematical Society, Providence (1992)
25. Zell, T.: Topology of definable Hausdorff limits. Discret. Comput. Geom. **33**, 423–443 (2005)

Enumeration in Convex Geometries and Associated Polytopal Subdivisions of Spheres

Louis J. Billera · Samuel K. Hsiao · J. Scott Provan

Abstract We construct CW spheres from the lattices that arise as the closed sets of a convex closure, the meet-distributive lattices. These spheres are nearly polytopal, in the sense that their barycentric subdivisions are simplicial polytopes. The complete information on the numbers of faces and chains of faces in these spheres can be obtained from the defining lattices in a manner analogous to the relation between arrangements of hyperplanes and their underlying geometric intersection lattices.

Keywords Abstract convexity · Quasisymmetric functions · Meet-distributive lattice · Join-distributive lattice

1 Introduction

A well known result due to Zaslavsky [29] shows that the numbers of faces in an arrangement of hyperplanes in a real Euclidean space can be read from the underlying geometric lattice of all intersections of these hyperplanes. This result was extended

The first author was supported in part by NSF grant DMS-0100323. The second author was supported by an NSF Postdoctoral Fellowship. The first two authors enjoyed the hospitality of the Mittag-Leffler Institute, Djursholm, Sweden, during the preparation of this manuscript.

L.J. Billera (✉)
Department of Mathematics, Cornell University, Ithaca, NY 14853-4201, USA
e-mail: billera@math.cornell.edu

S.K. Hsiao
Mathematics Program, Bard College, PO Box 5000, Annandale-on-Hudson, NY 12504-5000, USA
e-mail: hsiao@bard.edu

J.S. Provan
Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599-3260, USA
e-mail: Scott_Provan@UNC.edu

to the determination of the numbers of chains of faces in arrangements in [2, 8]. In [3], the numbers of chains in an arrangement were shown to depend only on the numbers of chains in the associated geometric lattice. A particularly simple form of this relationship, in terms of quasisymmetric functions, was given in [5].

Geometric lattices (matroids) are combinatorial abstractions of linear span in vector spaces. There is a different combinatorial model for convex span, known as convex geometries (or anti-matroids) [12–15], for which the corresponding lattices are the meet-distributive lattices. We show here that a similar situation exists for these; that is, for each convex geometry, we construct a regular CW sphere, whose enumerative properties are related to those of the underlying geometry in essentially the same way. Moreover, these spheres are nearly polytopes, in the sense that their first barycentric subdivisions are combinatorially simplicial convex polytopes.

We begin by establishing some notation. Our basic object of study is a combinatorial closure operation called a *convex* or *anti-exchange* closure. This is defined on a finite set, which we will take, without loss of generality, to be the set $[n] := \{1, 2, \ldots, n\}$.

Definition 1.1 A *convex closure* is a function $\langle\cdot\rangle : 2^{[n]} \to 2^{[n]}$, $A \mapsto \langle A\rangle$, such that, for $A, B \subseteq [n]$,

(1) $A \subseteq \langle A\rangle$
(2) if $A \subseteq B$ then $\langle A\rangle \subseteq \langle B\rangle$
(3) $\langle A\rangle = \langle\langle A\rangle\rangle$
(4) if $x, y \notin \langle A\rangle$ and $x \in \langle A \cup y\rangle$ then $y \notin \langle A \cup x\rangle$.

The last condition is often called the *anti-exchange* axiom, and the complements of the closed sets of such a closure system has been called an *anti-matroid*. We will call a set together with a convex closure operator on it a *convex geometry*. The set of *closed sets* of a convex geometry, that is, those sets A satisfying $A = \langle A\rangle$, form a lattice when ordered by set inclusion. Such lattices are precisely the meet-distributive lattices. A lattice L is *meet-distributive* if for each $y \in L$, if $x \in L$ is the meet of (all the) elements covered by y, then the interval $[x, y]$ is a Boolean algebra.

One example of an anti-exchange closure operator is ideal closure on a partially ordered set P; here, for $A \subseteq P$, $\langle A\rangle$ denotes the (lower) order ideal generated by A. The lattices of closed sets of these are precisely the distributive lattices. Another class of examples comes from considering convex closure on a finite point set in Euclidean space. Figure 1 illustrates the convex geometry formed by three collinear points a, b, c. Note that the set $\{a, c\}$ is not closed since its closure is $\{a, b, c\}$.

Meet-distributive lattices were first studied by Dilworth [11] and have reappeared in many contexts since then (see [21]). Their study in the context of theory of convex geometries was extensively developed about 20 years ago in a series of papers by Edelman and coauthors [12–15]. See also [20] and [10] for general discussions. An important (and characterizing) property of convex geometries is that every set has a unique minimal generating set, that is, for each $A \subseteq [n]$, there is a unique minimal subset $\mathrm{ext}(A) \subseteq A$ so that $\langle A\rangle = \langle \mathrm{ext}(A)\rangle$ [13, Theorem 2.1]. The elements of $\mathrm{ext}(A)$ are called the *extreme points* of A.

A *simplicial complex* on a finite set V is a family of subsets $\Delta \subseteq 2^V$ such that if $\tau \subseteq \sigma \in \Delta$ then $\tau \in \Delta$ and $\{v\} \in \Delta$ for all $v \in V$. The elements of Δ are called the

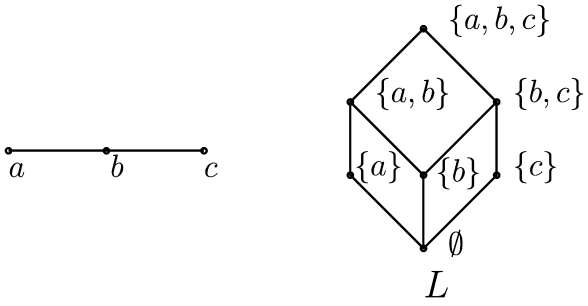


Fig. 1 The convex geometry of three collinear points and its associated meet-distributive lattice L

faces of the complex and the elements of V are its *vertices*. We will need an operation on simplicial complexes known as *stellar subdivision*.

Definition 1.2 The *stellar subdivision* of a simplicial complex Δ over a nonempty face $\sigma \in \Delta$ is the simplicial complex $sd_\sigma(\Delta)$ on the set $V \cup \{v_\sigma\}$, where v_σ is a new vertex, consisting of

(1) all $\tau \in \Delta$ such that $\tau \not\supseteq \sigma$, and
(2) all $\tau \cup \{v_\sigma\}$ where $\tau \in \Delta$, $\tau \not\supseteq \sigma$ and $\tau \cup \sigma \in \Delta$.

The use of stellar subdivision to describe order complexes of posets was begun in [22], where it was shown that the order complex of any distributive lattice can be obtained from a simplex by a sequence of stellar subdivisions. Although this result and some of its implications were discussed in [23], its proof was never published. We will give a generalization of this result to meet-distributive lattices in the next section. The proof is an adaptation of that in [22].

More recently, stellar subdivision was used in [9] to produce the order complex of a so-called Bier poset of a poset P from the order complex of P.

In Sect. 2, we describe the order complex of a meet-distributive lattice as a stellar subdivision of a simplex. We use this in Sect. 3 to construct the sphere associated with the lattice. Finally, in Sect. 4 we relate enumeration in this sphere to that of the lattice.

2 Order Complexes of Meet-Distributive Lattices

Let L be an arbitrary meet-distributive lattice. We can assume L is the lattice of closed sets of a convex closure $\langle\cdot\rangle$ on the set $[n]$, for some $n > 0$. L has unique maximal element $\hat{1} = \langle[n]\rangle$ and minimal element $\hat{0} = \langle\emptyset\rangle$ (we may assume $\langle\emptyset\rangle = \emptyset$, although this will not be important here). For simplicity of notation, we will write $\langle i\rangle$ for the *principal* closed set $\langle\{i\}\rangle$ whenever $i \in [n]$. These are precisely the *join-irreducible* elements of L, that is, those $x \in L \setminus \{\hat{0}\}$ that cannot be written as $y \vee z$, with $y, z < x$. (This follows, for example, from [13, Theorem 2.1(f)].)

In fact, the convex closure $\langle\cdot\rangle$ is uniquely defined from the lattice L: we take $[n]$ to be an enumeration of the join-irreducible elements of L and define, for $A \subseteq [n]$,

$$\langle A\rangle = \left\{ j \in [n] \,\middle|\, j \leq \bigvee_{i \in A} i \right\}.$$

Thus we are free, without loss of generality, to use the closure relation when making constructions concerning the lattice L.

Consider the simplex of all principal closed sets (join-irreducibles) $\{\langle i\rangle \mid i \in [n]\}$, and let Δ_0 be the simplicial complex consisting of this simplex and all its faces (subsets). Note that $\Delta(L \setminus \{\hat{0}\})$, the order complex of $L \setminus \{\hat{0}\}$, is a simplicial complex on the vertex set $V = \{\langle A\rangle \mid A \subseteq [n], A \neq \emptyset\}$.

Theorem 2.1 *For any meet-distributive lattice L, $\Delta(L \setminus \{\hat{0}\})$ can be obtained from the simplex of join-irreducible elements by a sequence of stellar subdivisions.*

Proof Suppose L is the lattice of closed sets of a convex closure $\langle\cdot\rangle$ on $[n]$. Let $A_1, A_2, \ldots, A_k$ be a reverse linear extension of $L \setminus \{\hat{0}\}$, that is, the A_i are all the nonempty closed subsets in $[n]$, ordered so that we never have $A_i \subseteq A_j$ if $i < j$. In particular, $A_1 = [n]$.

The order complex $\Delta(L \setminus \{\hat{0}\})$ can be obtained from Δ_0 by a sequence of stellar subdivisions as follows. For $i = 1, \ldots, k$, let

$$\Delta_i = sd_{\mathrm{ext}(A_i)}(\Delta_{i-1}),$$

where, by a slight abuse of notation, $\mathrm{ext}(A_i)$ will denote the face of Δ_{i-1} having vertices $\langle j\rangle$, $j \in \mathrm{ext}(A_i)$. The new vertex added at the ith step will be denoted simply by A_i. Note that because of the ordering of the A_i, the face $\mathrm{ext}(A_i)$ is in the complex Δ_{i-1}, so each of these subdivisions is defined.

We claim that $\Delta_k = \Delta(L \setminus \{\hat{0}\})$. The proof proceeds by induction on n. The case $n = 1$ is clear.

When $n > 1$, consider the complex $\Delta_1 = sd_{\mathrm{ext}(A_1)}(\Delta_0)$. Since Δ_0 is a simplex, the new vertex A_1 is a cone point, that is, it is in every maximal simplex of Δ_1. The base of this cone (the *link* of A_1) consists of all the facets $F_1, \ldots, F_m$ of Δ_0 that are opposite to vertices in $\mathrm{ext}(A_1)$. By relabeling if necessary, we can assume that $\mathrm{ext}(A_1) = \{1, 2, \ldots, m\}$, and $F_i = \{\langle j\rangle \mid j \neq i\}$. Since all further subdivisions are made on faces not containing A_1, the vertex A_1 remains a cone point in all Δ_i. So it is enough to consider the effect of further subdivisions on each of the facets F_i.

Now, by induction, the face F_i is subdivided so that it becomes the order complex of $L_i \setminus \{\hat{0}\}$, where L_i is the lattice of closed subsets of $[n] \setminus \{i\}$. Since A_1 is a cone point in Δ_k, and A_1 is in every maximal chain in L, it follows that Δ_k is the order complex of $L \setminus \{\hat{0}\}$. □

Notice that the stellar subdivisions over the principal closed sets (join-irreducibles) are redundant and can be omitted without loss. Figure 2 gives the sequence of subdivisions leading to the order complex of the meet-distributive lattice generated by the example of three collinear points. In this example, once Δ_3 is constructed, every subsequent subdivision is over a principal closed set and therefore has no effect on the complex.

By a *polyhedral ball* we will mean a simplicial complex that is topologically a d-dimensional ball and can be embedded to give a regular triangulation, that is, one that admits a strictly convex piecewise-linear function (see, for example, [4] for the definitions). Polyhedral balls are known to satisfy strong enumerative conditions [6].

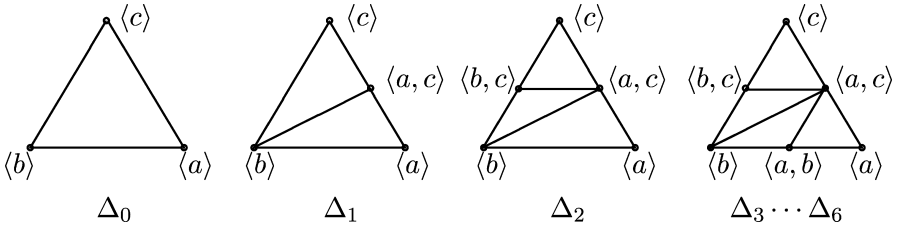


Fig. 2 The sequence of Δ_i for the example in Fig. 1

Corollary 2.2 *For any meet-distributive lattice $L \neq B_n$, the order complex $\Delta(L \setminus \{\hat{0}, \hat{1}\})$ is a polyhedral ball.*

Proof Let L be the lattice of closed sets of the convex closure $\langle\cdot\rangle$ on $[n]$. Since $L \neq B_n$, we have that $\mathrm{ext}([n]) \neq [n]$, and so every stellar subdivision that is involved in producing $\Delta(L \setminus \{\hat{0}\})$ takes place on the boundary of the simplex Δ_0.

Since being the boundary complex of a simplicial convex polytope is preserved under taking stellar subdivisions [23], we conclude that the boundary of $\Delta(L \setminus \{\hat{0}\})$ is the boundary of a simplicial convex polytope Q. By means of a projective transformation that sends the vertex $A_1 = [n] = \hat{1}$ to the point at infinity, we see that the image of Q under such a map is the graph of a strictly convex function over $\Delta(L \setminus \{\hat{0}, \hat{1}\})$. □

It was shown in [23] that stellar subdivision preserves the property of being *vertex decomposable*, which in turn implies shellability. As a consequence we get that both $\Delta(L \setminus \{\hat{0}\})$ and $\Delta(L \setminus \{\hat{0}, \hat{1}\})$ are vertex decomposable and hence shellable, as stated in [7, Theorem 8.1] (and its proof) in the language of greedoids. Theorem 2.1 was first proved for distributive lattices in [22] precisely to show that order complexes of distributive lattices were shellable. The result in [22] was stated for $\Delta(L)$, which is a cone over the complex we consider.

3 The Associated CW Spheres

We define now a triangulated sphere derived from the order complex of a meet-distributive lattice L. It will turn out that this triangulated sphere is the barycentric subdivision of a regular CW sphere that has the same enumerative relationship to L^* (the dual to L) as an arrangement of hyperplanes (oriented matroid) has to the underlying geometric lattice.

3.1 The complex $\pm\Delta$

For a meet-distributive lattice L, let $\Delta = \Delta(L \setminus \{\hat{0}\})$, a triangulation of the $(n-1)$-simplex Δ_0. We will define a triangulation $\pm\Delta$ of the n-dimensional crosspolytope O_n as follows. If the vertex set of the simplex Δ_0 is $[n] = \{1, 2, \ldots, n\}$, then that of the crosspolytope is $\pm[n] = \{\pm 1, \pm 2, \ldots, \pm n\}$. Faces of the crosspolytope are all $\sigma \subseteq \pm[n]$ such that not both i and $-i$ are in σ.

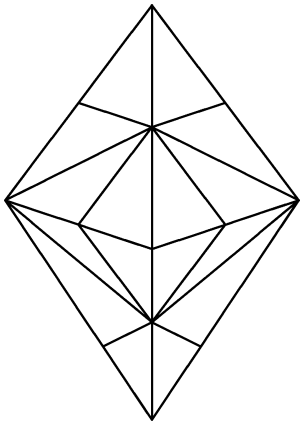


Fig. 3 Triangulation of the boundary of the octahedron induced by reflecting Δ_6

We reflect the triangulation Δ to obtain a triangulation $\pm\Delta$ of the crosspolytope, much as the crosspolytope can be built by reflecting the simplex generated by the unit vectors. We consider the simplex Δ_0 to be embedded as the convex hull of the unit vectors and define the triangulation $\pm\Delta$ by reflecting the triangulation Δ.

Formally, $\pm\Delta$ is the simplicial complex whose vertices are all equivalence classes of pairs (A, ε), where $A \in L \setminus \{\hat{0}\}$, ε is a map from $[n]$ to $\{\pm 1\}$, and we identify (A, ε) and (A, ε') when $\varepsilon|_{\text{ext}(A)} = \varepsilon'|_{\text{ext}(A)}$. For arbitrary $\varepsilon : [n] \to \{\pm 1\}$ and $\sigma = \{A_1, A_2, \ldots, A_k\} \in \Delta$, let

$$\sigma^{\varepsilon} := \{(A_1, \varepsilon), (A_2, \varepsilon), \ldots, (A_k, \varepsilon)\}$$

and

$$\Delta^{\varepsilon} = \{\sigma^{\varepsilon} | \sigma \in \Delta\}.$$

Δ^{ε} is essentially the triangulation Δ transferred to the face of the crosspolytope given by the sign pattern ε. Finally, define

$$\pm\Delta = \bigcup_{\varepsilon} \Delta^{\varepsilon},$$

the union being taken over all $\varepsilon : [n] \to \{\pm 1\}$.

Remark 3.1 Note that boundary faces of Δ can result in faces of $\pm\Delta$ having more than one name; in fact, $\sigma^{\varepsilon} = \sigma^{\rho}$ if and only if ε and ρ agree on the set

$$\text{ext}(\sigma) := \bigcup_{i=1}^{k} \text{ext}(A_i).$$

Figure 3 shows the complex $\pm\Delta$ for the example of three collinear points.

Theorem 3.2 *For any meet-distributive lattice L, $\pm\Delta(L \setminus \{\hat{0}\})$ can be obtained from the n-dimensional crosspolytope by a sequence of stellar subdivisions, and so it is combinatorially the boundary complex of an n-dimensional simplicial polytope, where n is the number of join-irreducibles of L.*

Proof As before, suppose L is the lattice of closed sets of a convex closure $\langle\cdot\rangle$ on $[n]$, and let $A_1, A_2, \ldots, A_k$ be a reverse linear extension of $L \setminus \{\hat{0}\}$. We can extend this order to an order on all pairs (A_i, ε), $\varepsilon : [n] \to \{\pm 1\}$, by ordering lexicographically, with the given order on the first coordinate and any order on the second.

It is now relatively straightforward to adapt the proof of Theorem 2.1 to show that the complex $\pm\Delta$ is obtained by carrying out stellar subdivisions over faces of O_n in the order given by the order of the (A_i, ε). The subdivision corresponding to (A_i, ε) is done over the face $\{\varepsilon(j) \cdot j \mid j \in \mathrm{ext}(A_i)\}$ of O_n; in $\pm\Delta$, every face containing $\{\varepsilon(j) \cdot j \mid j \in \mathrm{ext}(A_i)\}$ is subdivided as it would be by doing the subdivision in the boundary of O_n. Again, since stellar subdivision preserves the property of being the boundary complex of a polytope, the result follows. □

3.2 The poset Q_L

We construct a regular CW complex Σ having $\pm\Delta$ as its barycentric subdivision. Equivalently, if $\mathcal{F}(\Sigma)$ is the face poset of Σ, then $\Delta(\mathcal{F}(\Sigma)) = \pm\Delta$.

We begin by defining a poset Q_L associated to any meet-distributive lattice L. The elements of Q_L are all equivalence classes of pairs (A, ε), where $A \in L$ and ε is a map from $[n]$ to $\{\pm 1\}$ as before. We define the order relation on Q_L by $(A, \varepsilon) \leq (B, \delta)$ if and only if $A \subseteq B$ and the maps ε, δ agree on the set $\mathrm{ext}(A) \cap \mathrm{ext}(B)$. We include an element $\hat{1} \in Q_L$ for convenience; the element $(\hat{0}, \emptyset)$ corresponding to $\hat{0} \in L$ serves as $\hat{0}$ in Q_L. Note that when L is distributive, Q_L is the signed Birkhoff poset of [19].

Proposition 3.3 $\Delta(Q_L \setminus \{\hat{0}, \hat{1}\}) = \pm\Delta$.

Proof The maximal simplices in $\pm\Delta$ are the simplices

$$\sigma^{\varepsilon} := \{(A_1, \varepsilon), (A_2, \varepsilon), \ldots, (A_n, \varepsilon)\},$$

where $A_1 \subseteq A_2 \subseteq \cdots \subseteq A_n$ is a maximal chain in $L \setminus \{\hat{0}\}$. Then clearly,

$$(A_1, \varepsilon) < (A_2, \varepsilon) < \cdots < (A_n, \varepsilon)$$

is a maximal chain in $Q_L \setminus \{\hat{0}, \hat{1}\}$.

Conversely, if

$$(A_1, \varepsilon_1) < (A_2, \varepsilon_2) < \cdots < (A_n, \varepsilon_n)$$

is a maximal chain in $Q_L \setminus \{\hat{0}, \hat{1}\}$, then, if we let $\sigma = \{A_1, A_2, \ldots, A_n\} \in \Delta$, we have $\mathrm{ext}(\sigma) = [n]$ and so there is an $\varepsilon : [n] \to \{\pm 1\}$ such that $\varepsilon_i = \varepsilon|_{\mathrm{ext}(A_i)}$ for each i. Thus

$$\{(A_1, \varepsilon_1), (A_2, \varepsilon_2), \ldots, (A_n, \varepsilon_n)\} = \sigma^{\varepsilon}$$

is a maximal simplex in $\pm\Delta$. □

Next, we define a cell complex Σ_L from the lattice L (the underlying convex closure $\langle\cdot\rangle$ on $[n]$) and the simplicial complex $\pm\Delta$ as follows. For each $A \in L \setminus \{\hat{0}\}$ and $\varepsilon : [n] \to \{\pm 1\}$, we define a cell $C_{(A,\varepsilon)}$ that is a union of simplices in $\pm\Delta$. For

$A = [n]$, we take $C_{(A,\varepsilon)}$ to be the star of (A, ε) in the complex $\pm\Delta$, that is, the union of all maximal simplices containing the vertex (A, ε).

For proper closed sets $A \in L$, we consider the subgeometry $\langle\cdot\rangle$ restricted to subsets of A, with lattice $L_A = [\hat{0}, A]$ and order complex $\Delta_A = \Delta(L_A \setminus \{\hat{0}\})$. The complex Δ_A is the subcomplex of Δ subdividing the face of Δ_0 spanned by the vertices $i \in A$, and the corresponding complex $\pm\Delta_A$ is the subcomplex of $\pm\Delta$ subdividing the faces of the crosspolytope spanned by all vertices $\pm i$, $i \in A$. For any $A \in L$ and any $\varepsilon : [n] \to \{\pm 1\}$, we define the cell $C_{(A,\varepsilon)}$ to be the star of (A, ε) in the complex $\pm\Delta_A$.

Since $\pm\Delta_A$ is the boundary of a simplicial polytope by Theorem 3.2, each cell $C_{(A,\varepsilon)}$ is topologically a disk of dimension $|A| - 1$, and its boundary is the link of the vertex (A, ε) in the complex $\pm\Delta_A$ and so is a sphere. We define Σ_L to be the collection of all the cells $C_{(A,\varepsilon)}$, $A \in L \setminus \{\hat{0}\}$.

Lemma 3.4 *The boundary of $C_{(A,\varepsilon)}$ is the union of all cells $C_{(B,\delta)}$, where $B \subseteq A$, $B \neq A$ and the maps ε, δ agree on* $\mathrm{ext}(A) \cap \mathrm{ext}(B)$.

Proof By definition, we have

$$C_{(A,\varepsilon)} = \bigcup_{\substack{A \in \sigma \in \Delta \\ \gamma|_{\mathrm{ext}(A)} = \varepsilon|_{\mathrm{ext}(A)}}} \sigma^{\gamma}. \tag{3.1}$$

Since $\partial C_{(A,\varepsilon)}$ is the link of (A, ε) in $\pm\Delta_A$, that is,

$$\partial C_{(A,\varepsilon)} = \bigcup_{\substack{\tau \in lk_{\Delta_A}(A) \\ \gamma|_{\mathrm{ext}(A)} = \varepsilon|_{\mathrm{ext}(A)}}} \tau^{\gamma},$$

the statement of the lemma is equivalent to

$$\bigcup_{\substack{\tau \in lk_{\Delta_A}(A) \\ \gamma|_{\mathrm{ext}(A)} = \varepsilon|_{\mathrm{ext}(A)}}} \tau^{\gamma} = \bigcup_{\substack{B \subsetneq A \\ \delta|_{\mathrm{ext}(A)\cap\mathrm{ext}(B)} = \varepsilon|_{\mathrm{ext}(A)\cap\mathrm{ext}(B)}}} C_{(B,\delta)}. \tag{3.2}$$

Here the unions are over $\gamma : [n] \to \{\pm 1\}$ and $\delta : [n] \to \{\pm 1\}$, respectively, and $lk_{\Delta_A}(A) = \{\tau \in \Delta_A \mid A \notin \tau, \tau \cup \{A\} \in \Delta_A\}$ is the link of A in Δ_A.

To see the equality in (3.2), note that if τ^{γ}, $\tau \in lk_{\Delta_A}(A)$, $\gamma|_{\mathrm{ext}(A)} = \varepsilon|_{\mathrm{ext}(A)}$, appears on the left side, then $\tau^{\gamma} \subseteq C_{(B,\gamma)}$, where B is a maximal element of τ. Since $\gamma|_{\mathrm{ext}(A)\cap\mathrm{ext}(B)} = \varepsilon|_{\mathrm{ext}(A)\cap\mathrm{ext}(B)}$, the cell $C_{(B,\gamma|_{\mathrm{ext}(B)})}$ appears on the right side.

For the opposite inclusion, suppose τ^{γ} is a maximal simplex of $\pm\Delta_A$ in $C_{(B,\delta)}$, where $B \subsetneq A$ and $\delta|_{\mathrm{ext}(A)\cap\mathrm{ext}(B)} = \varepsilon|_{\mathrm{ext}(A)\cap\mathrm{ext}(B)}$. Then $\gamma|_{\mathrm{ext}(B)} = \delta|_{\mathrm{ext}(B)}$ by (3.1), and so

$$\gamma|_{\mathrm{ext}(A)\cap\mathrm{ext}(B)} = \delta|_{\mathrm{ext}(A)\cap\mathrm{ext}(B)} = \varepsilon|_{\mathrm{ext}(A)\cap\mathrm{ext}(B)}.$$

Since $i \in \mathrm{ext}(A) \cap B$ implies $i \in \mathrm{ext}(B)$ (otherwise $i \in \langle B \setminus \{i\}\rangle \subseteq \langle A \setminus \{i\}\rangle$), we have that the only places where $\gamma|\,\mathrm{ext}(A)$ and ε might not agree are outside of B. Since $\mathrm{ext}(\tau) \subseteq B$, we may, by Remark 3.1, adjust γ to γ' outside of B so that $\tau^{\gamma'} = \tau^{\gamma}$ and $\gamma'|_{\mathrm{ext}(A)} = \varepsilon|_{\mathrm{ext}(A)}$. Thus τ^{γ} appears on the left of (3.2), establishing the equality. □

We can now prove the main result of this section.

Theorem 3.5 *The cells in Σ_L form a regular CW sphere, with face poset $Q_L \setminus \{\hat{0}, \hat{1}\}$ and barycentric subdivision $\pm\Delta$.*

Proof Since each Δ^ε is a cone on $([n], \varepsilon)$,

$$|\Sigma_L| = \bigcup_{(A,\varepsilon)\in Q_L\setminus\{\hat{0},\hat{1}\}} C_{(A,\varepsilon)} = \bigcup_{\varepsilon:[n]\to\{\pm 1\}} C_{([n],\varepsilon)} = |\pm\Delta|,$$

so $|\Sigma_L|$ is a sphere by Theorem 3.2.

By construction, the only inclusions $C_{(B,\delta)} \subseteq C_{(A,\varepsilon)}$ possible among cells is when $C_{(B,\delta)} \subseteq \partial C_{(A,\varepsilon)}$, so $C_{(B,\delta)} \subseteq C_{(A,\varepsilon)}$ if and only if $(B, \delta) \leq (A, \varepsilon)$ in $Q_L \setminus \{\hat{0}, \hat{1}\}$ by Lemma 3.4.

To see that $|\Sigma_L|$ is a regular CW sphere, one can assemble $|\Sigma_L|$ according to a linear extension of the poset $Q_L \setminus \{\hat{0}, \hat{1}\}$. By Lemma 3.4, all the boundary faces of any cell $C_{(A,\varepsilon)}$ will be present when it comes time to attach it.

Since the poset of inclusions among the faces of Σ_L is $Q_L \setminus \{\hat{0}, \hat{1}\}$, it will have $\pm\Delta$ as barycentric subdivision by Proposition 3.3. □

3.3 Join-distributive lattices

We note briefly that everything in this section works for a *join-distributive lattices*, that is, a lattice L whose dual L^* (reverse all order relations) is meet-distributive. Here we have to reverse the roles of $\hat{0}$ and $\hat{1}$. In particular, both L and L^* have the same order complex, so we have $\Delta(L \setminus \{\hat{1}\}) = \Delta(L^* \setminus \{\hat{0}\}) = \Delta$, which gives rise to the same simplicial polytope $\pm\Delta$.

For join-distributive L, the poset $Q_L = (Q_{L^*})^*$, and so the corresponding spherical complex Σ_L is defined by defining the maximal cells to correspond to the maximal elements of $Q_L \setminus \{\hat{1}\}$ (the minimal elements of $Q_{L^*} \setminus \{\hat{0}\}$). Here, the CW sphere $\Sigma_L = (\Sigma_{L^*})^*$ is the dual to Σ_{L^*}.

Figure 4 shows the CW sphere for both the meet-distributive L from three collinear points and the corresponding join-distributive L^*. Note that both $\pm\Delta$ and Σ_L retain the full $(\mathbb{Z}/2\mathbb{Z})^n$ symmetry of the crosspolytope.

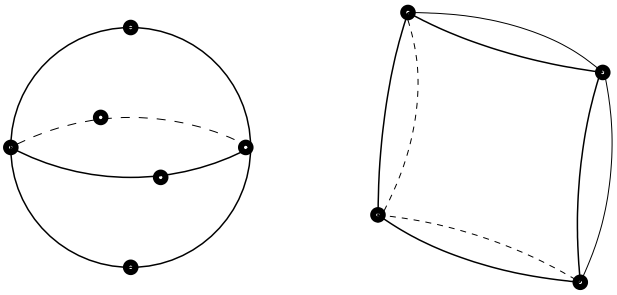


Fig. 4 The spheres Σ_L and Σ_{L^*} for L from three collinear points

We remark that if L is the lattice of the convex geometry on n collinear points, then one can verify that Q_{L^*} is isomorphic to the Tchebyshev poset T_n of [18]. Hetyei showed that T_n is the face poset of a regular CW sphere and its order complex subdivides a crosspolytope. In fact his proof of the latter assertion uses essentially the reflection construction discussed at the beginning of Sect. 3.1.

4 Enumerative Properties of Q_L

For a graded poset P (with $\hat{0}$ and $\hat{1}$) with rank function ρ, define

$$\nu(P) = \sum_{t \in P} (-1)^{\rho(t)} \mu(\hat{0}, t), \tag{4.1}$$

where μ denotes the Möbius function as defined in [24, Chap. 3]. If L is the intersection lattice of a real hyperplane arrangement, a well known result due to Zaslavsky [29] gives $\nu(L)$ as the number of connected components in the complement of the arrangement. He extended this to show how all the face numbers of an arrangement depend solely on the lattice of intersections. As a generalization, [8, Proposition 4.6.2] expresses the flag numbers of an arrangement, that is, the enumerators of chains of faces having prescribed rank sets, in terms of the functional ν applied to intervals in the intersection lattice.

We now show that for join-distributive L, the flag numbers of Q_L may be computed similarly from intervals in L. (For meet-distributive L, the flag numbers of Q_L can be obtained from this by duality.) Suppose that L consists of the closed sets of a convex geometry ordered by reverse inclusion. In analogy with the zero map on oriented matroids [8], we define the map $z : Q_L \setminus \{\hat{0}\} \to L$ by $z((A, \varepsilon)) = A$.

Proposition 4.1 *Let $c = \{A_1 < A_2 < \cdots < A_k = \hat{1}\}$ be a chain in the join distributive lattice L and $z^{-1}(c)$ denote the set of chains in Q_L that are mapped by z to c. Then*

$$|z^{-1}(c)| = \prod_{i=1}^{k-1} \nu([A_i, A_{i+1}]).$$

Proof Given a sign function $\varepsilon_i : [n] \to \{\pm 1\}$ for some $2 \le i \le k$, there are $2^{|\operatorname{ext}(A_{i-1}) \setminus \operatorname{ext}(A_i)|}$ essentially different sign functions $\varepsilon_{i-1} : [n] \to \{\pm 1\}$ such that $(A_{i-1}, \varepsilon_{i-1}) < (A_i, \varepsilon_i)$ in Q_L, since the only restriction on ε_{i-1} is that it agree with ε_i on $\operatorname{ext}(A_i) \cap \operatorname{ext}(A_{i-1})$. Thus, starting with $\varepsilon_k = \emptyset$, there are precisely $\prod_{i=1}^{k-1} 2^{|\operatorname{ext}(A_i) \setminus \operatorname{ext}(A_{i+1})|}$ ways to build a sequence of sign functions $\varepsilon_k, \ldots, \varepsilon_2, \varepsilon_1$ resulting in a chain $(A_1, \varepsilon_1) < \cdots < (A_k, \varepsilon_k)$ in Q_L.

To complete the proof it suffices to show that for $1 \le i \le k$,

$$\sum_{A_i \le B \le A_{i+1}} (-1)^{\rho(A_i, B)} \mu(A_i, B) = 2^{|\operatorname{ext}(A_i) \setminus \operatorname{ext}(A_{i+1})|}. \tag{4.2}$$

The Möbius function of a join-distributive lattice satisfies

$$(-1)^{\rho(A_i, B)} \mu(A_i, B) = \begin{cases} 1 & \text{if } [A_i, B] \text{ is a Boolean lattice,} \\ 0 & \text{otherwise.} \end{cases}$$

(This follows, for example, from [13, Theorems 4.2, 4.3].) By definition of join-distributivity, for $B \in [A_i, A_{i+1}]$ the interval $[A_i, B]$ is a Boolean lattice precisely when B is less than or equal to the join of atoms of $[A_i, A_{i+1}]$, which are those $A_i \setminus \{a\}$ such that $a \in \mathrm{ext}(A_i) \setminus \mathrm{ext}(A_{i+1})$. Hence the left side of (4.2) reduces to a sum of the form $\sum_B 1$ with B ranging over a Boolean lattice of rank $|\mathrm{ext}(A_i) \setminus \mathrm{ext}(A_{i+1})|$. □

The complete enumerative information on chains in a graded poset P is carried by the formal power series

$$F_P := \sum_{\substack{\hat{0}=t_0<t_1<\cdots<t_k=\hat{1} \\ 0<i_1<\cdots<i_k}} x_{i_1}^{\rho(t_0,t_1)} x_{i_2}^{\rho(t_1,t_2)} \cdots x_{i_k}^{\rho(t_{k-1},t_k)},$$

where $\rho(s,t) = \rho(t) - \rho(s)$. As P ranges over the family of graded posets, the F_P span the (Hopf) algebra of quasisymmetric functions, denoted $\mathcal{Q}$. The definition of F_P is due to Ehrenborg [16]. See [25, Sect. 7.19] for further background on quasisymmetric functions.

In the context of combinatorial Hopf algebras [1], the functional ν can be seen as the pullback of a certain "odd character" $\nu_{\mathcal{Q}}$ to the Hopf algebra of graded posets along the map $P \mapsto F_P$; that is, $\nu(P) = \nu_{\mathcal{Q}}(F_P)$. By the general theory there is an induced Hopf algebra map $\vartheta : \mathcal{Q} \to \mathcal{Q}$ satisfying

$$\vartheta(F_P) = \sum_{\substack{\hat{0}=t_0<t_1<\cdots<t_k=\hat{1} \\ 0<i_1<\cdots<i_k}} \nu([t_0,t_1]) \cdots \nu([t_{k-1},t_k]) x_{i_1}^{\rho(t_0,t_1)} \cdots x_{i_k}^{\rho(t_{k-1},t_k)}.$$

In fact ϑ is precisely the map introduced by Stembridge [26] to relate the quasisymmetric weight enumerator for P-partitions of a labeled poset to the enriched quasisymmetric weight enumerator of that poset. See [1, Examples 2.2, 4.4, 4.9].

The main result of this section is an extension of [19, Theorem 5.15]:

Theorem 4.2 *For a join-distributive lattice L, we have*

$$2F_{Q_L} = \vartheta(F_{L \cup \{\hat{0}\}}),$$

where $\hat{0}$ denotes a new minimum element adjoined to L.

Proof This is essentially the argument used to prove [3, Theorem 3.1]. Extend z to a map $z : Q_L \to L \cup \{\hat{0}\}$ by requiring $z(\hat{0}) = \hat{0}$. By Proposition 4.1,

$$\begin{aligned} F_{Q_L} &= \sum_{\substack{c=\{\hat{0}=A_0<\cdots<A_k=\hat{1}\} \subseteq L \cup \{\hat{0}\} \\ i_1<\cdots<i_k}} |z^{-1}(c)| x_{i_1}^{\rho(A_0,A_1)} \cdots x_{i_k}^{\rho(A_{k-1},A_k)} \\ &= \sum_{\substack{\hat{0}=A_0<\cdots<A_k=\hat{1} \\ i_1<\cdots<i_k}} \nu([A_1,A_2]) \cdots \nu([A_{k-1},A_k]) x_{i_1}^{\rho(A_0,A_1)} \cdots x_{i_k}^{\rho(A_{k-1},A_k)} \end{aligned}$$

$$= \frac{1}{2} \sum_{\substack{\hat{0}=A_0<\cdots<A_k=\hat{1} \\ i_1<\cdots<i_k}} \nu([A_0, A_1]) \cdots \nu([A_{k-1}, A_k]) x_{i_1}^{\rho(A_0,A_1)} \cdots x_{i_k}^{\rho(A_{k-1},A_k)}$$

$$= \frac{1}{2} \vartheta(F_{L \cup \{\hat{0}\}}).$$

The third equality holds because $\hat{0}$ is covered by only one element in $L \cup \{\hat{0}\}$, implying that $\mu(\hat{0}, A_1)$ vanishes if $\rho(A_1) > 1$; hence $\nu([\hat{0}, A_1]) = \mu(\hat{0}, \hat{0}) - \mu(\hat{0}, [n]) = 2$. □

Analogously, if Z is the face lattice of the zonotope associated with a hyperplane arrangement and L is the intersection lattice of the arrangement, then

$$2F_Z = \vartheta(F_{L \cup \{\hat{0}\}})$$

[3, 5, Proposition 3.5]. It is easy to see that a join-distributive lattice L must be semimodular, as are geometric lattices. One is led to speculate whether this relationship holds for *all* semimodular lattices, namely, whether for any semimodular lattice L, there exists a regular CW sphere Σ_L with face poset Q_L (with $\hat{0}$, $\hat{1}$ adjoined) such that

$$2F_{Q_L} = \vartheta(F_{L \cup \{\hat{0}\}}).$$

The role played by convex closures in this work might be played instead by *interval greedoids* (see [10, Theorem 8.8.7]; we are grateful to Anders Björner for suggesting this connection). Note that this would imply the existence of spheres Σ_L for geometric lattices that are not necessarily orientable. In nonorientable case, one might also ask for the relationship of $\vartheta(F_{L \cup \{\hat{0}\}})$ to the (dual) face counts of the homotopy-sphere arrangements of Swartz [28]. Simple examples suggest the former might provide lower bounds for the latter. In the orientable case, these bounds are clearly achieved by the results of [3, 5]. One could speculate further that achieving the bounds implies orientability.

There is a well-known bijection between multichains of a fixed length in a distributive lattice $J(P)$ and P-partitions (order-preserving maps) whose parts have a certain fixed upper bound [24, Proposition 3.5.1]. Edelman and Jamison [13, Theorem 4.7] extended this to a bijection between multichains in a meet-distributive lattice and extremal functions, which are generalizations of P-partitions. We will conclude with a discussion of an analogous correspondence between multichains in Q_L and a new class of functions called *enriched extremal functions*, which are generalizations of enriched P-partitions [26]. (We are grateful to Paul Edelman for suggesting that we seek such a correspondence.)

Let L be the lattice of closed sets of a convex closure $\langle\cdot\rangle$ on the set $[n]$. Consider the linear ordering $-1 \prec 1 \prec -2 \prec 2 \prec \cdots$ of the set of non-zero integers $\mathbb{Z} \setminus \{0\}$. For a function $f : [n] \to \mathbb{Z} \setminus \{0\}$ and a closed set A, let f_A denote the minimum element of $\{f(a) \mid a \in A\}$ with respect to $\prec$.

Definition 4.3 Given a convex closure on $[n]$, a function $f : [n] \to \mathbb{Z} \setminus \{0\}$ is called *enriched extremal function* provided that

(1) For every closed set A there exists $a \in \mathrm{ext}(A)$ such that $f(a) = f_A$, and

(2) For every $a \in [n]$, if $f(a) < 0$ then $a \in \text{ext}\{b \in [n] \mid f(b) \succeq f(a)\}$.

We remark that if f satisfies condition (1) then $\{a \in [n] \mid f(a) \succeq b\}$ is closed for any $b \in \mathbb{Z} \setminus \{0\}$. This justifies the notation used in (2).

For the convex closure on three collinear points there are four enriched extremal functions $f : \{a, b, c\} \to \{-1, 1\}$. They are given by $(f(a), f(b), f(c)) = (1, 1, 1), (-1, 1, 1), (1, 1, -1)$, and $(-1, 1, -1)$.

Notice that if $\langle\cdot\rangle$ is the upper-order-ideal closure on a poset $P = ([n], \leq_P)$, then f is an enriched extremal function if and only if for all $a <_P b$, we have (1) $f(a) \preceq f(b)$, and (2) $f(a) = f(b)$ implies $f(a) > 0$; in other words f is an *enriched P-partition* with respect to a natural labeling of P [26].

The zeta polynomial of a graded poset Q, denoted $Z(Q, t)$, is determined by the property that for a positive integer m, $Z(Q, m)$ is the number of multichains in Q_L of the form $\hat{0} = q_0 \leq q_1 \leq \cdots \leq q_m = \hat{1}$. It will be convenient to introduce another polynomial $\overline{Z}(Q, t)$ given by

$$\overline{Z}(Q, t) = \sum_{q \text{ maximal in } Q \setminus \{\hat{1}\}} Z([\hat{0}, q], t).$$

Thus $\overline{Z}(Q, m)$ is the number of multichains in Q of the form $\hat{0} = q_0 \leq \cdots \leq q_m$, where q_m is a maximal element in $Q \setminus \{\hat{1}\}$.

For a positive integer m, recall $\pm[m] = \{-m, \ldots, -2, -1, 1, 2, \ldots, m\}$.

Proposition 4.4

Suppose that L is the meet-distributive lattice of closed sets of a convex closure $\langle\cdot\rangle$ on $[n]$. Then for $m \geq 1$,

$$\overline{Z}(Q_L, m) = \#\textit{ of enriched extremal functions } f : [n] \to \pm[m].$$

Sketch of proof Given a multichain $([n], \varepsilon_0) = (A_0, \varepsilon_0) \geq (A_1, \varepsilon_1) \geq \cdots \geq (A_m, \varepsilon_m) = (\emptyset, \emptyset)$ in $Q_L \setminus \{\hat{1}\}$, we obtain an enriched extremal function $f : [n] \to \{\pm 1, \ldots, \pm m\}$ by setting, for $a \in A_{i-1} \setminus A_i$,

$$f(a) = \begin{cases} -i & \text{if } a \in \text{ext}(A_{i-1}) \text{ and } \varepsilon_{i-1}(a) = -1, \\ i & \text{otherwise.} \end{cases}$$

Conversely, for an enriched extremal function $f : [n] \to \pm[m]$, if we write $\{|f(1)|, |f(2)|, \ldots, |f(n)|\} = \{s_1 < s_2 < \cdots < s_k\}$ then we can recover the corresponding multichain by setting $A_i = \{a \in [n] \mid f(a) \succeq -s_j\}$ for $1 \leq j \leq k$ and $s_{j-1} \leq i < s_j$, where $s_0 = 0$, and setting $\varepsilon_i(a)$ equal to the sign of $f(a)$. □

To obtain a similar formula for the zeta polynomial $Z(Q_L, m)$, we extend $\langle\cdot\rangle$ to a convex closure on $[n+1]$ by declaring that $\langle n+1\rangle = [n+1]$. If L' denotes the lattice of closed sets for the new closure then it is easy to see that $\overline{Z}(Q_{L'}, m) = 2\, Z(Q_L, m)$. It follows that

$$2\, Z(Q_L, m) = \#\textit{ of enriched extremal functions } f : [n+1] \to \pm[m].$$

Proposition 4.4 may be viewed as the enriched analogue of [13, Theorem 4.7], which asserts that $Z(L, m)$ enumerates certain extremal functions and that by reciprocity $(-1)^n Z(L, -m)$ enumerates *strictly* extremal functions. It turns out that for Q_L we have self-reciprocity, that is,

$$Z(Q_L, -t) = (-1)^{n+1} Z(Q_L, t)$$

and

$$\overline{Z}(Q_L, -t) = (-1)^n \overline{Z}(Q_L, t)$$

for any meet-distributive lattice L (cf. [26, Proposition 4.2]). This follows, for example, from [24, Proposition 3.14.1] and the fact that Q_L is an Eulerian poset of rank $n + 1$.

A further corollary of Proposition 4.4 is that if L is the lattice of upper order ideals of a poset P and P_0 denotes the poset obtained from P by adjoining a new minimum element, then

$$2\, Z(Q_L, m) = \#\, \textit{of enriched } P_0\textit{-partition } f : P_0 \to \pm[m].$$

(See also [19, Corollary 5.3].) As an application we describe a way to translate the recent counterexamples of Stembridge's enriched poset conjecture [27] into new counterexamples of Gal's real root conjecture for flag triangulations of spheres [17].

It can be shown (e.g., [24, Chap. 3, Exercise 67b]) that the generating function for $Z(Q_L, m)$ satisfies

$$\sum_{m \geq 0} Z(Q_L, m) t^m = \frac{t \cdot h^{\Delta(Q_L \setminus \{\hat{0}, \hat{1}\})}(t)}{(1-t)^{n+1}},$$

where $h^{\Delta(Q_L \setminus \{\hat{0}, \hat{1}\})}(t)$ denotes the *h-polynomial* of the order complex $\Delta(Q_L \setminus \{\hat{0}, \hat{1}\})$. Stembridge found examples of a poset P such that the numerator of the rational generating function enumerating enriched P_0-partitions has non-real roots, thereby disproving an earlier conjecture of his [26]. From such a poset one can construct via our results a flag simplicial complex (meaning every minimal non-face has size two)—namely $\Delta(Q_L \setminus \{\hat{0}, \hat{1}\})$, where L is the lattice of upper order ideals of P—that barycentrically subdivides a regular CW sphere and whose h-polynomial has non-real roots. In fact, this simplicial sphere will be a simplicial polytope. We have thus provided additional counterexamples of Gal's conjecture that the h-polynomial of a flag simplicial triangulation of a sphere should have only real roots [17].

References

1. Aguiar, M., Bergeron, N., Sottile, F.: Combinatorial Hopf algebras and generalized Dehn–Sommerville relations. Compos. Math. **142**, 1–30 (2006)
2. Bayer, M., Sturmfels, B.: Lawrence polytopes. Can. J. Math. **42**, 62–79 (1990)
3. Billera, L.J., Ehrenborg, R., Readdy, M.: The **c**-2**d**-index of oriented matroids. J. Comb. Theory Ser. A **80**, 79–105 (1997)
4. Billera, L.J., Filliman, P., Sturmfels, B.: Constructions and complexity of secondary polytopes. Adv. Math. **83**, 155–179 (1990)

5. Billera, L.J., Hsiao, S.K., van Willigenburg, S.: Peak quasisymmetric functions and Eulerian enumeration. Adv. Math. **176**, 248–276 (2003)
6. Billera, L.J., Lee, C.W.: The numbers of faces of polytope pairs and unbounded polyhedra. Eur. J. Comb. **2**, 307–322 (1981)
7. Björner, A., Korte, B., Lovász, L.L.: Homotopy properties of greedoids. Adv. Appl. Math. **6**, 447–494 (1985)
8. Björner, A., Las Vergnas, M., Sturmfels, B., White, N., Ziegler, G.M.: Oriented Matroids. Cambridge University Press, New York (1993)
9. Björner, A., Paffenholz, A., Sjöstrand, J., Ziegler, G.: Bier spheres and posets. Discret. Comput. Geom. **34**, 71–86 (2005)
10. Björner, A., Ziegler, G.: Introduction to greedoids. In: White, N. (ed.) Matroid Applications, pp. 284–357. Cambridge University Press, Cambridge (1992)
11. Dilworth, R.P.: Lattices with unique irreducible decompositions. Ann. Math. **41**(2), 771–777 (1940)
12. Edelman, P.H.: Meet-distributive lattices and the anti-exchange closure. Algebra Universalis **10**, 290–299 (1980)
13. Edelman, P.H., Jamison, R.E.: The theory of convex geometries. Geometriae Dedicata **19**, 247–270 (1985)
14. Edelman, P.H.: Abstract convexity and meet-distributive lattices. In: Rival, I. (ed.) Combinatorics and Ordered Sets (Arcata, 1985). Contemporary Mathematics, vol. 57, pp. 127–150. American Mathematical Society, Providence (1986)
15. Edelman, P.H., Saks, M.E.: Combinatorial Representation and convex dimension of convex geometries. Order **5**, 23–32 (1988)
16. Ehrenborg, R.: On posets and Hopf algebras. Adv. Math. **119**(1), 1–25 (1996)
17. Gal, S.: Real root conjecture fails for five- and higher-dimensional spheres. Discret. Comput. Geom. **34**, 269–284 (2005)
18. Hetyei, G.: Tchebyshev posets. Discret. Comput. Geom. **32**, 493–520 (2004)
19. Hsiao, S.K.: A signed analog of the Birkhoff transform. J. Comb. Theory Ser. A **113**, 251–272 (2006)
20. Korte, B., Lovász, L.L., Schrader, R.: Greedoids. Springer, Berlin (1991)
21. Monjardet, B.: A use for frequently rediscovering a concept. Order **1**, 415–417 (1985)
22. Provan, J.S.: Decompositions, shellings, and diameters of simplicial complexes and convex polyhedra. Ph.D. Dissertation, Cornell University, Ithaca (1977)
23. Provan, J.S., Billera, L.J.: Decompositions of simplicial complexes related to diameters of convex polyhedra. Math. Oper. Res. **5**, 576–594 (1980)
24. Stanley, R.: Enumerative Combinatorics, vol. 1. Cambridge Studies in Advanced Mathematics, vol. 49. Cambridge University Press, Cambridge (1997)
25. Stanley, R.: Enumerative Combinatorics, vol. 2. Cambridge Studies in Advanced Mathematics, vol. 62. Cambridge University Press, Cambridge (1999)
26. Stembridge, J.R.: Enriched P-partitions. Trans. Am. Math. Soc. **349**(2), 763–788 (1997)
27. Stembridge, J.R.: Counterexamples to the poset conjectures of Neggers, Stanley, and Stembridge. Trans. Am. Math. Soc. **359**, 1115–1128 (2007)
28. Swartz, E.: Topological representations of matroids. J. Am. Math. Soc. **16**, 427–442 (2003)
29. Zaslavsky, T.: Facing up to arrangements: face count formulas for partitions of space by hyperplanes. Mem. Am. Math. Soc. **1**(154) (1975)

Isotopic Implicit Surface Meshing

Jean-Daniel Boissonnat · David Cohen-Steiner · Gert Vegter

Abstract This paper addresses the problem of piecewise linear approximation of implicit surfaces. We first give a criterion ensuring that the zero-set of a smooth function and the one of a piecewise linear approximation of it are isotopic. Then, we deduce from this criterion an implicit surface meshing algorithm certifying that the output mesh is isotopic to the actual implicit surface. This is the first algorithm achieving this goal in a provably correct way.

Keywords Topology · Triangulations · Morse theory · Algorithms

1 Introduction

Implicit equations are a popular way to encode geometric objects; See, e.g., [4, 25]. Typical examples are CSG models, where objects are defined as results of boolean operations on simple geometric primitives. Given an implicit surface, associated geometric objects of interest, such as contour generators, are also defined by implicit equations. Another advantage of implicit representations is that they allow for efficient blending of surfaces, with obvious applications in CAD or metamorphosis. Finally, this type of representation is also relevant to other scientific fields, such as level set methods or density estimation [8].

However, most graphical algorithms, and especially those implemented in hardware, cannot process implicit surfaces directly, and require that a piecewise linear approximation of the considered surface has been computed beforehand. As a consequence, polygonization of implicit surfaces has been widely studied in the literature.

J.-D. Boissonnat · D. Cohen-Steiner (✉)
Projet Géométrica, INRIA Sophia-Antipolis, Nice, France
e-mail: david.cohen-steiner@sophia.inria.fr

G. Vegter
Institute for Mathematics and Computing Science, RUG, Amsterdam, Netherlands

There are two general classes of methods devoted to this problem: continuation methods and adaptive enumeration methods. A *continuation algorithm* is surface based in the sense that it starts from a seed point on the surface, and computes successive vertices of the mesh while following the surface in some tangent direction. None of the algorithms in this category comes with topological guarantees: they might miss some connected components, or merge different components into a single one. *Adaptive enumeration methods*, also called *extrinsic polygonization methods* [25], are grid based, or, more generally, based on a tessellation of the ambient 3D space. They consist of two steps: first build a tessellation of space, and then analyze the intersection of the considered surface with each cell of the tessellation to construct the approximation. The celebrated marching cube algorithm [16] belongs to this category. The goal of an implicit surface polygonizer is twofold: its output should be geometrically close to the original surface, and have the same topology. While the former is achieved by several polygonization schemes [26], the latter has been barely addressed up to now.

Some algorithms achieve topological consistency, that is, ensure that the result is indeed a manifold, by taking more or less arbitrary decisions when a topologically ambiguous configuration is encountered. This implies that their output might have a topology different from the one of the original surface, except in very specific cases [15]. The problem of topologically correct polygonization of implicit curves in the plane is treated by Snyder in [24], who uses an adaptive enumeration method. His algorithm combines interval arithmetic with a quadtree tessellation of the domain of interest. It seems hard to generalize this method to implicit surfaces in three-space. Moreover, this algorithm seems to have high complexity due to the large number of calls to the interval version of Newton's method.

When the conference version of the present paper was published (Proceedings of STOC'04), there was only one paper devoted to the problem of homeomorphic polygonization of surfaces [19]. Since then there has been several papers [5, 7, 18] that solve the same problem as ours, or a related one. The main theoretical tool used in [19] is Morse theory. The authors first find a level set of the considered function that can be easily polygonized. This initial polygonization is then progressively transformed into the desired one, by computing intermediate level sets. This requires in particular to perform topological changes when critical points are encountered. This algorithm has an intuitive justification and seems to work on simple cases. Unfortunately, the authors do not give any proof of its correctness, and it is not clear to us whether it can deal with complex shapes in a robust way. In particular, the method does not guarantee that the mesh produced are self-intersection free.

In this paper, we give the first certified algorithm for the more difficult problem of *isotopic* implicit surface polygonization. This means that our output can be continuously deformed into the actual implicit surface without introducing self-intersections [14]. For instance, if the original implicit surface is knotted, then our output is guaranteed to be knotted in the same way, which would not be guaranteed by an algorithm ensuring only homeomorphic polygonization. Moreover, the whole algorithm can be implemented in the setting of interval analysis. We only assume that the considered isosurface is smooth, that is, does not contain any critical point. By Sard's theorem [22], this is a generic condition. Our polygonization is the zero-set of the linear interpolation of the implicit function on a mesh of $\mathbb{R}^3$. We first exhibit a set

of conditions on the mesh used for interpolation that ensure the topological correctness (Sect. 2). Then, we describe an algorithm for building a mesh satisfying these conditions, thereby leading to a provably correct isotopic polygonization algorithm (Sect. 3).

We note that since the publication of the conference version of the present paper, another method appeared that solves exactly the same problem as ours [18]. One difference between the two methods is that [18] uses octrees instead of triangulations. A more important difference is in the refinement stopping criterion: in [18], cells are subdivided until the intersection of the implicit surface with each cell is sufficiently flat. By contrast, we stop refinement as soon as a certain global criterion ensuring topological correctness is met. Hence, we may expect that our method is faster than [18]. This remains to be proved though, since we did not implement our method.

2 A Condition for Isotopic Meshing

Let f be a C^2 function from $\mathbb{R}^3$ to $\mathbb{R}$, and M be its zero-set. We assume that M, the surface we want to polygonize, is compact (condition a1). In what follows, T denotes a triangulation of a domain $\Omega \subset \mathbb{R}^3$ containing M and $\hat{f}$ the function that coincides with f at the vertices of T and that is linearly interpolated on the simplices of T. A vertex v will be said to be *larger* (resp. *smaller*) than a vertex u if $f(v)$ is *larger* (resp. *smaller*) than $f(u)$; the sign of f at a vertex will be referred to as the sign of that vertex. We set $\hat{M} = \hat{f}^{-1}(0)$.

2.1 Topological Background

Collapses Loosely speaking, a collapse [20] is an operation which consists of removing cells from a simplicial complex without changing its connectivity. More precisely:

Definition 1 If L is a simplicial complex and K a subcomplex of L, one says that there is an elementary collapse from L to K if there is a p-simplex s of L and a $(p-1)$-face t of s such that:

- s is not a face of any simplex of L.
- t is not a face of any simplex of L other than s.
- L is the union of K, s, and all the faces of s.
- $\partial s \setminus K$ is the relative interior of t.

Definition 2 If L is a simplicial complex and K a subset of L, one says that L collapses to K if there is a subdivision L' of L such that a subdivision of K can be obtained from L' by a sequence of elementary collapses.

Definition 2 is illustrated in Fig. 2. In Fig. 2, the complexes in the middle and on the right do not collapse to the bold curve because they would need to be "torn" in order to do so.

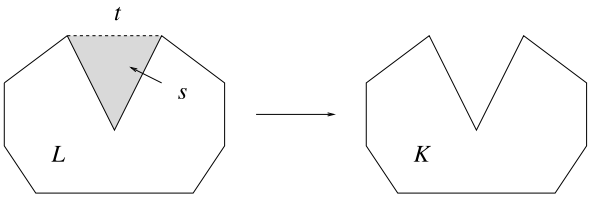


Fig. 1 Elementary collapse

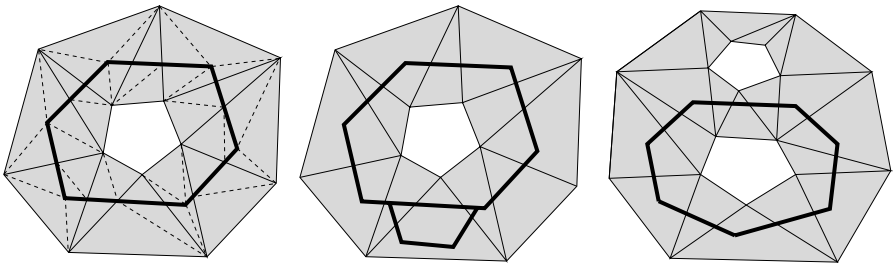


Fig. 2 The *grey complex L* on the *left collapses* to the *bold curve K* (*dashed edges* represent the subdivision L'). This is not true for the two other complexes

Smooth Morse theory The topology of implicit surfaces is usually investigated through Morse theory [17]. Given a real function f defined on a manifold, Morse theory studies the topological changes in the sets $f^{-1}(]-\infty, a])$ (*lower level-sets*) when a varies. In our case, as f is defined on $\mathbb{R}^3$, this amounts to studying how the topology of the part of the graph of f lying below a horizontal hyperplane changes as this hyperplane sweeps $\mathbb{R}^4$. Classical Morse theory assumes that f is of class C^2. In this case, as is well known, these topological changes are related to the *critical points* of f, that is, the points where the gradient ∇f of f vanishes. More precisely, the only topological changes occur when $f^{-1}(a)$ passes through a critical point p. The value a is then called a *critical value*. Generically, in the 2-dimensional case, the topology of $f^{-1}(]-\infty, a])$ can change in three possible ways, according to the type of the critical point p (see Fig. 3).

In Fig. 3, the sets $f^{-1}(]-\infty, a])$ are displayed as light grey regions. The leftmost column depicts the situation where p is a local maximum, that is, when the Hessian of f at p is positive. In this case, $f^{-1}(]-\infty, a+\varepsilon])$ is obtained from $f^{-1}(]-\infty, a-\varepsilon])$ by gluing a topological disk along its boundary. In the case of a saddle point (i.e. the Hessian has critical values of both signs), passing a critical value amounts to gluing a thickened topological line segment (in grey) along its "thickened" boundary (in bold). Finally, passing through a local minimum (negative Hessian) just amounts to adding a disk disconnected from $f^{-1}(]-\infty, a-\varepsilon])$. If p does not fall in any of these categories, that is, if the Hessian at p is degenerate, then classical Morse theory cannot be applied. C^2 functions the critical points of which all have non-degenerate Hessian are called *Morse functions*. From now on, we will assume that f is a Morse function (condition a2). Also, we require that 0 is not a critical value of f (condition a3), which implies that M is a manifold.

The number n of negative eigenvalues of the Hessian at p is classically called the index of p. However, for consistency reasons that will appear later, we call the *index*

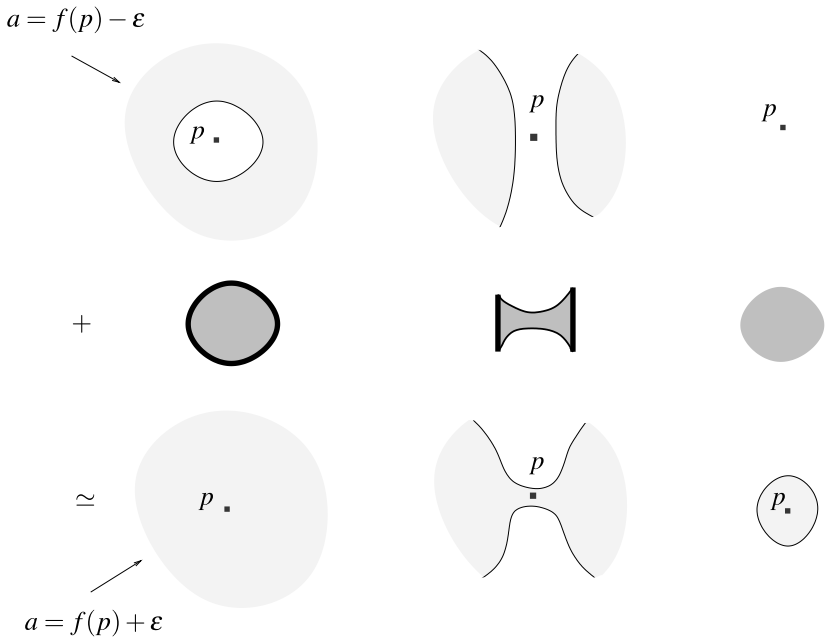


Fig. 3 Smooth Morse theory in 2D

of p the integer $(-1)^n$. The *index* of f on a region V is the sum of the indices of all critical points of f lying in V. The index satisfies the following important theorem:

Theorem 1 (Poincaré-Hopf index theorem) *The index of f on one of its lower level-sets is the Euler characteristic of that lower level-set.*

PL Morse theory Morse theory has been extended to a broad class of non-smooth functions by Goresky and McPherson [11]. We now outline the special case of PL functions, that is, we consider the case of $\hat{f}$. We assume from now on that no two neighboring vertices map to the same value under f, and that no vertex of T maps to 0 under f (conditions b1 and b2), which guarantees that $\hat{M}$ is a manifold. We refer to these assumptions as *genericity assumptions*. Let us first recall some well-known definitions [10, 11]:

Definition 1 The *star* of a vertex is the union of all simplices[1] containing this vertex. The *link* of a vertex is the boundary of its star.

Definition 2 The *lower star* $St^-(v)$ of $\hat{f}$ at a vertex v is the union of all simplices incident on v whose vertices other than v are smaller than v. The *lower link* $\mathrm{Lk}^-(v)$ of $\hat{f}$ at a vertex v is the union of all simplices of the link of v all vertices of which are smaller than v. The upper star $\mathrm{St}^+(v)$ and the upper link $\mathrm{Lk}^+(v)$ are defined similarly.

[1] By simplex we mean a closed cell of T of any dimension.

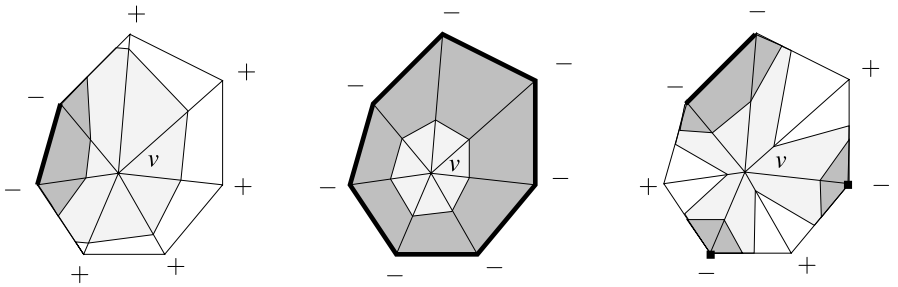


Fig. 4 Morse theory for PL functions in 2D. *Plus* and *minus signs* indicate whether neighbors of v are larger or smaller than v. Lower links are displayed in *bold*, sets $\hat{f}^{-1}(]-\infty, f(v)-\varepsilon])$ in *grey*, and sets $\hat{f}^{-1}([f(v)-\varepsilon, f(v)+\varepsilon])$ in *light grey*

Figure 4 shows that—for small ε—the topological changes between lower level-sets $\hat{f}^{-1}(]-\infty, f(v)-\varepsilon])$ and $\hat{f}^{-1}(]-\infty, f(v)+\varepsilon])$ are determined by the topology of $\mathrm{Lk}^-(v)$. In particular, in 2D, topological changes occur whenever $\mathrm{Lk}^-(v)$ is not connected or equals the link of v (right and middle cases in Fig. 4). This is what motivates the next definition in the higher dimensional case:

Definition 3 A *critical point* of $\hat{f}$ is a vertex whose lower link is not collapsible.[2] A vertex that is not a critical point of $\hat{f}$ will be called *regular*.

With this definition, topological changes in lower level-sets occur exactly at critical points, which is consistent with smooth Morse theory. The *index* of a vertex v is defined to be 1 minus the Euler characteristic of $\mathrm{Lk}^-(v)$ [2]. In particular, regular points all have index 0. The converse is not true however in dimension at least 3. Also, checking if a vertex is regular is easy for PL functions defined on three-dimensional meshes: it is sufficient to check that the lower link and the upper link are both non-empty and connected.[3] Define the index of $\hat{f}$ on a region V to be the sum of the indices of all critical points of $\hat{f}$ lying in V. Again, this definition is consistent with the smooth case, since the PL index can be shown to also satisfy the Poincaré-Hopf index theorem [2]. The following lemma will be used later:

Lemma 2 *If the gradients of $\hat{f}$ on tetrahedra incident to a vertex v all have a positive inner product with some vector, then v is regular.*

Proof By Proposition 1.2 page 450 in [1], $\hat{f}^{-1}(]-\infty, f(v)+\varepsilon])$ retracts by deformation on $\hat{f}^{-1}(]-\infty, f(v)-\varepsilon])$ for sufficiently small ε. Hence $\mathrm{Lk}^-(v)$ has the homology groups of a point, implying that it is collapsible since it is a subcomplex of the 2-sphere. □

[2] A complex is collapsible if it collapses to a point.

[3] This follows from Alexander duality together with the fact that contractible subcomplexes of the 2-sphere are collapsible.

2.2 Main Result

We assume throughout the paper that f and T satisfy conditions a1, a2, a3, b1, b2. That is, M is compact, f is a Morse function, 0 is not a critical value of f, no vertex of T map to 0 by f, and no two neighboring vertices of T map to the same value by f. Additionally, we assume that the following condition holds:

0. f does not vanish on any tetrahedron of T containing a critical point of f.

Theorem 3 *Let W be a subcomplex of T satisfying the following conditions*:

1. *f does not vanish on ∂W.*
2. *W contains no critical point of f.*
2′. *W contains no critical point of $\hat{f}$.*
3. *W collapses to $\hat{M}$.*
4. *f and $\hat{f}$ have the same index on each bounded component of $\Omega \setminus W$.*

Then M and $\hat{M}$ are isotopic in W. Moreover, the Hausdorff distance between M and $\hat{M}$ is smaller than the "width" of W, that is, the maximum over the components V of W of the Hausdorff distance between the subset of ∂V where f is positive and the one where f is negative.

Here, isotopic in W means that M can be continuously deformed into $\hat{M}$ while remaining a manifold embedded in W, so that M could not be a knotted torus if $\hat{M}$ is an unknotted one, for instance. We first prove that under the conditions of the theorem, M and $\hat{M}$ are homeomorphic. Under the assumptions of the theorem, the fact that they actually are isotopic will be a direct consequence of a result obtained in [6]. Before proving the theorem, we first show by some examples that none of its assumptions can be removed. In the three following pictures, (local) minima of f are represented by min, (local) maxima by max, and saddle points by s. Critical points of $\hat{f}$ are represented similarly but with a caret. The sign preceding a critical point symbol indicates the sign of the considered function (f or $\hat{f}$) at the critical point.

Figure 5 shows that condition 0 cannot be removed even in the 2D case. By allowing for critical points of f inside a triangle of T with positive vertices, one can build an example where M has an extra component with respect to $\hat{M}$ without violating

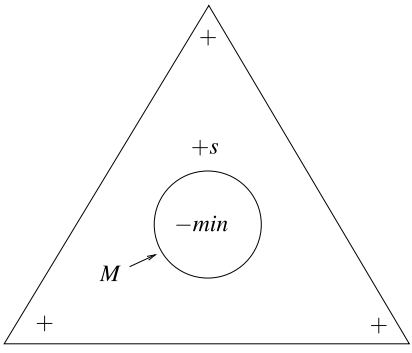


Fig. 5 Condition 0 is necessary

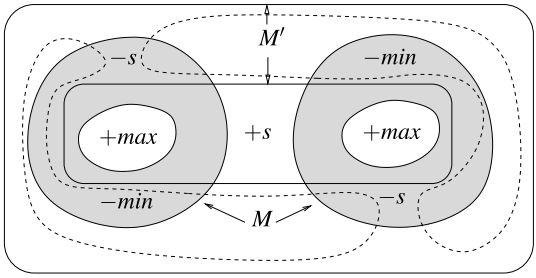


Fig. 6 Critical points do not determine the topology of level-sets

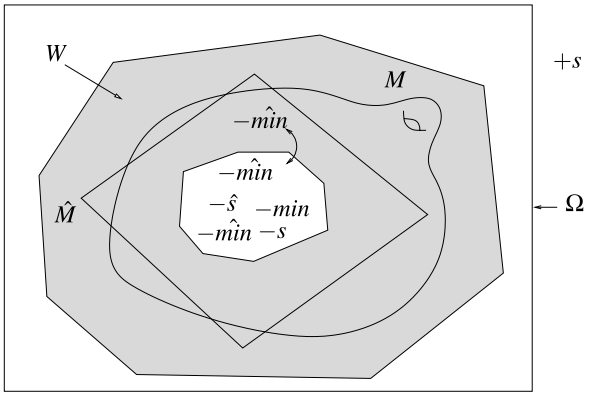


Fig. 7 Condition 2′ and 4 are necessary

conditions involving critical points and their indices. Indeed, in Fig. 5, f has index 0 on the triangle, since minima have index 1 and saddle points have index -1.

The situation in Fig. 6 is a 2D example of two zero-sets M (boundary of the grey region) and M' which are not homeomorphic, though their defining functions have the same critical points, with the same indices. The dashed curve represents a negative level-set of the function defining M'. Such an example can also be built such that $M' = \hat{M}$ for some mesh T. This shows the importance of the set W in the theorem. In particular, conditions 1 and 3 cannot be removed. Indeed, if one drops 1, taking for W any set satisfying 2 and 3 makes the theorem fail. On the other hand, if one drops 3, any W satisfying 2 and 1 also makes the theorem fail.

Figure 7 shows a 3D example where M is a torus whereas $\hat{M}$ is a sphere. This is because $\hat{f}$ has an extra negative minimum inside $\hat{f}^{-1}(]-\infty, 0])$ whereas f has an index 1 saddle point outside the bounding box Ω. Depending on whether this extra minimum lies in W or not (see the circle arc with arrows at both ends in Fig. 7), one obtains counterexamples to the theorem if assumptions 2′ or 4 are dropped. One can build similar examples showing that condition 2 is also needed.

We now return to the proof of Theorem 3.

2.3 Proof of the Theorem

Lemma 4 *Let S and T be two subsets of a topological space X that meet (i.e. $S \cap T \neq \emptyset$). Assume the boundary of S, as well as T and $X \setminus T$, are connected. If*

$X \setminus S$ and $X \setminus T$ meet but their boundaries do not, then S is contained in the interior of T or the other way around.

Proof The boundary of S is the disjoint union of $\partial S \cap \text{int}(T)$ and $\partial S \cap \text{int}(X \setminus T)$ since $\partial S \cap \partial T$ is empty. So we have a partition of ∂S in two relatively open sets. As it is connected, one has to be empty. If $\partial S \cap \text{int}(T)$ is empty then $\partial S \subset \text{int}(X \setminus T)$ that is, $T \cap \partial S$ is empty. As a consequence, T is included in $\text{int}(S)$ or in $\text{int}(X \setminus S)$ by connectedness. Since S and T meet, we have that $T \subset \text{int}(S)$.
Now if $\partial S \cap \text{int}(X \setminus T)$ is empty then $X \setminus T$ is contained in $\text{int}(S)$ or in $\text{int}(X \setminus S)$ by connectedness again. Similarly as above it has to be contained in $\text{int}(X \setminus S)$, which implies that $S \subset T$. Thus $\text{int}(S) \subset \text{int}(T)$ so $\partial S \supset S \setminus \text{int}(T) = S \cap \partial T$. If S would meet ∂T, then ∂S and ∂T would meet, which is impossible. Hence, S is included in the interior of T. □

Lemma 5 *Let V be a connected component of W. $M \cap V$ is a connected smooth compact manifold without boundary.*

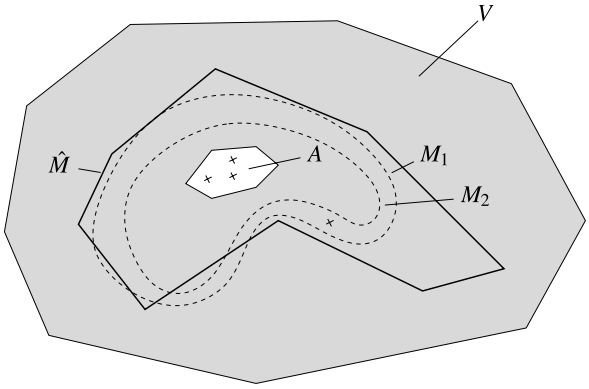
Proof Condition 3 implies easily that V collapses to $\hat{M} \cap V$. Therefore V contains a simplex having positive and negative vertices. As a consequence, f vanishes on V. Since f does not vanish on ∂W (condition 1), M intersects V. Also, M does not meet the boundary of V (condition 1), so $M \cap V$ is a smooth compact manifold without boundary.

Because V, which is connected, collapses to $\hat{M} \cap V$, $\hat{M} \cap V$ is a connected closed surface. Therefore, the complement of $\hat{M} \cap V$ has exactly two components, one of which is bounded. Because V collapses to $\hat{M} \cap V$, $\mathbb{R}^3 \setminus V$ also has exactly one bounded component which we denote by A and one unbounded component we denote by B (see Fig. 8). The complement of A, which is $B \cup V$, is connected, because B and V are connected. For the same reason, $A \cup V$ is also connected. Moreover, the complement of $A \cup V$, being equal to B, is also connected. In summary, A is connected as well as its complement, and the same is true for $A \cup V$.

Call now M_i, $i = 1, \ldots, n$ the connected components of $M \cap V$ (see Fig. 8). For each i, let N_i be the bounded component of $\mathbb{R}^3 \setminus M_i$. $M_i = \partial N_i$ does not meet $\partial(A \cup V) \subset \partial W$ (1), and $A \cup V$ is connected as is its complement. So N_i is included in $A \cup V$ thanks to Lemma 4. Now N_i contains at least one critical point of f. But as $N_i \subset A \cup V$, such a point has to lie in A, by 2. So N_i meets A, but since $\partial N_i = M_i$ does not meet $\partial A \subset \bar{W}$, N_i contains A by Lemma 4 again. Suppose $M \cap V$ is not connected. Then N_1 and N_2 both contain A so they intersect. Because M is smooth, their boundaries do not intersect. So one has w.l.o.g. $N_2 \subset N_1$. Now f vanishes on $\partial(N_1 \setminus N_2) = \partial N_1 \cup \partial N_2$, and therefore has an extremum in $N_1 \setminus N_2$, which is impossible by 2 because $N_1 \setminus N_2 \subset V$. □

So $M \cap V$ and $\hat{M} \cap V$ are connected compact surfaces without boundary. As seen in the preceding proof, A contains all critical points of f enclosed by $M \cap V$. Also, A contains all critical points of $\hat{f}$ enclosed by $\hat{M} \cap V$ by $2'$. From condition 4, we deduce that the volumes enclosed by $M \cap V$ and by $\hat{M} \cap V$ have the same Euler characteristic, since the Euler characteristic of a lower level set is the index of the considered function on that lower level set (Theorem 1). So $M \cap V$ and $\hat{M} \cap V$ have

Fig. 8 Proof of Lemma 5



the same genus and are thus homeomorphic. To complete the proof that M and $\hat{M}$ are homeomorphic, it remains to check that:

Lemma 6 *M is included in W.*

Proof Let D be some component of $\Omega \setminus W$. We claim that $M \cap D$ is empty. First $\hat{M} \cap D$ is empty by condition 1 so w.l.o.g. vertices lying in the closure of D are all positive. If $M \cap D$ is not empty then some component E of $f^{-1}(]-\infty, 0])$ meets D. Moreover, ∂D does not meet E. Indeed, f is positive at vertices of ∂D, and does not vanish on $\partial D \subset \partial W \cup \partial\Omega$ by condition 1. So E, being connected, is included in the interior of D. But then E is compact and thus f reaches its minimum on E, implying that E contains a (negative) critical point of f. This is impossible since the tetrahedron containing this critical point would have negative vertices by condition 0, though being included in D. □

The proof of the bound on the Hausdorff distance between M and $\hat{M}$ is not difficult. Pick any point p in $\hat{M}$ and let V be the component of W containing it. Assume w.l.o.g. that $f(p) > 0$ and let p' be the closest point of p on the component of ∂V where f is negative. By the intermediate value theorem, the line segment pp' meets M at a point q. The distance between p and q is smaller than the distance between p and p' which is smaller than the Hausdorff distance between the two components of ∂V. This shows one part of the bound. The other part can be proved in a similar way.

Now that we know that M and $\hat{M}$ are homeomorphic, the fact that they are isotopic is a consequence of Proposition 7, which is proved in [6].

Proposition 7 *Let $\hat{M}$ be an orientable compact surface without boundary and let M be a surface such that*

- *$\hat{M}$ is homeomorphic to M,*
- *M separates the sides of a topological thickening*[4] *$\tilde{W}$ of $\hat{M}$.*

Then M is isotopic to $\hat{M}$ in $\tilde{W}$.

[4]This means that there is a homeomorphism $\Phi : \tilde{W} \to \hat{M} \times [0, 1]$ mapping $\hat{M}$ to $\hat{M} \times \{1/2\}$.

Indeed, considering a regular neighborhood of W [20] yields the desired topological thickening $\tilde{W}$, as can be seen from the uniqueness theorem for regular neighborhoods from piecewise-linear topology [20].

3 Algorithm

In the algorithm, we take as W a set that is related to the notion of watershed from topography. This set satisfies properties $2'$ and 3 by construction. In Sect. 3.1, we give its definition, basic properties, and construction algorithms. Section 3.2 describes the meshing algorithm itself, which ensures that W fulfills also conditions 0, 1, 2, and 4, and proves its correctness.

3.1 PL Watersheds

We first assume that the mesh T conforms to $\hat{M}$, i.e. $\hat{M}$ is contained in a union of triangles of T. We will see later how to remove this assumption, which is in contradiction with the genericity assumptions. Define W^+ as the result of the following procedure:

Positive Watershed Algorithm

set $W^+ = \hat{M}$.
mark all vertices of $\hat{M}$.
while there is a positive regular unmarked vertex v of T
  such that the vertices of $Lk^-(v)$ are marked
do
  set $W^+ = W^+ \cup St^-(v)$.
  mark v.
end while
return W^+

W^- is defined as the result of the same algorithm applied to $-f$. We set $W = W^+ \cup W^-$. Note that W contains no critical point of $\hat{f}$. Also, positive marked vertices are exactly the vertices of W^+.

Lemma 8 *W collapses to $\hat{M}$.*

Proof It is sufficient to show the result for W^+. Let W_i^+ be the state of W^+ after i steps of the algorithm, and let v_i be the i-th marked vertex. As $W_0^+ = \hat{M}$, the only thing we have to show is that W_{i+1}^+ collapses to W_i^+ for all i. Let us first show that $\mathrm{Lk}^-(v_i)$ is included in W_i^+. If it is not the case, let u be the largest vertex of some simplex s of $\mathrm{Lk}^-(v_i)$ that is not in W_i^+. Simplex s is in $\mathrm{St}^-(u)$ which is therefore not included in W_i^+. This is a contradiction since v_i is marked. Therefore $\mathrm{Lk}^-(v_i) \subset W_i^+$. Now since v_i is regular, $\mathrm{Lk}^-(v_i)$ is collapsible. Consider a sequence of elementary collapses allowing to collapse $\mathrm{Lk}^-(v_i)$ to p and let $s_j \subset \mathrm{Lk}^-(v_i)$, $j = 1, \ldots, n$

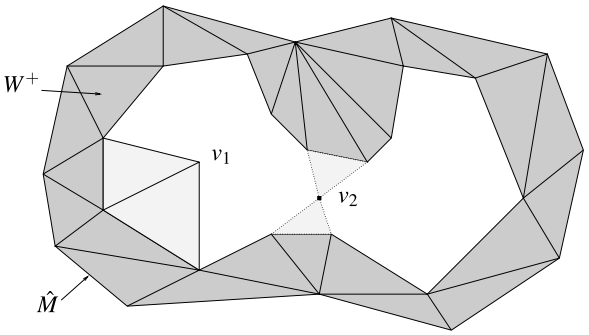


Fig. 9 Construction of W^+: lower stars of regular vertices (such as v_1) are added one by one. Lower stars of critical vertices (v_2) are discarded

be the sequence of simplices defining these elementary collapses. The simplices $\mathrm{conv}(s_j \cup v_i)$, $j = 1, \ldots, n$ and the edge pv_i define a valid sequence of elementary collapses allowing to collapse $W^+_{i+1} = W^+_i \cup \mathrm{St}^-(v_i)$ to W^+_i, which concludes the proof. □

One may prefer a more intrinsic definition of W^+. In the same spirit as in [9], one can define a partial order on the vertices of T by the closure of the acyclic relation $\prec$ defined by $u \prec v$ if $u \in \mathrm{Lk}^-(v)$ or $u = v$. We will denote this order $\prec$ again and say that v *flows* into u whenever $u \prec v$. The next lemma shows that the vertices of W^+ do not depend on the order in which the vertices are considered in the construction.

Lemma 9 *The vertices of W^+ are exactly the positive vertices that do not flow into any positive critical point of $\hat{f}$.*

Proof The vertices of W^+ have this property by construction. Let p be a positive vertex not belonging to W^+ and assume p does not flow into any positive critical point. In particular, p is regular by reflexivity. Hence, as $p \notin W^+$, the lower link of p, which is not empty, has to contain an unmarked vertex. It cannot contain a critical point because as T conforms to $\hat{M}$, vertices in $\mathrm{Lk}^-(p)$ are all non-negative, and so p would flow into a positive critical point. There is thus an unmarked vertex in $\mathrm{Lk}^-(p)$. If we can choose an unmarked positive vertex p_1 in $\mathrm{Lk}^-(p)$, then p_1 does not belong to W^+, and flows into a positive critical point. Repeating this process with p replaced by p_1, we find a strictly decreasing sequence of positive vertices, that thus has to end. Let p_k be its last term. The lower link $\mathrm{Lk}^-(p_k)$ contains no positive unmarked vertices. But as T conforms to $\hat{M}$, vertices in $\mathrm{Lk}^-(p_k)$ are all non-negative. Since vertices of $\hat{M}$ are marked, we get a contradiction. □

Note that W is the union of simplices with all their vertices in W. As a result, we get an intrinsic definition of W, and not only of its vertices. From an algorithmic point of view, it may be efficient to examine the vertices in increasing order in the construction of W^+. One can for instance maintain the ordered list of vertices neighboring W, always consider the first element of this list for marking, and discard it if it cannot be marked. Indeed, with this strategy, a vertex that cannot be marked at some point will never be marked.

Another consequence of Lemma 9, which will be useful later, goes as follows. Let c be the minimum of $|\hat{f}(v)|$, and hence the minimum of $|f(v)|$ over all critical points v of $\hat{f}$.

Lemma 10 *W contains all vertices the image of which under $|f|$ is smaller than c.*

Proof Let p be such that $|f(p)| < c$. Without loss of generality, assume that p is positive. Any critical point v into which p flows satisfies $f(v) < f(p)$. So it cannot be positive by definition of c: by Lemma 9, p lies in W^+. □

Non conforming case We now drop the assumption that T conforms to $\hat{M}$ and assume genericity again. From T and $\hat{M}$ one can build a mesh S that is finer than T, conforms to $\hat{M}$, and has all its extra vertices on $\hat{M}$. Indeed, it suffices to triangulate the overlay of $\hat{M}$ and T without adding extra vertices except those of $\hat{M} \cap T$. This can be done as the cells of the overlay are convex. The construction of W described above can then be applied to S. A positive vertex of T has its lower link in S containing only vertices of $\hat{M}$ if and only if its lower link in T contains only negative vertices. Thus, in order to find the positive vertices of $W \cap T$, one can apply the positive watershed algorithm described above to T, if at the initialization step one marks all negative vertices having a positive neighbor instead of those of $\hat{M}$. Still, note that if a negative critical point has a positive neighbor, then this neighbor will not be marked by this modified algorithm, whereas it could have been marked by the standard algorithm applied to S. However, if we assume that vertices having a neighbor of opposite sign are regular (condition c), then this does not happen and the result W' of the modified algorithm is equal to W. The negative vertices of $W \cap T$ are determined similarly. In our meshing algorithm, we will not build the mesh S, but rather make sure condition c holds, and apply the modified algorithm.

Updating W' The intrinsic definition of W—or W'—given above yields an efficient way of updating W when T undergoes local transformations. It is sufficient to describe the algorithm for updating the vertices of W^+. Let T_1 be a mesh obtained from T by removing some set of tetrahedra E and remeshing the void left by E. Call A the set of positive critical points of the linear interpolation of f on T_1 that lie in E. Then the vertex set of the positive watershed W_1^+ associated with T_1 can be computed from the vertex set of W^+ by performing the following two operations. To begin with, the set of vertices of T_1 that flow into A must be removed from W^+ (Lemma 9), which amounts to a graph traversal. The remaining vertices of T_1 all belong to W_1^+. Then, mark these vertices and apply the positive watershed algorithm loop to get the other vertices of W_1^+.

Remark The watershed we compute is in general strictly included in the 'true watershed'. The 'true watershed' seems hard to compute, though, and can intersect a triangle in a very complicated way. There might be interesting intermediate definitions between ours and the true one, for instance based on the PL analog of the Morse complex introduced in [10].

3.2 Main Algorithm

Theorem 3 enables us to build a mesh isotopic to M using two simple predicates, *vanish* and *vanish'*. The predicate *vanish* (resp. *vanish'*) takes a triangle or a box and return true if f (resp. ∇f) vanishes on that triangle or that box. We actually do not even need predicates, but rather filters. More precisely, *vanish* (or *vanish'*) may return true even if f does not vanish on the considered element, but not the other way around. Still, we require that *vanish* returns the correct answer if the input triangle or box is sufficiently small. Such filters can be designed using interval analysis.

Our algorithm also requires to build a refinable triangulation of space such that $\hat{f}$ (resp. $\nabla \hat{f}$) converges to f (resp. ∇f) when the size of the elements tends to 0. As noticed by Shewchuk [23], this is guaranteed provided all tetrahedra have dihedral and planar angles bounded away from π. In [3], Bern, Eppstein and Gilbert described an octree-based algorithm yielding meshes the angles of which are bounded away from 0. In our case, which is much easier, the desired triangulation can simply be obtained by adding a vertex at the center of each square and each cube of the octree, triangulating the squares radially from their center, and doing the same with the cubes. Indeed, resulting planar and dihedral angles are all bounded away from π. One can expect that this scheme does not produce too many elements upon refinement, because the size of elements is allowed to change rapidly as we do not require that these have a bounded aspect ratio (see Fig. 10). The main algorithm uses an octree O, the associated triangulation T, and the watershed W'. We will say that two (closed) boxes of O are neighbors if they intersect. O is initialized to a bounding box Ω of M. Such a bounding box can be found by computing the critical points of the coordinate functions restricted to M, if possible, or by using interval analysis. Besides, we maintain five sets of boxes ordered by decreasing size. *Critical1* is a certain set of boxes obtained by interval analysis (see below). This set has the property that the union of its boxes, which we call the *critical set*, encloses all critical points of f but does not intersect M. *Critical2* contains all boxes containing a critical point of $\hat{f}$ that is not in a box belonging to *Critical1*. *Index* contains all boxes neighboring a box b in *Critical1* such that f and $\hat{f}$ have different indices on the connected component of the critical set that contains b. We defer the description of a method that computes the

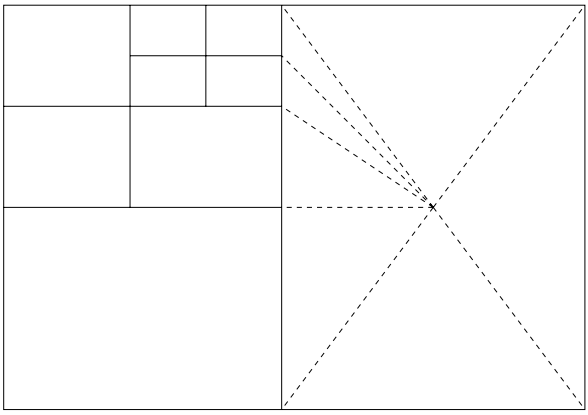


Fig. 10 Octree and triangulation used in the algorithm. In this 2D example, only the edges of the triangulation of the box on the right are shown (*dashed*)

index of f on a box in a certified way to the appendix. *Boundary1* contains all boxes containing two neighboring vertices of opposite signs one of which is critical for $\hat{f}$ (condition c, see paragraph **Non conforming case**). *Boundary2* contains all boxes that are not included in W', and that contain a triangle t of $\partial W'$ such that *vanish*(t) is true. Finally, for our algorithm to work, we need to introduce a slight modification of the watershed W', which we call W''. The modification consists of taking as W''^{+} vertices—and the same for W''^{-}—the positive vertices that do not flow into positive critical points of $\hat{f}$ nor into vertices lying in a box containing a positive critical point of f. With this modification, Lemma 8 still holds and Lemma 10 holds if one replaces c by the minimum c' of c and the minimum of $|f|$ on the boxes containing a critical point of f. Also, c' is positive as f does not vanish on these boxes.

Main Algorithm

Initialization Refine O until all boxes b satisfy either *vanish*(b) is false or *vanish'*(b) is false. Insert all boxes b such that *vanish'*(b) is true in *Critical1*. compute T and W'', and the four sets *Critical2*, *Boundary1*, *Boundary2*, and *Index*.
while (**true**) **do**
  update T, W'', and the four sets.
  if *Critical2* $\neq \emptyset$ **then**
    split its first element.
  else if *Boundary1* $\neq \emptyset$ **then**
    split its first element.
  else if *Boundary2* $\neq \emptyset$ **then**
    split its first element.
  else if f and $\hat{f}$ have different indices on some component of the critical set **then**
    split the first element of *Index*.
  else
    return $\hat{M}$
  end if
end while

Thanks to Theorem 3 applied to W'', the correctness of this algorithm amounts to its termination. We now show that the main algorithm terminates. First note that after the initialization step, no box containing a critical point of f is split, because such boxes belong to *Critical1*. The magnitude of ∇f is thus larger than a certain constant $g_{\min}$ on the complement C of the union of these boxes. Let us show that the size of the boxes of *Critical2* that are split at some point is bounded from below. As $\nabla \hat{f}$ converges to ∇f, there is a number s_1 such that for each tetrahedron with diameter smaller than s_1, $\|\nabla f - \nabla \hat{f}\|$ is smaller than $g_{\min}/2$ on the interior of that tetrahedron. If the tetrahedron is included in C, $\|\nabla f\| > g_{\min}$, which implies that $\nabla \hat{f}$ and ∇f make an angle smaller than $\pi/6$.

Lemma 11 *Let $A \subset \mathbb{R}^3$ be such that ∂A is a manifold included in C and containing no vertex of T. Suppose that all boxes meeting ∂A are smaller than s_1. Then f and $\hat{f}$ have the same index on A.*

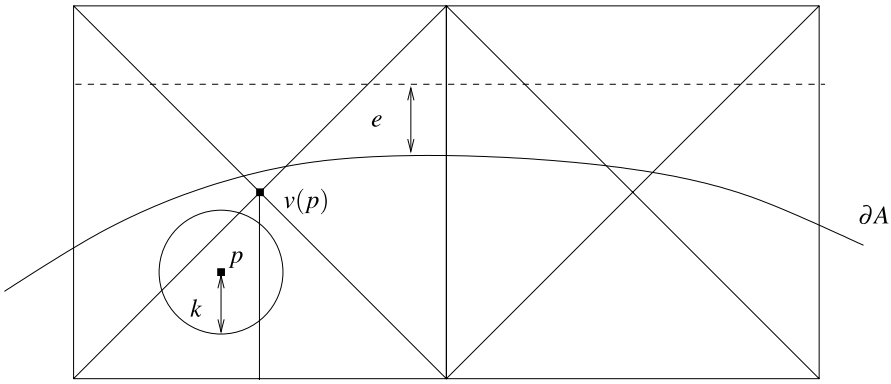


Fig. 11 Proof of Lemma 11

The proof of Lemma 11 resorts to stratified Morse theory, which is an extension of both the smooth and PL Morse theory to the case of piecewise smooth functions. We refer to [11] for a complete exposition of this subject.

Proof For $p \in \partial A$, let $d(p)$ denote the largest number such that the simplices of T that meet the open ball centered at p of radius $d(p)$ all share a vertex, $v(p)$. The quantity $d(p)$ is the 3-dimensional analog of the local feature size function introduced by Ruppert [21]. We call $d_{\min}$ the minimum of d, which is known to be positive, and set k equal to the minimum of $d_{\min}$ and e, where e is half the distance from ∂A to the closest box that does not meet ∂A.

Let us now consider a smooth nonnegative function $\phi : \mathbb{R}^3 \to \mathbb{R}$ with support included in the open ball centered at 0 of radius k. The convolution of $\hat{f}$ and ϕ is a smooth function $\tilde{f}$. Let p be a point at distance less than e from ∂A. The gradient of $\tilde{f}$ at p is a weighted average of the gradients of $\hat{f}$ at points lying in the open ball centered at p and with radius k. All gradients involved in this average are gradients of $\hat{f}$ on tetrahedra incident on $v(p)$. Moreover, the size of these tetrahedra is smaller than s_1 because $k \leq e$. As a consequence, all gradients considered make an angle smaller than $\pi/6$ with the gradient of f at $v(p)$. As the weights in the average are nonnegative, we have that the angle between $\nabla\tilde{f}(p)$ and $\nabla f(v(p))$ is smaller than $\pi/6$. Also, the angle between $\nabla f(v(p))$ and $\nabla f(p)$ is less than $\pi/3$ since both vectors make an angle smaller than $\pi/6$ with the gradient of $\hat{f}$ on some tetrahedron containing p and $v(p)$. Finally, we get that $\nabla\tilde{f}(p)$ and $\nabla f(p)$ have a positive inner product.

Let now U_1 be a neighborhood of ∂A whose closure does not contain any vertex of T and let U_2 be an open set such that $U_1 \cup U_2 = \mathbb{R}^3$. We also require that the Hausdorff distance between U_1 and ∂A is smaller than e and that $U_2 \cap \partial A = \emptyset$. Denote by $\{u_1, u_2\}$ a partition of unity subordinate to the covering $\{U_1, U_2\}$. This means that u_1 and u_2 are nonnegative smooth function defined on $\mathbb{R}^3$, with support in U_1 and U_2 respectively, and such that $u_1 + u_2 = 1$. In particular, u_2 equals 1 on the complement of U_1, and u_1 equals 1 on the complement of U_2. So the function $g = u_2\hat{f} + u_1\tilde{f}$ coincide with $\hat{f}$ on $\mathbb{R}^3 \setminus U_1$ and with $\tilde{f}$ on $\mathbb{R}^3 \setminus U_2$. Now recall that $\nabla\tilde{f}$ and ∇f have a positive inner product on ∂A, which is contained in the complement of U_2. Hence the linear homotopy between both vector fields does not vanish on ∂A: by normaliza-

tion, one gets a homotopy between $\nabla\tilde{f}/\|\nabla\tilde{f}\|$ and $\nabla f/\|\nabla f\|$, considered as maps from ∂A to the unit sphere. Because the degree (see [13] p. 134 for a definition) is invariant under homotopy, we deduce that these maps have the same degree, which shows that f and $\tilde{f}$ have the same index on A. Now as g and $\tilde{f}$ coincide in a neighborhood of ∂A, f and g have the same index on A. To complete the proof, it thus suffices to show that g and $\hat{f}$ also have the same index on A. Now the critical points of $\hat{f}$ are critical for g, with the same index, as U_1 contains no such point. Potential other critical points of g can only lie in U_1. But the gradient of g at any point p of U_1 where it is defined is a convex combination of $\nabla\tilde{f}(p)$ and $\nabla\hat{f}(p)$: it thus has a positive inner product with $\nabla f(p)$. By the result of [1] which we mentioned when we stated Lemma 2, this implies that the index of p is 0. We thus proved the announced claim. □

Suppose that some box b of *Critical2* of size smaller than s_1 is split. Let v be a critical point of $\hat{f}$ included in b. All the boxes containing v are in *Critical2* and their size is smaller than s_1 since we consider boxes in decreasing order. Now the gradients of $\hat{f}$ on tetrahedra incident on v all have a positive inner product with $\nabla f(v)$ (recall ∇f and $\nabla\hat{f}$ make an angle less than $\pi/6$), which is a contradicts Lemma 2, implying that v is not critical. So the conclusion is that *Critical2* becomes—at least temporarily—empty after a finite number of consecutive splittings of boxes in *Critical2*.

Now if the algorithm splits a box b in *Boundary1*, then b contains a critical point of $\hat{f}$. This critical point, which we assume to be positive, belongs to a box containing a critical point of f as *Critical2* is empty. So the maximum of $|f|$ on b is larger than the minimum of $|f|$ on the boxes containing a critical point of f (i.e. c'). On the other hand, f vanishes on b since b contains a negative vertex. This cannot happen if the size of b is below a certain value, so that boxes in *Boundary1* cannot be split indefinitely.

Suppose that the algorithm splits arbitrarily small boxes in *Boundary2*. If a small enough box b is split, then b contains a triangle t of W'' on which f vanishes. So, if the size of b is small enough, the maximum of $|f|$ on b will be smaller than c'. By Lemma 10, all vertices of b will then belong to W'' so $b \subset W''$ which is a contradiction. Thus the size of split boxes in *Boundary2* is also bounded from below.

To complete the proof of termination, we need to prove that *Index* does not contain boxes that are too small. This is true by applying Lemma 11 to smooth neighborhoods of each connected component of the critical set. Finally:

Theorem 12 *The main algorithm returns an isotopic piecewise linear approximation of M.*

If one wishes to guarantee in addition that the Hausdorff distance between M and its approximation is less than say ε, by Theorem 3 it is sufficient to modify the positive watershed algorithm so as to control that the width of W is smaller than ε.

4 Conclusion

We have given an algorithm that approximates regular level sets of a given function with piecewise linear manifolds having the same topology. Though no implementation has been carried out, we believe that it should be rather efficient due to the simplicity of the involved predicates and the relative coarseness of the required space decomposition.

Appendix

We now briefly explain how to compute the index of a generic smooth function $f : \mathbb{R}^3 \to \mathbb{R}$ on a box $B \subset \mathbb{R}^3$ in a certified way. Without loss of generality, we assume that $B = [0, 1]^3$. Our approach is based on a recursive definition of the index of a vector field introduced in [12]. The central formula in this work is the following (see Fig. 12). If V denotes a vector field (in our case, $V = \nabla f$) defined on a compact smooth n-manifold M and not vanishing on ∂M, then the index of V satisfies:

$$\mathrm{Ind}(V) = \chi(M) - \mathrm{Ind}(\partial_- V).$$

Here $\partial_- V$ is a vector field defined on $\partial_- M$, which is the set of boundary points where V points inwards. On $\partial_- M$, $\partial_- V$ coincides with the projection of V on the tangent space of ∂M. Now suppose we can find a $(n-1)$-submanifold $M_1 \subset \partial_- M$ that contains all zeroes of $\partial_- V$. Then, to compute the index of V on M, it is sufficient to compute the index of $\partial_- V$ on M_1 (and the Euler characteristic of M_1). By repeated application of this principle, we can express the index of V as a sum of Euler characteristics and indices of vector fields defined over 1-manifolds, which are trivial to compute.

To apply this strategy to our case, in which $M = B$ has edges and corners, we conceptually consider offsets of M, which are smooth, and let the offset parameter go to 0. Almost by definition, in this setting the zeros of $\partial_- V$ are the points where V belongs to the normal cone and points inwards. Using interval analysis, it is not difficult to find a subset B_1 of $\partial_- B$ that contains all such points, and such that $\partial_- V$

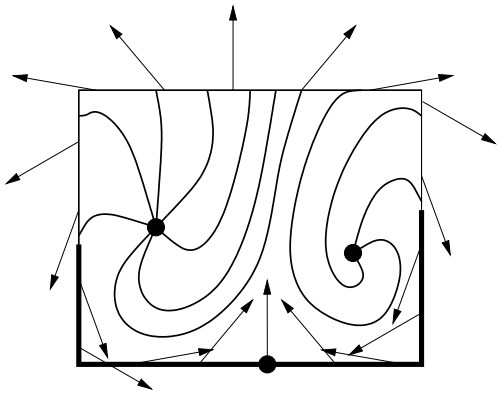


Fig. 12 An index 2 vector field V on a *square* C represented by a *few flow lines*. $\partial_- C$ is in *bold*. The *dot* on $\partial_- C$ represents the unique zero of $\partial_- V$, which has index -1

does not vanish on ∂B_1. To do this, we recursively subdivide the faces of the cube until all cells satisfy one of the two following conditions: either the cell does not contain a zero of $\partial_- V$, or it is included in $\partial_- B$. The union of the cells of the latter type will then provide a suitable B_1. For a square C lying on the face supported by, say, the plane $z = 1$, sufficient conditions ensuring that C does not contain any zero of $\partial_- V$ are

$$(V_z(C) > 0) \quad \text{or} \quad (0 \notin V_x(C)) \text{ or } (0 \notin V_y(C)).$$

Here $V_z(C) > 0$ for instance means that the z-coordinate of V is positive on C. The condition under which C is included in $\partial_- B$ is obviously $V_z(C) < 0$. Edges of the cube might also have to be subdivided. Without loss of generality we assume that edge E is supported by the line with equation $x = y = 1$. Then sufficient conditions under which E cannot contain a zero are as follows:

$$(V_x(E) > 0) \quad \text{or} \quad (V_y(E) > 0) \text{ or } (0 \notin V_z(E)).$$

Also, the condition under which E is included in $\partial_- B$ is $(V_x(E) < 0)$ and $(V_y(E) < 0)$. It can be checked that this subdivision process terminates if V has no zeroes on the surface of the cube, which is a generic condition. Upon termination of the subdivision process, we obtain a set B_1 to which the formula can be applied. It thus remains to recursively subdivide the boundary edges of B_1 in a similar way as above to complete the computation of the index of V.

References

1. Agrachev, A.A., Pallaschke, D., Scholtes, S.: On Morse theory for piecewise smooth functions. J. Dyn. Control Syst. **3**, 449–469 (1997)
2. Banchoff, Th.: Critical points and curvature for embedded polyhedra. J. Differ. Geom. **1**, 245–256 (1967)
3. Bern, M., Eppstein, D., Gilbert, J.: Provably good mesh generation. J. Comput. Syst. Sci. **48**, 384–409 (1994)
4. Bloomenthal, J.: Introduction to Implicit Surfaces. Morgan Kaufmann Series in Computer Graphics and Geometric Modeling. Morgan Kaufmann, San Mateo (1997)
5. Boissonnat, J.D., Oudot, S.: Provably good sampling and meshing of surfaces. Graph. Models **67**, 405–451 (2005)
6. Chazal, F., Cohen-Steiner, D.: A condition for isotopic approximation. In: Proc. ACM Symp. Solid Modeling and Applications, 2004
7. Cheng, S.-W., Dey, T.K., Ramos, E., Ray, T.: Sampling and meshing a surface with guaranteed topology and geometry. In: Proc. 20th Sympos. Comput. Geom., pp. 280–289, 2004
8. Dobkin, D.P., Levy, S.V.F., Thurston, W.P., Wilks, A.R.: Contour tracing by piecewise linear approximations. ACM Trans. Graph. **9**(4), 389–423 (1990)
9. Edelsbrunner, H.: Surface reconstruction by wrapping finite point sets in space. In: Aronov, B., Basu, S., Pach, J., Sharir, M. (eds.) Discrete and Computational Geometry. The Goodman–Pollack Festschrift, pp. 379–404. Springer, Berlin (2003)
10. Edelsbrunner, H., Harer, J., Zomorodian, A.: Hierarchical Morse complexes for piecewise linear 2-manifolds. In: Proc. 17th Annu. ACM Sympos. Comput. Geom., pp. 70–79, 2001
11. Goresky, M., MacPherson, R.: Stratified Morse Theory. Springer, Berlin (1988)
12. Gottlieb, D., Samaranayake, G.: The index of discontinuous vector fields. N. Y. J. Math. **1**, 130–148 (1995)
13. Hatcher, A.: Algebraic Topology. Cambridge University Press, Cambridge (2002)
14. Hirsch, M.: Differential Topology. Springer, Berlin (1976)

15. Lopez, A., Brodlie, K.: Improving the robustness and accuracy of the marching cubes algorithm for isosurfacing. IEEE Trans. Vis. Comput. Graph. 9(1), 2003
16. Lorensen, W.E., Cline, H.E.: Marching Cubes: a high resolution 3D surface construction algorithm. Comput. Graph. **21**(4), 163–169 (1987)
17. Milnor, J.: Morse Theory. Ann. of Math. Studies, vol. 51. Princeton University Press, Princeton (1963)
18. Plantinga, S., Vegter, G.: Isotopic approximation of implicit curves and surfaces. In: Proceeding Symposium on Geometry Processing, pp. 251–260. Nice, France (2004)
19. Stander, B.T., Hart, J.C.: Guaranteeing the Topology of an Implicit Surface Polygonizer for Interactive Modeling. In: Proceedings of SIGGRAPH 97, pp. 279–286
20. Rourke, C.P., Sanderson, B.J.: Introduction to Piecewise-Linear Topology. Springer, Berlin (1982)
21. Ruppert, J.: A Delaunay refinement algorithm for quality 2-dimensional mesh generation. J. Algorithms **18**, 548–585 (1995)
22. Sard, A.: The measure of the critical values of differentiable maps. Bull. Am. Math. Soc. **48**, 883–890 (1942)
23. Shewchuk, J.R.: What is a good linear finite element? Interpolation, conditioning, and quality measures. In: Eleventh International Meshing Roundtable, Ithaca, New York, pp. 115–126, Sandia National Laboratories, September 2002
24. Snyder, J.M.: Interval analysis for computer graphics. In: Proceedings of SIGGRAPH 92, pp. 121–130
25. Velho, L., Gomes, J., de Figueiredo, L.H.: Implicit Objects in Computer Graphics. Springer, Berlin (2002)
26. Velho, L.: Simple and efficient polygonization of implicit surfaces. J. Graph. Tools **1**(2), 5–24 (1996). ISSN 1086-7651

Line Transversals to Disjoint Balls

Ciprian Borcea · Xavier Goaoc · Sylvain Petitjean

Abstract We prove that the set of directions of lines intersecting three disjoint balls in $\mathbb{R}^3$ in a given order is a strictly convex subset of $\mathbb{S}^2$. We then generalize this result to n disjoint balls in $\mathbb{R}^d$. As a consequence, we can improve upon several old and new results on line transversals to disjoint balls in arbitrary dimension, such as bounds on the number of connected components and Helly-type theorems.

Keywords Transversal · Geometric permutation · Convexity

1 Introduction

Helly's theorem [12] of 1923 opened a large field of inquiry designated now as *geometric transversal theory*. A typical concern is the study of all k-planes (also called k-flats) which intersect all sets of a given family of subsets (or *objects*) in $\mathbb{R}^d$. These are the *k-transversals* of the given family and they define a certain subspace of the corresponding Grassmannian. True to its origin, transversal theory usually implicates *convexity* in some form, either in its assumptions, its proofs or most likely, both.

In what follows, $k = 1$ and the objects will be pairwise disjoint closed balls with arbitrary radii in $\mathbb{R}^d$. Our main result is the following convexity theorem:

C. Borcea (✉)
Rider University, Lawrenceville, NJ 08648, USA
e-mail: borcea@rider.edu

X. Goaoc
LORIA–INRIA Lorraine, Nancy, France
e-mail: goaoc@loria.fr

S. Petitjean
LORIA–CNRS, Nancy, France
e-mail: petitjea@loria.fr

Theorem 1 *The directions of all oriented lines intersecting a given finite family of disjoint balls in* $\mathbb{R}^d$ *in a specific order form a strictly convex subset of the sphere* $\mathbb{S}^{d-1}$.

As a first consequence, the connected components in the space of line transversals correspond to the possible *geometric permutations* of the given family, where a geometric permutation is understood as a pair of orderings defined by a single line transversal with its two orientations. This is not true in general, not even for $n \geq 4$ disjoint line segments in $\mathbb{R}^3$.

Before discussing other implications, we want to emphasize that the *key* to our theorem resides in the case of *three disjoint balls in* $\mathbb{R}^3$, and the approach we use to settle this case is geometrically quite revealing, in that it shows the nuanced dependency of the convexity property on the *curve of common tangents* to the three bounding spheres.

1.1 Relation to Previous Work

Helly's theorem [12] states that a finite family $\mathcal{S}$ of convex sets in $\mathbb{R}^d$ has non-empty intersection if and only if any subfamily of size at most $d+1$ has non-empty intersection. Passing from $k=0$ to $k=1$, one of the early results is due to Danzer [7] who proved that n disjoint *unit* disks in the plane have a line transversal if and only if every five of them have a line transversal. Hadwiger's theorem [11], which allows arbitrary disjoint convex sets in the plane as objects, showed the importance of the *order* in which oriented line transversals meet the objects: when every three objects have an oriented line transversal respecting some fixed order of the whole family, there must be a line transversal for the family.

This stimulated interest in comparing, for arbitrary dimension, two equivalence relations for line transversals: a coarse one, *geometric permutation*, determined by the order in which the given disjoint objects are met (up to reversal of orientation) and a finer one, *isotopy*, determined by the connected components of the space of transversals.

In general, for $d \geq 3$, the gap between the two notions may be wide [8], and families for which the two notions coincide are thereby "remarkable". The first examples of such families are "thinly distributed" balls[1] in arbitrary dimension, as observed by Hadwiger [9, 10]. Then, the work of Holmsen et al. [14] showed that disjoint *unit* balls in $\mathbb{R}^3$ provide remarkable cases as well. They verified the convexity property in the case of equal radii, and their method can be extended to the larger class of "pairwise inflatable" balls[2] in arbitrary dimension [6], inviting the obvious question regarding disjoint balls of arbitrary radii. The significance of this problem is also discussed in the recent notes [19, p. 191–195] where one can find ample references to related literature.

[1] A family of balls is *thinly distributed* if the distance between the centers of any two balls is at least twice the sum of their radii.

[2] A family of balls is *pairwise inflatable* if the squared distance between the centers of any two balls is at least twice the sum of their squared radii.

Our solution for the case of arbitrary radii is based on a new approach, suggested by the detailed study of the curve of common tangents to three spheres in $\mathbb{R}^3$ [2]. The main ideas are outlined in Sect. 3 as a preamble to the detailed proof in Sects. 4 to 6.

In dimension three particularly, there are connections with other problems in visibility and geometric computing. Changes of visibility (or "visual events") in a scene made of smooth obstacles typically occur for multiple tangencies between a line and some of the obstacles [20]. Tritangent and quadritangent lines play a prominent role in this picture as they determine the 1- and 0-dimensional faces of visibility structures. An attractive case is that of four balls in $\mathbb{R}^3$ which allow, generically, up to twelve common real tangents [17]. Degenerate configurations are identified in [3]. Variations on such problems, where reliance on algebraic geometry comes to the forefront, are surveyed in [22]. See also a brief account in [1].

1.2 Further Implications

Danzer's theorem [7] motivated several other attempts to generalize Helly's result for $k = 1$, that is, for line transversals. Whereas Helly's theorem only requires convexity, the case $k = 1$ appears to be more sensitive to the geometry of the objects. In particular, Holmsen and Matoušek [15] showed that no such theorem holds in general for families of disjoint translates of a convex set, not even with restriction on the ordering *à la* Hadwiger. Our Theorem 1 has consequences in this direction, presented below in Sect. 7.

Hadwiger's proof of his Transversal Theorem [11] relies on the observation that any *minimal pinning configuration*, that is, any family of objects with an isolated line transversal that would become non-isolated should any of the objects be removed, has size 3 if the objects are disjoint convex sets in the plane. Theorem 1 implies that any minimal pinning configuration of disjoint balls in $\mathbb{R}^d$ has size at most $2d - 1$ (Corollary 14). A generalization of Hadwiger's theorem for families of disjoint balls then follows (Corollary 15).

2 Preliminaries

2.1 Notations and Prerequisites

For any two vectors $\mathbf{a}$, $\mathbf{b}$ of $\mathbb{R}^3$, we denote by $\langle \mathbf{a}, \mathbf{b} \rangle$ their dot product and by $\mathbf{a} \times \mathbf{b}$ their cross product. These expressions will retain their algebraic meaning when $\mathbf{a}$ and $\mathbf{b}$ are complex vectors.

The space of directions in $\mathbb{R}^3$ is the real projective space $\mathbb{P}^2 = \mathbb{P}^2(\mathbb{R})$ envisaged either as the space of lines through the origin (and then the direction of a line is given by its parallel through the origin) or as the "plane at infinity" in the completion $\mathbb{P}^3 = \mathbb{R}^3 \cup \mathbb{P}^2$ (and then the direction of a line is simply its point of intersection with the plane at infinity). A non-zero vector $\mathbf{u} \in \mathbb{R}^3$ may also stand for the direction $(u_1 : u_2 : u_3)$ it defines in $\mathbb{P}^2$.

Convexity in $\mathbb{P}^2$ is relative to the metric induced by the standard metric of the sphere through the identification $\mathbb{S}^2/\mathbb{Z}_2 = \mathbb{P}^2$. All considerations can be pulled-back to $\mathbb{S}^2$ by orienting the lines.

In following our convexity arguments related to three disjoint balls in $\mathbb{R}^3$, it may be helpful to bear in mind that the regions of $\mathbb{P}^2$ determined by directions of line transversals are always contained in the simply-connected side of some smooth conic[3]. When testing convexity, one may use affine charts $\mathbb{R}^2$, and verify locally, then globally, that the boundary curve "stays on the same side of its tangent". If this property were to fail at some point, one must have an *inflection point* there or, in one word, a *flex*.

We denote by B_0, B_1, B_2 three balls in $\mathbb{R}^3$ with respective centers $\mathbf{c}_0, \mathbf{c}_1, \mathbf{c}_2$ and squared radii s_0, s_1, s_2, $s_k = r_k^2$. Since degenerate cases are eventually shown to follow from the generic case (Lemma 10), we assume here that we have a non-degenerate *triangle of centers*.

2.2 Direction-sextic

The directions of common tangent lines to B_0, B_1, B_2 make up an algebraic curve of degree six in $\mathbb{P}^2$, which we call the *direction-sextic* and denote by σ. To take advantage of symmetries in expressing σ, we introduce the edge vectors $\mathbf{e}_{ij} = \mathbf{c}_j - \mathbf{c}_i$ and denote by $\delta_{ij} = \langle \mathbf{e}_{ij}, \mathbf{e}_{ij} \rangle$ their squared norms. For a direction $\mathbf{u} \in \mathbb{R}^3 \setminus \{(0,0,0)\}$, we put:

$$q = q(\mathbf{u}) = \langle \mathbf{u}, \mathbf{u} \rangle,$$

$$t_{ij} = t_{ji} = \langle \mathbf{e}_{ij} \times \mathbf{u}, \mathbf{e}_{ij} \times \mathbf{u} \rangle = \delta_{ij} q - \langle \mathbf{e}_{ij}, \mathbf{u} \rangle^2.$$

Thus in $\mathbb{P}^2(\mathbb{C})$, the equation $t_{ij} = 0$ gives the two tangents from e_{ij} to the imaginary conic $q = 0$.

Proposition 2 *The direction-sextic for B_0, B_1, B_2 can be given by means of the Cayley determinant*:

$$\sigma = \sigma(\mathbf{u}) = \det \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & qs_0 & qs_1 & qs_2 \\ 1 & qs_0 & 0 & t_{01} & t_{02} \\ 1 & qs_1 & t_{01} & 0 & t_{12} \\ 1 & qs_2 & t_{02} & t_{12} & 0 \end{pmatrix}.$$

Proof One way to find the equation of the direction curve is to begin with a description of lines in $\mathbb{R}^3$ by parameters $(\mathbf{p}, \mathbf{u}) \in \mathbb{R}^3 \times \mathbb{P}^2$, where $\mathbf{p}$ is the orthogonal projection of the origin on the given line, and $\mathbf{u}$ is the direction of the line. With $\mathbf{c}_0 = 0$ and abbreviations:

$$a_i = a_i(\mathbf{u}) = \langle \mathbf{c}_i \times \mathbf{u}, \mathbf{c}_i \times \mathbf{u} \rangle + (s_0 - s_i)\langle \mathbf{u}, \mathbf{u} \rangle = t_{0i} + (s_0 - s_i)q, \quad i = 1, 2,$$

[3]The complement of any proper non-empty conic in the real projective plane consists of two connected components, one homeomorphic to a Möbius strip and the other to a disc.

affine common tangents obey the system (see e.g. [3] or [17]):

$$\langle \mathbf{p}, \mathbf{c}_i \rangle = \frac{a_i(\mathbf{u})}{2\langle \mathbf{u}, \mathbf{u} \rangle}, \quad i = 1, 2, \qquad \langle \mathbf{p}, \mathbf{u} \rangle = 0, \qquad \langle \mathbf{p}, \mathbf{p} \rangle = s_0.$$

The direction-sextic is obtained by eliminating $\mathbf{p}$ from this system. The fact that the resulting equation allows the stated Cayley determinant expression is given a natural explanation in [2], but can be directly verified by computation. □

The direction of an *oriented line* can be represented either by a point on the unit sphere or, by the whole *ray* emanating from the origin and passing through that point. Our expression "cone of directions" stems from the latter representation, which converts questions of convexity in $\mathbb{S}^2$ into equivalent questions of convexity in $\mathbb{R}^3$. In the projective context, it will be understood that we mean the image via $\mathbb{S}^2/\mathbb{Z}_2 = \mathbb{P}^2$.

2.3 Cone of Directions

The *cone of directions* $K(B_0B_1B_2)$ of B_0, B_1, B_2 is the set of directions of all oriented line transversals to these balls which meet them in the stated order: $B_0 \prec B_1 \prec B_2$. The boundary of $K(B_0B_1B_2)$ consists of [6, Lemma 9] certain arcs of the direction-sextic σ and certain arcs of directions of *inner special bitangents* i.e. tangents to two of the balls passing through their inner similitude center [13]. Figure 1 offers an illustration of a cone of directions. The plane of the picture must be conceived as an affine piece $\mathbb{R}^2 \subset \mathbb{P}^2$.

We recall the fact that a common tangent (here called bitangent) for two disjoint spheres (more precisely, the boundary of two disjoint balls) passes through their inner similitude center if and only if it is contained in a common tangent plane which has the two spheres on opposite sides. If a transversal for the two balls has the direction of an inner special bitangent, it must actually be that bitangent. The cone of directions

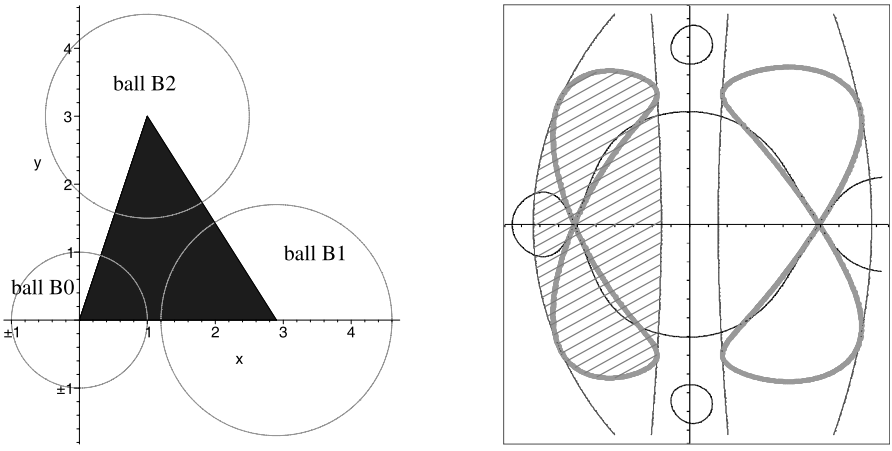


Fig. 1 *Left*: The trace of three balls B_0, B_1, B_2 on their plane of centers. *Right*: A planar depiction (hatched area) of $\mathcal{K}(B_1B_0B_2)$. The direction-sextic is drawn in *thick grey*, the Hessian in *black*, and the conics of inner special bitangents in *thin grey*

for a pair of disjoint balls is bounded precisely by their inner special bitangents. In $\mathbb{P}^2$ they trace a (circular) conic.

The points of σ that appear on the boundary $\partial K(B_0B_1B_2)$ can be characterized as follows:

Proposition 3 *The direction of a tritangent ℓ meeting the three balls B_0, B_1, B_2 in the prescribed order belongs to $\partial K(B_0B_1B_2)$ if and only if ℓ intersects the triangle of centers $\mathbf{c}_0\mathbf{c}_1\mathbf{c}_2$.*

Proof The set of directions of common transversals to disjoint balls is a proper subset of $\mathbb{P}^2$.

Assume that ℓ is neither parallel to the plane of centers, nor contained in it.

If ℓ does not intersect the triangle of centers, then, in the projected configuration on $\ell^\perp$, there is a line λ through two of the projected centers, separating the foot of ℓ from the third projected center. When moving ℓ parallel to itself and closer to λ, along a perpendicular to the latter, all distances to centers decrease. This shows that there are lines parallel to ℓ intersecting the open balls, and therefore the direction of ℓ is not on the boundary.

On the other hand, when the tritangent ℓ intersects the triangle of centers in a point P, there is no motion of ℓ parallel to itself which can decrease all distances to the centers. Indeed, reasoning in $\ell^\perp$ with respect to the triangle of projected centers, this would decrease all areas over edges, while these areas have a constant sum. This shows that no other transversal but ℓ can have its direction.[4] Looking now in the plane spanned by ℓ and the normal ν to the plane of centers at P, the rotation of ℓ, with center P, brings its direction inside $K(B_0B_1B_2)$ when approaching the plane of centers, and takes it outside $K(B_0B_1B_2)$ when approaching ν. Indeed, when rotating towards the plane of centers all distances to centers decrease, while increasing in the opposite sense. Some other transversal with direction between ℓ and ν (and parallel to the ℓ, ν-plane) cannot exist since by the same argument of rotating towards the plane of centers, one would obtain a realization of the direction of ℓ not passing through P. Thus, the direction of ℓ is in $\partial K(B_0B_1B_2)$.

If ℓ is parallel to the plane of centers (but not contained in it), we may consider any parallel plane which is closer to $\mathbf{c}_0\mathbf{c}_1\mathbf{c}_2$ than ℓ is, and find in this plane transversals to the open balls parallel to ℓ. Thus, ℓ cannot be on the boundary.

Finally, if ℓ is in the plane of centers, we look at the "section configuration" traced in that plane. *Either* all three discs are on one side of ℓ and then ℓ does not cross the triangle of centers and is not on the boundary, *or* ℓ has two discs on one side with the third on the other side and must cross the triangle of centers. Then, it is actually an inner special bitangent for two pairs of balls (and an outer special bitangent for the third pair) and belongs to the boundary. □

[4] One could conclude from here using [6, Lemma 9], which shows that a direction of $K(B_0B_1B_2)$ is in the interior if and only if there is a line transversal to the open balls with that direction.

Proposition 4 *For three disjoint balls, we have*:

(i) *The cone of directions $K(B_0B_1B_2)$ consists of a single point if and only if there is a tritangent contained in the plane of centers and tracing in it a pinned planar configuration, that is, the disc traced by B_1 is on the opposite side of the tritangent from the discs traced by B_0 and B_2;*
(ii) *In all other cases, the cone of directions $K(B_0B_1B_2)$ is the closure of its interior.*

Proof (i) Sufficiency: the plane intersecting the plane of centers along the tritangent and perpendicular to it, will have B_1 on one side, and B_0 and B_2 on the other. An oriented transversal meeting B_0 first, then B_1, and then B_2 must be contained in this separating perpendicular plane, and thus coincide with the given tritangent. Necessity is covered by our arguments in (ii).

(ii) Suppose we are not in case (i), and the centers are not aligned. If we have a transversal ℓ with direction belonging to the boundary of $K(B_0B_1B_2)$, we may assume the transversal is not in the plane of centers, since a non-pinned planar case is clear. But then ℓ and its reflection in the plane of centers define a plane perpendicular to the latter and all lines between them (passing through their intersection) have directions belonging to the interior, because all distances from centers decrease.

The case of collinear centers is trivial; there is only one geometric permutation (given by the line of centers) and the cone of directions is a disc-like region bounded by a conic. □

Corollary of the proof Cones of directions and connected components of transversals for three disjoint balls in $\mathbb{R}^3$ are *contractible*.

Indeed, the argument above shows that we may contract first to the segment in $K(B_0B_1B_2)$ consisting of directions in the plane of centers, and then contract this segment.

Obviously, the same holds true at the level of the connected components in the space of transversals.

2.4 Hessian and Flexes

The *Hessian* of the direction-sextic σ is defined as the determinant of the matrix of second derivatives:

$$H(\sigma) = H(\sigma)(\mathbf{u}) = \det\left(\frac{\partial^2\sigma}{\partial u_i \partial u_j}\right).$$

The Hessian curve, or simply "the Hessian", is the projective curve defined by the zero-set of this determinant.

The Hessian of a direction-sextic for three balls in $\mathbb{R}^3$ is thus an algebraic curve of degree twelve. The intersection between σ and its Hessian $H(\sigma)$ consists of all singular points of σ and all flexes of σ [4].

3 Outline of the Proof

For $d = 2$ the convexity theorem is elementary, and for $d \geq 3$ it is easily reduced to the case of three disjoint balls in $\mathbb{R}^3$. The *key property* used to settle this case is the following:

Proposition 5 *For disjoint balls B_0, B_1, B_2, any arc of their direction-sextic σ which belongs to the boundary $\partial K(B_0B_1B_2)$ contains no flex or singularity of σ between its endpoints.*

The convexity of the cone of directions $K(B_0B_1B_2)$ can then be inferred from the known fact that a simple C^1-loop in $\mathbb{R}^2 \subset \mathbb{P}^2$ with no inflection (in Euclidean terms: with positive curvature on its algebraic arcs) bounds a convex interior [23].

Thus, what is essential for this approach, is to obtain sufficient control over the flexes of σ. At first sight, the fact that the intersection of σ and the Hessian $H(\sigma)$ in $\mathbb{P}^2(\mathbb{C})$ has, counting multiplicities, $6 \times 12 = 72$ points, leaves little hope for the possibility of "tracking" all flexes. However, there is another way to exploit the Hessian: fix a direction and consider the ball configurations which have a tritangent with that direction and give the same planar configuration of four points when projecting, tangent and centers, on some orthogonal plane; evaluate the Hessians of the corresponding direction-sextics and determine which can vanish for the given direction.

The important point is that one can anticipate, from the form of the equations, that the computations must result in polynomials of low degree, which will be subject, in their turn, to geometric control.

The unfolding of this scenario is presented below and involves a certain amount of explicit computations. Although no part is too complicated to be done by hand, we have relied on Maple [18] in a few instances.

4 Absence of Flexes and Singularities

4.1 The Hessian Test

Following Proposition 3, we need only consider directions of tangents to the three balls that cross the triangle of centers and are not directions of inner special bitangents. When projecting along such a tangent on a perpendicular plane, the projected centers form a triangle containing the point image of the tangent as an interior point. One may start with the latter planar configuration, a triangle and an interior point, and ask which ball configurations yield this picture (by projection along a common tangent intersecting at the interior point)? Since the radii of the balls are given, one has only to "lift" the vertices of the triangle in the normal direction and obtain all the desired configurations.

We equip $\mathbb{R}^3$ with a coordinate frame such that the triangle lies in the plane $\mathbf{e}_3^\perp \subset \mathbb{R}^3$ and has its vertices at $\tilde{\mathbf{c}}_0 = 0, \tilde{\mathbf{c}}_1, \tilde{\mathbf{c}}_2$, with the understanding that there is a point inside, with squared distances s_i to these vertices. Then, we use three real parameters, x_0, x_1 and x_2, to describe the possible positions of the three centers:

$$\mathbf{c}_0 = \tilde{\mathbf{c}}_0 + x_0\mathbf{e}_3, \qquad \mathbf{c}_1 = \tilde{\mathbf{c}}_1 + x_1\mathbf{e}_3, \qquad \mathbf{c}_2 = \tilde{\mathbf{c}}_2 + x_2\mathbf{e}_3.$$

We use Proposition 2 to express the corresponding direction-sextic σ and its Hessian $H(\sigma)$ as functions of $\mathbf{x} = (x_0, x_1, x_2) \in \mathbb{R}^3$ depending on $\tilde{\mathbf{c}}_0, \tilde{\mathbf{c}}_1, \tilde{\mathbf{c}}_2, s_0, s_1, s_2$. Proposition 5 is now equivalent to proving that

$$H(\sigma)(0, 0, 1) \neq 0$$

holds for all initial data (triangle and interior point) and all (x_0, x_1, x_2) corresponding to disjoint balls.

4.2 A Quadric and a Quartic

We have reduced the probe for flexes to the study of a polynomial function of $\mathbf{x}$ (and parameters) which can be explicitly computed.

The parameters involved are the following:

$$\tilde{\mathbf{c}}_0 = (0, 0, 0), \qquad \tilde{\mathbf{c}}_1 = (a, 0, 0), \qquad \tilde{\mathbf{c}}_2 = (b, c, 0),$$

the triangle of centers $(\tilde{\mathbf{c}}_0, \tilde{\mathbf{c}}_1, \tilde{\mathbf{c}}_2)$ having interior point:

$$p = \frac{\sum p_i \tilde{\mathbf{c}}_i}{\sum p_i} = \frac{p_1 \tilde{\mathbf{c}}_1 + p_2 \tilde{\mathbf{c}}_2}{\sum p_i}, \quad p_0, p_1, p_2 > 0.$$

Let $\mathbf{v}_k = \mathbf{p} - \tilde{\mathbf{c}}_k$. Then $s_k = r_k^2 = \langle \mathbf{v}_k, \mathbf{v}_k \rangle$.

The computation gives the result:

$$H(\sigma)(0, 0, 1) = \frac{2^{12} 5^2 a^6 c^6}{(\sum p_i)^5} [H_2(\mathbf{x}) + H_4(\mathbf{x})],$$

where H_2 and H_4 have degree respectively 2 and 4 in $\mathbf{x} = (x_0, x_1, x_2)$:

$$H_2 = H_2(\mathbf{x}) = -a^2 c^2 \left(\prod p_k \right) \sum p_i p_j (x_i - x_j)^2,$$

$$H_4 = H_4(\mathbf{x}) = \sum p_k^3 s_k (x_i - x_k)^2 (x_j - x_k)^2,$$

with cyclic products and sums for $\{i, j, k\} = \{0, 1, 2\}$. Thus, away from $(0, 0, 0)$, H_2 is negative and H_4 is positive. The aim is now to show that ball disjointness is enough to ensure the positivity of $H_2 + H_4$.

4.3 Hyperboloid and Octant

We can further transform these expressions by retaining as parameters the (positive numbers) p_i and $q_j = p_j r_j$, and renaming the squares $z_k = (x_i - x_j)^2$. This gives:

$$H_2 = H_2(\mathbf{z}) = -a^2 c^2 \left(\prod p_k \right) \sum p_i p_j z_k,$$

$$H_4 = H_4(\mathbf{z}) = \sum p_k q_k^2 z_i z_j.$$

From now on, assume that $\sum p_i = 1$. We have to replace $\Delta = a^2 c^2$, which is four times the squared area of the triangle $\tilde{\mathbf{c}}_0, \tilde{\mathbf{c}}_1, \tilde{\mathbf{c}}_2$, by its expression in terms of p_i and q_j.

Lemma 6 *We have*:

$$\Delta = a^2 c^2 = \frac{Q}{4 \prod p_k^2}, \quad \textit{with } Q = \sum (2 q_i^2 q_j^2 - q_k^4).$$

Proof This is an elementary computation, which may be conducted as follows. By the definition of $\mathbf{v}_i$, we have

$$\sum p_i \mathbf{v}_i = 0.$$

From $\langle \sum p_i \mathbf{v}_i, \mathbf{v}_j \rangle = 0$, we obtain a linear system for $\langle \mathbf{v}_i, \mathbf{v}_j \rangle$, $i \neq j$:

$$p_i \langle \mathbf{v}_i, \mathbf{v}_k \rangle + p_j \langle \mathbf{v}_j, \mathbf{v}_k \rangle = -p_k \langle \mathbf{v}_k, \mathbf{v}_k \rangle = -p_k s_k,$$

with solutions:

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle = \frac{p_k^2 s_k - p_i^2 s_i - p_j^2 s_j}{2 p_i p_j} = \frac{q_k^2 - q_i^2 - q_j^2}{2 p_i p_j}.$$

Four times the squared area of a triangle $\mathbf{p}, \tilde{\mathbf{c}}_i, \tilde{\mathbf{c}}_j$ is a Gram determinant:

$$\begin{vmatrix} \langle \mathbf{v}_i, \mathbf{v}_i \rangle & \langle \mathbf{v}_i, \mathbf{v}_j \rangle \\ \langle \mathbf{v}_i, \mathbf{v}_j \rangle & \langle \mathbf{v}_j, \mathbf{v}_j \rangle \end{vmatrix} = s_i s_j - \langle \mathbf{v}_i, \mathbf{v}_j \rangle^2 = \frac{Q}{4 p_i^2 p_j^2},$$

where $Q = \sum (2 q_i^2 q_j^2 - q_k^4)$. Hence the area of the triangle $\tilde{\mathbf{c}}_0, \tilde{\mathbf{c}}_1, \tilde{\mathbf{c}}_2$ is:

$$\frac{1}{4} Q^{1/2} \sum \frac{1}{p_i p_j} = \frac{Q^{1/2}}{4 \prod p_k},$$

resulting in:

$$\Delta = a^2 c^2 = \frac{Q}{4 \prod p_k^2}.$$ □

Several new substitutions will be in order for the study of $H_2 + H_4$. Since a positive factor won't affect sign considerations, we will use the symbol $*H$ for any positive multiple of $H_2 + H_4$. We have found above:

$$*H = *H(\mathbf{z}) = -\frac{1}{4} Q \sum \frac{z_k}{p_k} + \sum p_k q_k^2 z_i z_j,$$

with the shorthand $Q = \sum (2 q_i^2 q_j^2 - q_k^4)$. We put $p_i p_j z_k = q_k^2 w_k$ and obtain, up to a positive factor:

$$*H = *H(\mathbf{w}) = -\frac{1}{4} Q \sum q_k^2 w_k + \prod q_k^2 \sum w_i w_j.$$

With one more positive rescaling, and $a_k = \frac{Q}{4 q_i^2 q_j^2}$, we have:

$$*H = *H(\mathbf{w}) = \sum w_i w_j - \sum a_k w_k.$$

We can turn now to the conditions expressing the fact that the spheres with centers $\mathbf{c}_i = \tilde{\mathbf{c}}_i + x_i \mathbf{e}_3$ and radii r_i are disjoint. They are:

$$z_k = (x_i - x_j)^2 > (r_i + r_j)^2 - \delta_{ij} = (r_i + r_j)^2 - \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_i - \mathbf{v}_j \rangle,$$

that is,

$$z_k > \frac{q_k^2 - (q_i - q_j)^2}{p_i p_j}.$$

In **w**-coordinates, the "*disjointness conditions*" become

$$w_k > 1 - \left(\frac{q_i - q_j}{q_k}\right)^2.$$

Note that from $\sum p_i \mathbf{v}_i = 0$ it follows that $q_k = \|p_i \mathbf{v}_i\| > 0$ are the lengths of the three edges in a triangle, and therefore the latter expressions are positive by the triangle inequality.

The purpose now is to study the position of the octant defined by the disjointness conditions relative to the affine quadric in $\mathbb{R}^3$ defined by $*H(\mathbf{w}) = 0$. We use first a translation by β, in order to absorb the linear part in $*H$:

$$*H = *H(\mathbf{w}) = \sum (w_i - \beta_i)(w_j - \beta_j) - \sum \beta_i \beta_j,$$

with β respecting:

$$\beta_i + \beta_j = a_k, \quad \text{that is } \beta_k = \frac{1}{2}(a_i + a_j - a_k).$$

This makes

$$\sum \beta_i \beta_j = \frac{1}{4} \sum (a_k + a_i - a_j)(a_k - a_i + a_j) = \frac{1}{4} \sum (2a_i a_j - a_k^2),$$

and results in

$$\sum \beta_i \beta_j = \frac{1}{4} \left(\frac{Q}{4 \prod q_k^2}\right)^2 \sum (2q_i^2 q_j^2 - q_k^4) = \frac{Q^3}{4^3 \prod q_k^4} > 0.$$

Thus, with translated coordinates $t_k = w_k - \beta_k$ we have a *hyperboloid of two sheets*:

$$*H = *H(\mathbf{t}) = \sum t_i t_j - \frac{Q^3}{4^3 \prod q_k^4} = 0,$$

which lies on the positive side of its asymptotic cone $\sum t_i t_j = 0$.

Lemma 7 *$\sum t_i t_j = 0$ is a circular cone with axis $t_0 = t_1 = t_2$. The two components of its smooth points circumscribe the positive and negative open octants, which are both contained in the positive part $\sum t_i t_j > 0$.*

The open octant defined by our disjointness conditions $w_k > 1 - (\frac{q_i - q_j}{q_k})^2$ is a translate of the open positive octant, and its position relative to the hyperboloid $*H(\mathbf{w}) = 0$ is determined by the position of its vertex $\mathbf{V}$. Continuing to refer here to **w**-coordinates, we have:

Lemma 8 *The point* $\mathbf{V} = (1 - (\frac{q_i - q_j}{q_k})^2)_{0 \le k \le 2}$ *is on the "positive side" of the hyperboloid* $*H(\mathbf{w}) = 0$ *and on the "positive side" of the plane* $\sum t_k = \sum (w_k - \beta_k) = 0$, *that is*:

$$*H(\mathbf{V}) > 0 \quad \textit{and} \quad \sum \left(1 - \left(\frac{q_i - q_j}{q_k}\right)^2\right) > \frac{Q}{8 \prod q_k^2} \sum q_k^2.$$

Proof A Maple assisted computation shows that $*H(\mathbf{V})$ factors as

$$*H(\mathbf{V}) = \frac{3 \prod (q_i + q_j - q_k)^2}{4 \prod q_k^2},$$

from which the first inequality follows.

The second inequality, which determines on which of the two components of the positive side of the hyperboloid $\mathbf{V}$ lies, is satisfied for $q_0 = q_1 = q_2$, and by continuity, must be satisfied for any other triangle edges, since vertex $\mathbf{V}$ cannot "jump" from one component to the other. □

It is now clear, geometrically, that the octant where the disjointness conditions are satisfied and the hyperboloid indicating a flex or a singularity for the corresponding configuration *have no point in common*. This completes the proof of Proposition 5.

5 Convexity of the Cone for 3 Balls in $\mathbb{R}^3$

We consider now three *disjoint* closed balls B_0, B_1, B_2 described by parameters: centers $\mathbf{c}_0, \mathbf{c}_1, \mathbf{c}_2$ and radii r_0, r_1, r_2. We shall prove first the convexity of any cone of directions in the *generic* case i.e. when the centers and radii are in the complement of a proper algebraic subset. Then, we will show that the generic case implies the general case.

Lemma 9 *The direction cone* $K(B_0 B_1 B_2)$ *of a generic triple of disjoint balls in* $\mathbb{R}^3$ *is strictly convex.*

Proof If $\partial K(B_0 B_1 B_2)$ is made only of directions of inner special bitangents, strict convexity is immediate, since $K(B_0 B_1 B_2)$ is then an intersection of convex regions bounded by conics. Otherwise, genericity allows us to assume that the direction-sextic σ is non-singular at all its contacts with any of the three conics determined by inner special tangents. Since the direction-sextic necessarily lies on the simply-connected side of each of the three conics, these contacts are tangency points at which $\partial K(B_0 B_1 B_2)$ is locally convex. Thus, if we start at some point of $\partial K(B_0 B_1 B_2)$ and follow the boundary curve, we obtain, by Proposition 5, a differentiable simple loop of class C^1, which is, locally, always on the same side of its tangent. For any affine plane $\mathbb{R}^2 \subset \mathbb{P}^2$ covering the loop, and any Euclidean metric in it, this means positive curvature on all its algebraic arcs and this implies [23] that our simple loop bounds a compact convex set. In fact *strictly convex*, because of non-vanishing curvature. By Proposition 4 and its Corollary, this strictly convex set is $K(B_0 B_1 B_2)$. □

The passage from the generic case to the general case is based on:

Lemma 10 *Let $\mathcal{B} = (B_0, B_1, B_2)$ be a configuration of three disjoint closed balls, and suppose $K(B_0B_1B_2)$ has non-empty interior. If $\mathcal{B}$ is the limit of a sequence of configurations $\mathcal{B}^{(\nu)}$ with a convex cone of directions for the given ordering, then $K(B_0B_1B_2)$ is convex as well.*

Proof By Proposition 4, it is enough to prove that, for any two points in the interior, the (geodesic) segment joining them is contained in $K(B_0B_1B_2)$.

Take two interior points. By assumption, for sufficiently large ν, the segment joining them is contained in all corresponding cones for $\mathcal{B}^{(\nu)}$. Consider one point of the segment, and project the sphere configuration along the direction defined by the point, on a perpendicular plane. We have to prove that the disks representing the projected balls have at least one point in common.

Suppose they don't. Then so would discs with the same centers and radii increased by a small $\epsilon > 0$. But then we can find, for sufficiently large ν, configurations $\mathcal{B}^{(\nu)}$ with centers projecting less than $\epsilon/2$ away from those of $\mathcal{B}$ and corresponding radii with less than $\epsilon/2$ augmentation. Then the point of the segment cannot be in the respective cones of directions, a contradiction.

Note that strict convexity still follows from non-zero curvature on smooth arcs for non-collinear centers, while for collinear centers it is obvious because of rotational symmetry. □

Lemmas 9 and 10 immediately imply Theorem 1 for the case of three balls in $\mathbb{R}^3$:

Proposition 11 *The directions of all oriented lines intersecting three disjoint balls in $\mathbb{R}^3$ in a specific order form a strictly convex subset of the sphere $\mathbb{S}^2$.*

6 Convexity of the Cone for n Balls in $\mathbb{R}^d$

The convexity result of Proposition 11 generalizes to arbitrary n and d as follows:

Proof of Theorem 1 Recall that, for any collection of balls in $\mathbb{R}^3$, a direction will be realized by some transversal if and only if the orthogonal projection of the balls on a perpendicular plane has non-empty intersection. By Helly's Theorem in the plane, the direction cone for a sequence of $n \geq 3$ balls is the intersection of the direction cones of all its triples. Thus, the direction cone of n ordered 3-dimensional disjoint balls is strictly convex for any n.

Given a sequence $\mathcal{S}$ of n disjoint balls in $\mathbb{R}^d$, let K be its direction cone for a prescribed order of intersection. Let $\mathbf{u}$ and $\mathbf{v}$ be two directions in K, $\ell_{\mathbf{u}}$ and $\ell_{\mathbf{v}}$ be two corresponding line transversals and let E denote the 3-dimensional affine space these two lines span (or a 3-space containing their planar span, should the lines be coplanar).

$E \cap \mathcal{S}$ is a collection of 3-dimensional disjoint balls whose corresponding direction cone is convex on $\mathbb{S}^2$. Thus, for any direction on the small arc of great circle joining $\mathbf{u}$ and $\mathbf{v}$ there exists an order-respecting transversal to $\mathcal{S}$, because it already exists

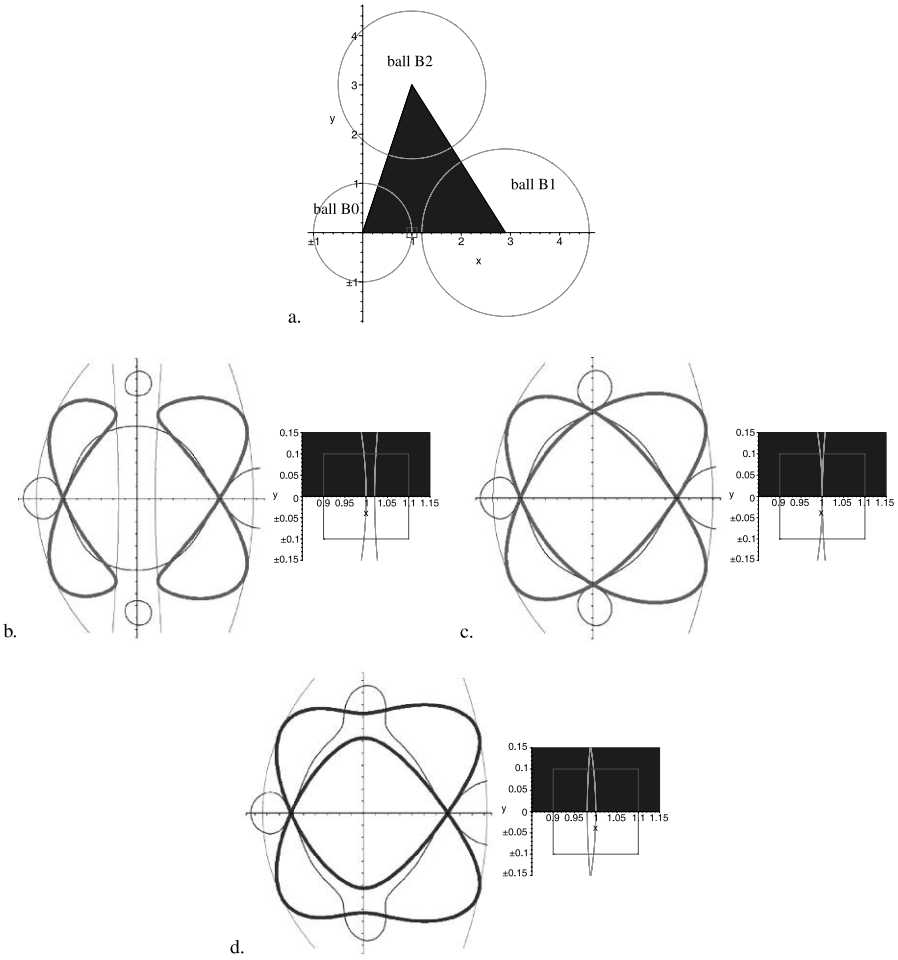


Fig. 2 **a** The trace of three disjoint balls *on the plane of centers*, with ball B_1 moving *on the horizontal axis* towards ball B_0. *The small square* is used for close-ups below. **b**, **c**, **d** The direction-sextic (*in thick gray*), its Hessian (*in black*) and arcs of inner special bitangent conics, when balls B_0 and B_1 are disjoint (**b**), tangent (**c**) and intersecting (**d**)

in E. It follows that K is convex, and again, from the three dimensional case, strictly convex. □

Let us emphasize the importance of the assumption that the balls are disjoint. Figure 2 illustrates a transition from convex to non-convex direction cones as three disjoint balls move and allow an overlap.

7 Implications

This section explores some consequences of Theorem 1. Similar results were proven for the case of unit balls in [6] and, with Theorem 1, the proofs carry through. We thus omit all arguments here and point to the relevant lemmata in [6].

7.1 Isotopy and Geometric Permutations

An immediate corollary of Theorem 1 is the correspondence of isotopy and geometric permutations for line transversals to disjoint balls:

Corollary 12 *The set of line transversals to n disjoint balls in $\mathbb{R}^d$ realizing the same geometric permutation is contractible.*

The proof given by Cheong et al. [6, Lemma 14] for disjoint unit balls immediately extends, with Theorem 1, to the case of disjoint balls.

Smorodinsky et al. [21] showed that in the worst case n disjoint balls in $\mathbb{R}^d$ admit $\Theta(n^{d-1})$ geometric permutations. The same bound thus applies for the number of connected components of line transversals, improving on the previous bounds of $O(n^{3+\epsilon})$ for $d = 3$ and of $O(n^{2d-2})$ for $d \geq 4$ due to Koltun and Sharir [16]. If the radii of the balls are in some interval $[1, \gamma]$ where γ is independent of n and d, then the number of components of transversals is $O(\gamma^{\log \gamma})$, following the bound on the number of geometric permutations obtained by Zhou and Suri [24]. These results are summarized as follows:

Corollary 13 *In the worst case, n disjoint balls in $\mathbb{R}^d$ have $\Theta(n^{d-1})$ connected components of line transversals. If the radii of the balls are in the interval $[1, \gamma]$, where γ is independent of n and d, this number becomes $O(\gamma^{\log \gamma})$.*

7.2 Minimal Pinning Configurations

A *minimal pinning configuration* is a collection of objects having an isolated line transversal that ceases to be isolated if any of the objects is discarded. An important step in the proof of Hadwiger's transversal theorem [11] is the observation that, in the plane, any minimal pinning configuration consisting of disjoint convex objects has cardinality 3. Cheong et al. [6, Proposition 13] proved that any minimal pinning configuration consisting of disjoint unit balls in $\mathbb{R}^d$ has cardinality at most $2d - 1$. With Theorem 1, the same holds for disjoint balls of arbitrary radii:

Corollary 14 *Any minimal pinning configuration consisting of disjoint balls in $\mathbb{R}^d$ has cardinality at most $2d - 1$.*

7.3 A Hadwiger-Type Result

A result in the flavor of Hadwiger's Transversal Theorem [6, Theorem 1] generalizes to disjoint balls of arbitrary radii:

Corollary 15 *A sequence of n disjoint balls in $\mathbb{R}^d$ has a line transversal if any subsequence of size at most $2d$ has an order-respecting line transversal.*

The "pure" generalizations [6, 14] of Helly's theorem, i.e. without additional constraints on the ordering à la Hadwiger, use the fact that $n \geq 9$ disjoint unit balls have at most 2 geometric permutations [5]. Since the latter is not true for balls of arbitrary radii [21], obtaining a Helly-type theorem for line transversals in this case requires different arguments.

References

1. Borcea, C.: Algebraic geometry for constraint problems. In: Pae, S.i., Park, H. (eds.) Proc. 7th Asian Symposium in Computer Math., pp. 112–114 (2005)
2. Borcea, C.: Involutive sextics and tangents to spheres (2006, manuscript)
3. Borcea, C., Goaoc, X., Lazard, S., Petitjean, S.: Common tangents to spheres in $\mathbb{R}^3$. Discrete Comput. Geom. **35**(2), 287–300 (2006)
4. Brieskorn, E., Knörrer, H.: Plane Algebraic Curves. Birkhaüser, Basel (1986)
5. Cheong, O., Goaoc, X., Na, H.-S.: Geometric permutations of disjoint unit spheres. Comput. Geom. Theory Appl. **30**, 253–270 (2005)
6. Cheong, O., Goaoc, X., Holmsen, A., Petitjean, S.: Hadwiger and Helly-type theorems for disjoint unit spheres. Discrete Comput. Geom. (2008, to appear). Special issue for the 20th anniversary of the journal
7. Danzer, L.: Über ein Problem aus der kombinatorischen Geometrie. Arch. Math. **8**, 347–351 (1957)
8. Goodman, J.E., Pollack, R., Wenger, R.: Geometric transversal theory. In: Pach, J. (ed.) New Trends in Discrete and Computational Geometry. Algorithms and Combinatorics, vol. 10, pp. 163–198. Springer, Heidelberg (1993)
9. Hadwiger, H.: Problem 107. Nieuw Arch. Wiskd. **3**, 4–57 (1956)
10. Hadwiger, H.: Solution. Wiskd. Opg. **20**, 27–29 (1957)
11. Hadwiger, H.: Über Eibereiche mit gemeinsamer Treffgeraden. Port. Math. **6**, 23–29 (1957)
12. Helly, E.: Über Mengen konvexer Körper mit gemeinschaftlichen Punkten. Jahresber. Dtsch. Math. Verein. **32**, 175–176 (1923)
13. Hilbert, D., Cohn-Vossen, S.: Geometry and the Imagination. Chelsea, New York (1952)
14. Holmsen, A., Katchalski, M., Lewis, T.: A Helly-type theorem for line transversals to disjoint unit balls. Discrete Comput. Geom. **29**, 595–602 (2003)
15. Holmsen, A., Matoušek, J.: No Helly theorem for stabbing translates by lines in $\mathbb{R}^d$. Discrete Comput. Geom. **31**, 405–410 (2004)
16. Koltun, V., Sharir, M.: The partition technique for overlays of envelopes. SIAM J. Comput. **32**, 841–863 (2003)
17. Macdonald, I.G., Pach, J., Theobald, T.: Common tangents to four unit balls in $\mathbb{R}^3$. Discrete Comput. Geom. **26**, 1–17 (2001)
18. The maple system. Waterloo maple software. http://www.maplesoft.com
19. Pach, J., Sharir, M.: Combinatorial Geometry with Algorithmic Applications—The Alcala Lectures, Alcala, Spain, August 31–September 5, 2006
20. Platonova, O.A.: Singularities of the mutual disposition of a surface and a line. Russ. Math. Surv. **36**, 248–249 (1981)
21. Smorodinsky, S., Mitchell, J.S.B., Sharir, M.: Sharp bounds on geometric permutations for pairwise disjoint balls in $\mathbb{R}^d$. Discrete Comput. Geom. **23**, 247–259 (2000)
22. Sottile, F., Theobald, T.: Line problems in nonlinear computational geometry. ArXiv math.MG 0610407 (2006)
23. Toponogov, V.: Differential Geometry of Curves and Surfaces: A Concise Guide. Birkhäuser, Basel (2006)
24. Zhou, Y., Suri, S.: Geometric permutations of balls with bounded size disparity. Comput. Geom. Theory Appl. **26**, 3–20 (2003)

Norm Bounds for Ehrhart Polynomial Roots

Benjamin Braun

Abstract M. Beck et al. found that the roots of the Ehrhart polynomial of a d-dimensional lattice polytope are bounded above in norm by $1 + (d + 1)!$. We provide an improved bound which is quadratic in d and applies to a larger family of polynomials.

Keywords Lattice polytopes · Polynomial roots · Ehrhart theory

Let P be a convex polytope in R^n with vertices in Z^n and affine span of dimension d; we refer to such polytopes as *lattice polytopes* and to elements of Z^n as *lattice points*. A remarkable theorem due to Ehrhart [5] is that the number of lattice points in the tth dilate of P, for non-negative integers t, is given by a polynomial in t of degree d called the *Ehrhart polynomial* of P. We denote this polynomial by $L_P(t)$, and let $\mathrm{Ehr}_P(x) = \sum_{t\geq 0} L_P(t)x^t$ denote its associated rational generating function. For more information regarding Ehrhart theory, see [2].

In [1] it was shown that for a lattice polytope P of dimension d, the roots of $L_P(t)$ are bounded above in norm by $1 + (d + 1)!$. However, the authors suggested that a bound that is polynomial in d should exist and questioned whether this is a property of Ehrhart polynomials in particular or of a broader class of polynomials (see Remark 4.4 on p. 26 of [1]). Our answer is the following:

Theorem 1 *If f is a nonzero polynomial of degree d with real-valued, non-negative coefficients when expressed with respect to the polynomial basis*

$$B_d := \left\{ \binom{t+d-j}{d} : 0 \leq j \leq d \right\},$$

B. Braun (✉)
Department of Mathematics, Washington University in St. Louis, Campus Box 1146, St. Louis, MO 63130-4899, USA
e-mail: bjbraun@math.wustl.edu

then all the roots of f lie inside the disc with center $-1/2$ and radius $d(d-\frac{1}{2})$.

The link between this situation and Ehrhart polynomials is that for a polynomial f of degree d over the complex numbers, there always exist complex values h_j so that

$$\frac{\sum_{j=0}^{d} h_j x^j}{(1-x)^{d+1}} = \sum_{t\geq 0} f(t)x^t.$$

As a result, f can be expressed as

$$f(t) = \sum_{j=0}^{d} h_j \binom{t+d-j}{d}.$$

This is easily seen by expanding the rational function as a formal power series. We then apply the following theorem, originally due to Stanley:

Theorem 2 (See [7] and [2]) *If P is a d-dimensional lattice polytope with*

$$\mathrm{Ehr}_P(x) = \frac{\sum_{j=0}^{d} h_j x^j}{(1-x)^{d+1}},$$

then the h_j are non-negative integers.

Thus, our result applies to Ehrhart polynomials and more generally to Hilbert polynomials of certain Cohen–Macaulay modules (see Corollary 4.1.10 of [3]).

Proof of Theorem 1 Let d be a positive integer, let $D_d := \{z\colon |z+\frac{1}{2}| \leq d(d-\frac{1}{2})\}$, and let f be as given in the theorem. It is enough to show that for any complex number z not in D_d there exists an open half-plane with zero on the boundary containing $B_d(z) := \{\binom{z+d-j}{d}\colon\ 0 \leq j \leq d\}$, since this implies that $f(z)$ is a nontrivial, non-negative linear combination of elements in a common open half-plane and is hence nonzero.

Each element of $B_d(z)$ is given by the product of $1/d!$ and d consecutive members of $M := \{(z+d), (z+d-1), \ldots, (z-d+2), (z-d+1)\}$. The elements of M are contained in a disk $D(z)$ of diameter $2d-1$ centered at $z+\frac{1}{2}$. We claim that if $|z+\frac{1}{2}| > d(d-\frac{1}{2})$, which holds for $z \notin D_d$, then the angular width of $D(z)$ is less than $\frac{\pi}{d}$. To see this, consider one of the lines through the origin tangent to $D(z)$. The triangle formed by the origin, the point of tangency, and $z+\frac{1}{2}$ is a right triangle with hypotenuse of length $|z+\frac{1}{2}|$ and a side of length $d-\frac{1}{2}$ opposite the interior angle formed at the origin. Hence, the interior angle at the origin is $\sin^{-1}(d-\frac{1}{2}/|z+\frac{1}{2}|)$, and thus the total angular width of $D(z)$ is $2\sin^{-1}(d-\frac{1}{2}/|z+\frac{1}{2}|)$. Finally, we see that

$$2\sin^{-1}\left(\frac{d-\frac{1}{2}}{|z+\frac{1}{2}|}\right) < 2\sin^{-1}\left(\frac{d-\frac{1}{2}}{d(d-\frac{1}{2})}\right) = 2\sin^{-1}\left(\frac{1}{d}\right) < \frac{\pi}{d}.$$

Therefore, the elements of M all lie in a cone in the plane with apex the origin and angle width less than π/d. Thus, the angular difference between $(z+d-j)\cdots(z-j+1)$ and $(z+d-j-1)\cdots(z-j)$ is less than π/d for any j, $0 \le j < d$. Hence, $B_d(z)$ lies in an open half-plane and our proof is complete. □

All the polynomials in B_d have roots contained in $\{-d, -d+1, \ldots, d-1\}$. For $1 \le j \le d$, the number of polynomials in B_d with $-j$ as a root is equal to the number with $-1+j$ as a root. Thus, the location of the center of the disc in our theorem should not come as a surprise since the roots of the elements of B_d are highly symmetric with respect to the point $-1/2$. The line $x = -1/2$ also plays a prominent role for Ehrhart polynomials of cross-polytopes, as shown in [4] and [6].

It is interesting that our result only depends on f having a "nice" representation with respect to B_d. In our situation, the reason that B_d is better than the standard monomial basis is that each of the polynomials in B_d is of full degree d, and hence each such polynomial has d roots. By adapting our method, one can obtain root bounds for any polynomial in the non-negative real span of any basis for degree d polynomials containing only polynomials of degree d having positive real leading coefficients and known roots.

Acknowledgements Thanks to John Shareshian for suggestions and advice, Matthias Beck and Sinai Robins for introducing me to Ehrhart theory, an anonymous referee for thoughtful comments, and Laura Braun for support and encouragement.

References

1. Beck, M., De Loera, J., Develin, M., Pfeifle, J., Stanley, R.: Coefficients and roots of Ehrhart polynomials. In: Integer Points in Polyhedra—Geometry, Number Theory, Algebra, Optimization. Contemporary Mathematics, vol. 374, pp. 15–36. American Mathematical Society, Providence (2005). arxiv:math.CO/0402148
2. Beck, M., Robins, S.: Computing the Continuous Discretely. Springer, New York (2007)
3. Brunz, W., Herzog, J.: Cohen–Macaulay Rings. Cambridge University Press, Cambridge (1993)
4. Bump, D., Choi, K.-K., Kurlberg, P., Vaalar, J.: A local Riemann hypothesis. I. Math. Z. **233**(1), 1–19 (2000)
5. Ehrhart, E.: Sur les polyèdres rationnels homothétiques à n dimensions. C. R. Acad. Sci. Paris **254**, 616–618 (1962)
6. Rodriguez-Villegas, F.: On the zeros of certain polynomials. Proc. Am. Math. Soc. **130**(8), 2251–2254 (2002)
7. Stanley, R.: Decompositions of rational convex polytopes. Ann. Discrete Math. **6**, 333–342 (1980)

Helly-Type Theorems for Line Transversals to Disjoint Unit Balls

Otfried Cheong · Xavier Goaoc · Andreas Holmsen · Sylvain Petitjean

Abstract We prove Helly-type theorems for line transversals to disjoint unit balls in $\mathbb{R}^d$. In particular, we show that a family of $n \geq 2d$ disjoint unit balls in $\mathbb{R}^d$ has a line transversal if, for some ordering $\prec$ of the balls, any subfamily of $2d$ balls admits a line transversal consistent with $\prec$. We also prove that a family of $n \geq 4d - 1$ disjoint unit balls in $\mathbb{R}^d$ admits a line transversal if any subfamily of size $4d - 1$ admits a transversal.

Keywords Geometric transversal theory · Helly-type theorem · Hadwiger-type theorem · Spheres · Balls · Line transversal

1 Introduction

Helly's celebrated theorem, published in 1923, states that a finite family of convex sets in $\mathbb{R}^d$ has non-empty intersection if and only if any subfamily of size at most

Andreas Holmsen was supported by the Research Council of Norway, prosjektnummer 166618/V30. Otfried Cheong and Xavier Goaoc acknowledge support from the French-Korean Science and Technology Amicable Relationships program (STAR).

O. Cheong
Division of Computer Science, KAIST, Daejeon, South Korea
e-mail: otfried@kaist.ac.kr

X. Goaoc (✉)
LORIA–INRIA Lorraine, Nancy, France
e-mail: goaoc@loria.fr

A. Holmsen
Department of Mathematics, University of Bergen, Bergen, Norway
e-mail: andreash@mi.uib.no

S. Petitjean
LORIA–CNRS, Nancy, France
e-mail: petitjea@loria.fr

$d + 1$ has non-empty intersection. Subsequent results of similar flavor (that is, if every subset of size k of a set $\mathcal{S}$ has property $\mathcal{P}$ then $\mathcal{S}$ has property $\mathcal{P}$) have been called *Helly-type theorems* and the minimal such k is known as the associated *Helly number*. Helly-type theorems and tight bounds on Helly numbers have been the object of active research in combinatorial geometry. In this paper, we investigate Helly-type theorems for the existence of line transversals to a family of objects, i.e. lines that intersect every member of the family.

History The earliest Helly-type theorems in geometric transversal theory appeared about five decades ago. In 1957, Hadwiger [14] showed that an ordered family $\mathcal{S}$ of compact convex sets in the plane admits a line transversal if every triple admits a line transversal compatible with the ordering. (Note that a line transversal to $\mathcal{S}$ may not respect the ordering on $\mathcal{S}$; to prove the existence of a line transversal that respects the ordering on $\mathcal{S}$ one needs the assumption that any *four*-tuple admits an order-respecting line transversal.) In what follows, we shall talk about a Hadwiger-type theorem when the family of objects under consideration is ordered.

The same year, Danzer [6] proved the following result concerning families of pairwise disjoint unit discs in the plane: if such a family consists of at least 5 discs, and if any 5 of these discs are met by some line, then there exists a line meeting all the discs of the family. This answered a question of Hadwiger [11], who gave an example (5 circles, almost touching and with centers forming a regular pentagon) which shows that 5 cannot be replaced by 4. Grünbaum [9] showed that the same result holds if "unit disc" is replaced by "unit square", and conjectured that the result holds for families of disjoint translates of any compact convex set in the plane. This long-standing conjecture was finally proved by Tverberg [21]. A weaker form of the conjecture which assumed 128 instead of 5 had been established earlier by Katchalski [18].

Danzer [6] conjectured that Helly-type theorems exist for line transversals to disjoint unit balls in arbitrary dimension. The first positive result was obtained by Hadwiger [12, 13] for the case of families of "thinly distributed" balls, where the distance between any two balls is at least the sum of their radii. This result was extended by Ambrus et al. [1] to disjoint unit balls, in arbitrary dimension, the centers of which are at distance at least $2\sqrt{2+\sqrt{2}}$. Danzer's conjecture for three-dimensional disjoint unit balls, without additional assumption on their distribution, was only settled in 2001 by Holmsen et al. [17]. It should be stressed that in dimension three (and higher), neither Hadwiger nor Helly-type theorems exist for line transversals to general convex objects, not even for translates of a convex compact set [16].

In his paper [6], Danzer also asked whether the Helly number for line transversals to disjoint unit balls in $\mathbb{R}^d$ is a strictly increasing function of d. The only known lower bound is the planar example of Hadwiger [11]. This number was proved to be at most d^2 for thinly distributed balls in $\mathbb{R}^d$ by Hadwiger [12, 13], a bound improved to $2d - 1$ by Grünbaum [10] using the topological Helly theorem. For disjoint unit balls in dimension three, Holmsen et al. [17] proved bounds of respectively 12 and 46 for the Hadwiger-type and Helly-type theorems, which were later improved to 12 and 18 by Cheong et al. [5].

We refer the reader to the recent survey by Wenger [22] for a broader discussion of geometric transversal theory.

Our Results In this paper we complete the proof of Danzer's conjecture. More precisely, we show that Helly-type theorems exist for line transversals to families of *pairwise-inflatable* balls in $\mathbb{R}^d$. A family $\mathcal{F}$ of balls in $\mathbb{R}^d$ is called pairwise-inflatable if for every pair of balls $B_1, B_2 \in \mathcal{F}$ we have $\gamma^2 > 2(r_1^2 + r_2^2)$, where r_i is the radius of B_i, and γ is the distance between their centers. A family of disjoint unit balls is pairwise-inflatable, since $\gamma^2 > 2(r_1^2 + r_2^2)$ implies $\gamma > r_1 + r_2$ when $r_1 = r_2$, and so is a family of balls that is "thinly distributed" in Hadwiger's sense. Pairwise-inflatable families of balls are not only more general than families of disjoint congruent balls but allow to generalize most of our proofs obtained in three or four dimensions to arbitrary dimension; the key property, which we prove in this paper, is that the set of pairwise-inflatable families is closed under intersection with affine subspaces, unlike the set of families of disjoint congruent balls.

An order-respecting line transversal to a subset of an ordered family is a line transversal that respects the order induced by the family on that subset. An ordered family $\mathcal{F}$ of pairwise-inflatable balls is said to have property $(OR)T$ if it admits a (order-respecting) line transversal. If every k or fewer members of $\mathcal{F}$ admit a (order-respecting) line transversal then $\mathcal{F}$ is said to have property $(OR)T(k)$. Our first main result requires that the line transversals to the subfamilies induce consistent orderings:

Theorem 1 *For any ordered family of pairwise-inflatable balls in* $\mathbb{R}^d$, $ORT(2d)$ *implies* T *and* $ORT(2d+1)$ *implies* ORT.

We then remove the condition on the ordering at the cost of increasing the Helly number to $4d-1$ and restricting ourselves to disjoint unit balls:

Theorem 2 *For any family of disjoint unit balls in* $\mathbb{R}^d$, $T(4d-1)$ *implies* T.

Our results are thus both qualitative and quantitative: we generalize Danzer's result to arbitrary dimension and prove that the Helly number grows at most linearly with the dimension. We build on the work of Holmsen et al. [17] who obtained results similar to Theorems 1 and 2 for disjoint unit balls in three dimensions, albeit with larger bounds on Helly numbers (12 and 46 instead of 6 and 11, respectively). A previous version of this paper, also restricted to disjoint unit balls in three dimensions, appeared in the Symposium on Computational Geometry 2005 [4].

Paper Outline To prove Theorem 1, we start with a family of balls having property $ORT(2d)$ and continuously shrink them until that property no longer holds, following Hadwiger's approach [14]. Before the set of order-respecting line transversals to a $2d$-tuple of balls disappears, it first reduces to a single line (Corollary 12) and this line is an isolated line transversal to $2d-1$ of the balls (Proposition 13). That line has then to be a line transversal to the whole family and Theorem 1 follows; considerations on geometric permutations yield Theorem 2.

Proving the two properties mentioned above (Corollary 12 and Proposition 13) is elementary in the plane but requires considerably more work in higher dimension. Our proofs rely on Proposition 4, the cornerstone of this paper, which shows that the directions of order-respecting line transversals to a family of pairwise-inflatable

balls form a strictly convex subset of $\mathbb{S}^{d-1}$. This directly implies Corollary 12 and yields that order-respecting line transversals form a contractible set in line space. From there, a well-known topological analogue of Helly's theorem (Theorem 3) leads to a weaker version of Theorem 1 sufficient to prove Proposition 13.

2 Preliminaries

Transversals Let $\mathcal{F}$ be a finite family of disjoint compact convex sets $\mathcal{F}$ in $\mathbb{R}^d$ with a given linear order $\prec_{\mathcal{F}}$. We will call $\mathcal{F}$ a *sequence* to stress the existence of this order. A *line transversal* to a *sequence* $\mathcal{F}$ is an oriented line that intersects all the objects of $\mathcal{F}$ *in the order prescribed by* $\prec_{\mathcal{F}}$. A line transversal is *strict* if it intersects the *interior* of each object in $\mathcal{F}$.

For a sequence $\mathcal{F}$, let $\mathcal{K}(\mathcal{F}) \subset \mathbb{S}^{d-1}$ denote the set of directions of line transversals to $\mathcal{F}$. That is, a direction vector $v \in \mathbb{S}^{d-1}$ is in $\mathcal{K}(\mathcal{F})$ if there is a line transversal to $\mathcal{F}$ with direction v. Note that the direction vector of a line transversal determines the order in which it intersects a family of disjoint convex objects. Thus, if sequences $\mathcal{F}_1$ and $\mathcal{F}_2$ are two distinct orderings of the same collection of objects, then $\mathcal{K}(\mathcal{F}_1)$ and $\mathcal{K}(\mathcal{F}_2)$ are disjoint. We will call $\mathcal{K}(\mathcal{F})$ the *cone of directions* of $\mathcal{F}$. Similarly, let $\mathcal{K}^\circ(\mathcal{F})$ be the set of directions of *strict* line transversals to $\mathcal{F}$.

Note that all our line transversals must respect a given order. Only in Sect. 5 will we consider line transversals without order restriction. For clarity, let us call such a line transversal an *unordered* line transversal.

We consider the natural topology over the set of oriented lines in $\mathbb{R}^d$: U is a neighborhood of a line ℓ if and only if for some $\delta > 0$ it contains all lines ℓ' such that the shortest distance between ℓ and ℓ' and the angle between their direction vectors are both less than δ. An *isolated* line transversal to a family of objects $\mathcal{F}$ is an isolated point of the set of line transversals to $\mathcal{F}$, that is, a line transversal ℓ which is a connected component of the line transversals to $\mathcal{F}$.

Given a ball A and a direction v in $\mathbb{R}^d$, we denote by $P_v(A)$ the $(d-1)$-dimensional ball obtained by projecting A orthogonally on an hyperplane with normal v. Observe that a sequence of balls $\mathcal{F}$ has a line transversal with direction v if and only if the balls $P_v(\mathcal{F}) := \{P_v(A) \mid A \in \mathcal{F}\}$ have non-empty intersection. Similarly, $\mathcal{F}$ has a strict line transversal with direction v if and only if the intersection of $P_v(\mathcal{F})$ has non-empty interior.

Inflatable Balls A collection $\mathcal{F}$ of balls in $\mathbb{R}^d$ is called *pairwise-inflatable* if for every two balls $B_1, B_2 \in \mathcal{F}$ we have $\gamma^2 > 2(r_1^2 + r_2^2)$, where r_i is the radius of B_i, and γ is the distance between their centers. Note that for balls of equal radius, this condition only enforces that they are disjoint (and so any family of disjoint congruent balls is pairwise-inflatable). The more unequal the radius of the balls, however, the stronger the distance constraint. At the limit, when $r_1 = 0$, the constraint is $\gamma > \sqrt{2}r_2$. Pairwise-inflatability is less restrictive than Hadwiger's notion of "thinly distributed" balls, which can be defined as $\gamma^2 > 4(r_1 + r_2)^2$ for each pair of balls.

The class of families of pairwise-inflatable balls is closed under intersection with affine subspaces (as proved in Lemma 5). This property (which does not hold for unit-radius balls) will allow us to carry results proved in three dimensions over to $\mathbb{R}^d$.

Topological Machinery We use a few notions from topology that we now review (these can be found, for instance, in the introductory chapter of Matoušek's book [19]). Given a topological space A and a subset $B \subset A$, B is a *deformation retract* of A if there exists a continuous map $F : A \times [0,1] \to A$ such that

$$\begin{cases} F(a,0) = a & \text{for any } a \in A, \\ F(b,t) = b & \text{for any } b \in B \text{ and } t \in [0,1], \\ F(a,1) \in B & \text{for any } a \in A. \end{cases}$$

Two topological spaces A, B are *homotopy equivalent* if there exists a third space C such that both A and B are deformation retracts of C. A space that is homotopy equivalent to a single point is said to be *contractible*. A homology cell is a non-empty set with trivial homology, e.g. a point. Since homology is invariant under homotopy equivalence, any contractible space is a homology cell. A generalization of Helly's theorem based on topology instead of convexity was originally given by Helly himself [15]. We will use a version proved by Debrunner using modern tools (singular homology) [7], as it allows us to work with open sets.

Theorem 3 (Topological Helly Theorem [7]) *Let $\{X_j\}_{j\in J}$ be a finite family of open subsets of Euclidean d-space $\mathbb{R}^d$ such that the intersection $X_{j_1} \cap \cdots \cap X_{j_r}$ of each r sets of this family is nonempty for $r \leq d+1$ and is even a homology cell for $r \leq d$. Then $\bigcap_{j\in J} X_j$ is a homology cell.*

In fact, we only use a weaker version of this theorem where "homology cell" is replaced by "contractible".

Compatible Directions Let $\mathcal{D}$ be a set of directions in $\mathbb{R}^d$ completely contained in the interior of a hemisphere of $\mathbb{S}^{d-1}$, and let $\mathcal{L}(\mathcal{D})$ be the set of lines with directions in $\mathcal{D}$. We parametrize $\mathcal{L}(\mathcal{D})$ as a subset of $\mathbb{R}^{2d-2}$, using the points of intersection of a line $\ell \in \mathcal{L}(\mathcal{D})$ with two parallel hyperplanes that are not parallel to any direction in $\mathcal{D}$. Our aim is to apply the Topological Helly Theorem to sets of line transversals to pairwise-inflatable balls. Unfortunately, such sets are not necessarily homology cells, and may in fact even be disconnected: two lines intersecting disjoint objects in different orders cannot be in the same connected component of transversals to these objects. We overcome this difficulty by restricting the set of directions that we allow for transversals. For a sequence $\mathcal{F}$ of pairwise-inflatable balls in $\mathbb{R}^{d-1}$, let

$$U(\mathcal{F}) := \big\{ c(Y) - c(X) \;\big|\; X, Y \in \mathcal{F};\ X \prec_{\mathcal{F}} Y \big\},$$

where $c(X)$ denotes the center of ball X. Let $\mathcal{D}_{\mathcal{F}}$ be the set of directions making a positive dot-product with each $u \in U(\mathcal{F})$. Note that $\mathcal{D}_{\mathcal{F}}$ is an open convex set on the sphere of directions $\mathbb{S}^{d-1}$. Clearly a line transversal $\ell \in \mathcal{L}(\mathcal{D}_{\mathcal{F}})$ for a subset $\mathcal{F}' \subset \mathcal{F}$ respects the order on $\mathcal{F}'$. Such a line transversal is called a transversal to $\mathcal{F}'$ *compatible* with $\mathcal{F}$.

3 The Cone of Directions is Strictly Convex

We now establish the cornerstone of this paper, a generalization of the first lemma by Holmsen et al. [17] to arbitrary dimension:

Proposition 4 *Let $\mathcal{F}$ be a sequence of pairwise-inflatable balls in $\mathbb{R}^d$. Then $\mathcal{K}(\mathcal{F})$ is strictly convex.*

The proof of this proposition is based on Lemma 7, which shows that some well-chosen fibers over 1-dimensional slices of the cone of directions of unit balls in $\mathbb{R}^4$ are convex. We also need some properties of families of pairwise-inflatable balls. We start by showing that this class is closed under intersection with affine subspaces.

Lemma 5 *Let $\mathcal{F}$ be a family of pairwise-inflatable balls in $\mathbb{R}^d$, and let E be an affine subspace of dimension $k < d$. Then $\mathcal{F}' := \{B \cap E \mid B \in \mathcal{F}\}$ is a family of pairwise-inflatable balls in E.*

Proof We prove the claim for $k = d - 1$ and the lemma follows by induction. Let $B_1, B_2 \in \mathcal{F}$ with respective radii r_1 and r_2 and centers at distance γ apart. Since $\mathcal{F}$ is pairwise-inflatable we have $\gamma^2 > 2(r_1^2 + r_2^2)$. For $i = 1, 2$ let $B_i' = B_i \cap E$, ρ_i denote the radius of B_i' and δ_i be the distance between the center of B_i and that of B_i'. First, observe that

$$\gamma^2 \leq \Delta^2 + (\delta_1 + \delta_2)^2,$$

where Δ is the distance between the centers of B_1' and B_2'. If E separates the centers of B_1 and B_2 the equality holds. If E does not separate the centers, then replacing B_2 by its mirror image with respect to E increases γ while leaving all other quantities unchanged, hence the inequality. Then from $(\delta_1 - \delta_2)^2 \geq 0$ we deduce $(\delta_1 + \delta_2)^2 \leq 2(\delta_1^2 + \delta_2^2)$ and since $r_i^2 = \rho_i^2 + \delta_i^2$ we finally obtain

$$\Delta^2 \geq \gamma^2 - (\delta_1 + \delta_2)^2 > 2(r_1^2 + r_2^2) - 2(\delta_1^2 + \delta_2^2) = 2(\rho_1^2 + \rho_2^2)$$

and the claim follows. □

The following lemma shows that two pairwise-inflatable balls in dimension d can always be "inflated"[1] to two disjoint equal-radius balls in dimension $d + 1$.

Lemma 6 *Let E be a d-dimensional subspace of $\mathbb{R}^{d+1}$, and let $B_1', B_2' \subset E$ be pairwise-inflatable d-dimensional balls in E. Then there exist two disjoint $(d+1)$-dimensional balls B_1, B_2 of* equal radius *in $\mathbb{R}^{d+1}$ such that $B_1' = B_1 \cap E$ and $B_2' = B_2 \cap E$.*

Proof Let q_i and ρ_i be the center and radius of B_i', for $i = 1, 2$. Consider the line orthogonal to E through q_i. Pick a point p_i on this line at distance δ_i from q_i, in such a way that p_1 and p_2 are on opposite sides of E. Let also B_i be the ball with center p_i and radius $r_i = \sqrt{\delta_i^2 + \rho_i^2}$. Clearly $B_i' = B_i \cap E$ and it remains to pick δ_i such that $r_1 = r_2$ and B_1 and B_2 are disjoint.

[1] Hence the name "pairwise-inflatable".

Let Δ be the distance between q_1 and q_2. Without loss of generality, we assume $\rho_1 > \rho_2$. Since $\Delta^2 > 2(\rho_1^2 + \rho_2^2)$, there exists $\sigma > 0$ such that

$$\sigma^2 < \min\{\Delta^2 - 2(\rho_1^2 + \rho_2^2),\ \rho_1^2 - \rho_2^2\}$$

and we can define

$$\delta_1 = (\rho_1^2 - \rho_2^2 - \sigma^2)/(2\sigma) \quad \text{and} \quad \delta_2 = \delta_1 + \sigma.$$

Now, since $2\sigma\delta_1 + \sigma^2 = \rho_1^2 - \rho_2^2$ we have that $\delta_2^2 = \delta_1^2 + \rho_1^2 - \rho_2^2$, and it follows that B_1 and B_2 have equal radius $r = r_1 = r_2$. Now, the distance γ between their centers satisfies

$$\gamma^2 = \Delta^2 + (\delta_1 + \delta_2)^2 = (\Delta^2 + 2\delta_1\delta_2) + \delta_1^2 + \delta_2^2.$$

Since

$$\Delta^2 - 2(\rho_1^2 + \rho_2^2) > \sigma^2 = (\delta_2 - \delta_1)^2 = \delta_1^2 + \delta_2^2 - 2\delta_1\delta_2$$

it follows that

$$\Delta^2 + 2\delta_1\delta_2 > \delta_1^2 + \delta_2^2 + 2(\rho_1^2 + \rho_2^2)$$

and finally

$$\gamma^2 > 2(\delta_1^2 + \rho_1^2) + 2(\delta_2^2 + \rho_2^2) = 4r^2.$$

This shows that B_1 and B_2 are disjoint. □

Let now $F = (O, x, y, z, w)$ be an orthogonal frame in four-dimensional space $\mathbb{R}^4$. Let H denote the plane (O, x, y), and let $H(z, w)$ be the translated copy of H going through the point[2] $(0, 0, z, w)$. Given two disjoint convex sets A and B in $\mathbb{R}^4$, we denote by $Q_{AB}^F \subset \mathbb{R}^2 \times \mathbb{S}^1$ the set of all (z, w, α) such that there is an oriented line in $H(z, w)$ that intersects A before B and that makes an angle α with the x-axis.

Lemma 7 *If A and B are disjoint congruent balls in $\mathbb{R}^4$ then Q_{AB}^F is convex for any orthogonal frame F of $\mathbb{R}^4$.*

We prove this lemma by showing that Q_{AB}^F is the volume under the graph of a concave function of two variables, which involves showing that the Hessian of this function is negative definite. We thus follow the approach of Holmsen et al. [17, proof of Lemma 1] but the details (postponed to Appendix) are more involved.

We proceed to prove the convexity of $\mathcal{K}(\mathcal{F})$ (but not yet its strict convexity) for the 3-dimensional case.

Lemma 8 *Let $\mathcal{F}$ be a sequence of pairwise-inflatable balls in $\mathbb{R}^3$. Then $\mathcal{K}(\mathcal{F})$ is convex.*

[2]By abuse of notation, we use the letters z and w to label the coordinate axes and to represent the coordinates of some specific point, the meaning being clear from the context.

Proof We need to show that for any pair $v_1, v_2 \in \mathcal{K}(\mathcal{F})$ the great circle arc joining them on $\mathbb{S}^2$ lies in $\mathcal{K}(\mathcal{F})$ (since $\mathcal{K}(\mathcal{F})$ is contained in an open hemisphere of $\mathbb{S}^2$, there is a unique such arc of length less than π). We thus let ℓ_1, ℓ_2 be line transversals to $\mathcal{F}$ with directions v_1, v_2, and pick a plane H parallel to both ℓ_1 and ℓ_2. We embed the 3-dimensional space as an affine 3-space of $\mathbb{R}^4$, and equip $\mathbb{R}^4$ with a frame $F = (O, x, y, z, w)$ such that $\{w = 0\}$ is our original 3-dimensional space, and such that (O, x, y) coincides with H.

For any pair of balls (B_1', B_2') from $\mathcal{F}$ with $B_1' \prec_{\mathcal{F}} B_2'$, Lemma 6 gives us two balls $B_1, B_2 \subset \mathbb{R}^4$ of equal radius such that $B_i' = B_i \cap \{w = 0\}$. By Lemma 7, $Q^F_{B_1 B_2}$ is convex and so $Q^F_{B_1' B_2'} = Q^F_{B_1 B_2} \cap \{w = 0\}$ is convex as well. It follows that

$$Q := \bigcap_{A, B \in \mathcal{F},\, A \prec_{\mathcal{F}} B} Q^F_{AB}$$

is a convex set.

Each point in Q corresponds to a family of parallel and coplanar lines such that each pair (A, B) in $\mathcal{F}$ is intersected by at least one of them in the correct order. Helly's theorem (in one dimension) implies that there is a line transversal to $\mathcal{F}$ in this family and this transversal is trivially order-respecting. Let $q_1, q_2 \in Q$ be the points representing the line transversals ℓ_1 and ℓ_2. For any direction v on the great circle arc $v_1 v_2$ there is a point q on the segment $q_1 q_2$ whose associated line transversal has direction v. □

We now characterize the boundary of $\mathcal{K}(\mathcal{F})$. This will allow us to show that $\mathcal{K}(\mathcal{F})$ is not only convex, but even strictly convex. The result will then carry over rather effortlessly to arbitrary dimension. Recall that $\mathcal{K}^\circ(\mathcal{F})$ is the set of directions of *strict* transversals to $\mathcal{F}$. The next lemma shows that $\mathcal{K}^\circ(\mathcal{F})$ is the interior of $\mathcal{K}(\mathcal{F})$.

Lemma 9 *Let $\mathcal{F}$ be a sequence of disjoint balls in $\mathbb{R}^3$, $v \in \mathbb{S}^2$ and $D := \bigcap P_v(\mathcal{F})$. Then $v \in \partial\mathcal{K}(\mathcal{F})$ if and only if D is a point and $v \in \mathrm{int}(\mathcal{K}(\mathcal{F}))$ if and only if D has non-empty interior.*

Proof Clearly $v \in \mathcal{K}(\mathcal{F})$ if and only if D is non-empty. Since $P_v(\mathcal{F})$ is a family of discs, D is either empty, a point, or has non-empty interior. If D has non-empty interior, then a small perturbation of the direction v cannot cause D to become empty, and so $v \in \mathrm{int}(\mathcal{K}(\mathcal{F}))$. It remains to show that if D is a point, then $v \in \partial\mathcal{K}(\mathcal{F})$.

We thus assume that D is a point. Let $k \geq 2$ be the number of discs that have this point on their boundary, and let ℓ be the (unique) transversal of $\mathcal{F}$ with direction v. If $k = 2$ then ℓ lies in a plane separating two balls and there are directions v' arbitrarily close to v such that no line transversal with direction v' to these two balls exists (see Fig. 1). Thus, $v \in \partial\mathcal{K}(\mathcal{F})$. If $k \geq 3$ then by Helly's theorem in the plane there are three balls whose projections intersect in a single point. Let A denote the middle one with respect to $\prec_{\mathcal{F}}$ and let ℓ' be the line through the center of A and its tangency point with ℓ (see Fig. 2). Consider a rotation of v by a small angle δ around ℓ'. This rotation leaves $P_v(A)$ invariant and moves the centers of the two other projections along lines orthogonal to $P_v(\ell')$, either both away from $P_v(\ell')$ or both towards $P_v(\ell')$, depending on the sign of δ. Any sufficiently small rotation that moves the centers away from

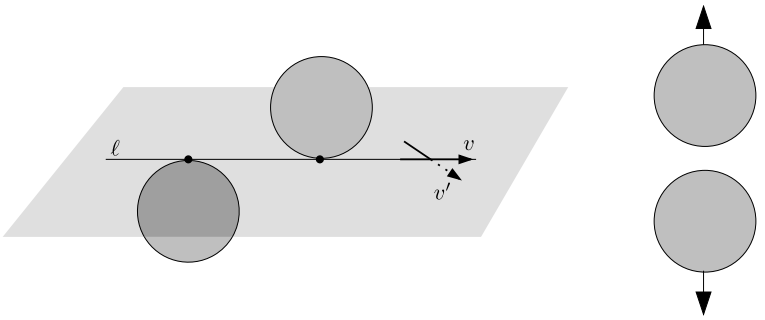


Fig. 1 Perturbation removing all transversals when $k = 2$: 3D view (*left*) and projections (*right*)

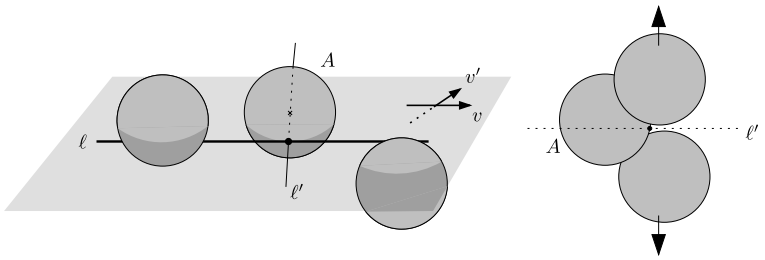


Fig. 2 Perturbation removing all transversals when $k = 3$: 3D view (*left*) and projections (*right*)

$P_v(\ell')$ turns v into a direction v' such that no transversal to the three balls exists in the direction v'. In that case we again have $v \in \partial\mathcal{K}(\mathcal{F})$. □

Lemma 10 *If $\mathcal{F}$ is a sequence of pairwise inflatable balls in $\mathbb{R}^3$ then $\mathcal{K}(\mathcal{F})$ is strictly convex.*

Proof We already know that $\mathcal{K}(\mathcal{F})$ is convex. If $\mathcal{K}(\mathcal{F})$ is not strictly convex then it has to contain on its boundary a great circle arc. By the previous lemma, if $v \in \partial\mathcal{K}(\mathcal{F})$ then $P_v(\mathcal{F})$ is a point. This implies, by Helly's theorem, that the boundary of $\mathcal{K}(\mathcal{F})$ consists of (finitely many) curve arcs that are either (a) directions of bitangent lines lying in bitangent planes or (b) directions of tritangent lines. The directions of bitangent lines lying in bitangent planes to two balls contain a great circle arc only if the two balls are tangent, which cannot occur in our situation.

Therefore, if $\mathcal{K}(\mathcal{F})$ is not strictly convex then it contains in its boundary a great circle arc of directions of lines tangent to three balls. These directions, being on a great circle arc, are parallel to a given plane. In projective geometry, parallels to a plane are recast as lines intersecting the "line at infinity" of that plane. Thus, if $\mathcal{K}(\mathcal{F})$ is not strictly convex, $\mathcal{F}$ contains three balls with infinitely many common tangents that intersect a fixed line at infinity. Such configurations were tabulated by Megyesi and Sottile [20]. Their cases (i), (iii), and (iv) cannot arise with disjoint spheres and the fixed line at infinity. The remaining possibility (case (ii)) is that the three spheres are tangent to a cone whose apex lies on the fixed line. In our case, that line is at

infinity so this cone is a cylinder and the spheres have equal radii and aligned centers; all common tangents then have the same direction and cannot form a great circle arc. □

We will need the generalization of Lemma 9 to arbitrary dimension.

Lemma 11 *If $\mathcal{F}$ is a sequence of disjoint balls in $\mathbb{R}^d$, then $\mathcal{K}^\circ(\mathcal{F}) = \text{int}(\mathcal{K}(\mathcal{F}))$.*

Proof As in the proof of Lemma 9 we observe that $\mathcal{K}^\circ(\mathcal{F}) \subset \text{int}(\mathcal{K}(\mathcal{F}))$, and it remains to prove the other inclusion. Let $v \in \text{int}(\mathcal{K}(\mathcal{F}))$ and pick $v_1, v_2 \in \mathcal{K}(\mathcal{F})$ in a neighborhood of v such that v lies in the interior of the great circle arc v_1v_2. Let ℓ_1, ℓ_2 be two line transversals to $\mathcal{F}$ with directions v_1, v_2, and let E be an affine subspace of dimension three containing both lines (E is unique if the lines are skew). By Lemma 5, the section of $\mathcal{F}$ by E is a sequence $\mathcal{F}'$ of pairwise-inflatable balls. Since v_1 and v_2 belong to $\mathcal{K}(\mathcal{F}')$ and v is interior to the great circle arc they span, Lemma 10 implies that $v \in \text{int}(\mathcal{K}(\mathcal{F}')) = \mathcal{K}^\circ(\mathcal{F}')$ and, by Lemma 9, there is a strict transversal to $\mathcal{F}'$ with direction v. This line is also a strict transversal to $\mathcal{F}$ and Lemma 9 yields that $v \in \mathcal{K}^\circ(\mathcal{F})$. □

We can now finally prove the main result of this section.

Proof Proposition 4 Let $v_1, v_2 \in \mathcal{K}(\mathcal{F})$ with $v_1 \neq v_2$. Since $\mathcal{K}(\mathcal{F})$ is a closed convex set contained in an open hemisphere of $\mathbb{S}^{d-1}$, there is a unique great circle arc of length less than π connecting v_1 and v_2. We need to show that all interior points of this great circle arc lie in the interior of $\mathcal{K}(\mathcal{F})$.

Let ℓ_1, ℓ_2 be two line transversals to $\mathcal{F}$ with directions v_1 and v_2. Let E be an affine subspace of dimension three containing both transversals. The space E intersects every ball in $\mathcal{F}$ and, by Lemma 5, the section of $\mathcal{F}$ by E is a sequence $\mathcal{F}'$ of pairwise-inflatable balls.

Let v be an interior point of the great circle arc v_1v_2. The direction v lies in E, and since $\mathcal{K}(\mathcal{F}')$ is strictly convex by Lemma 10, we have $v \in \text{int}(\mathcal{K}(\mathcal{F}')) = \mathcal{K}^\circ(\mathcal{F}')$. A strict transversal to $\mathcal{F}'$ is a strict transversal to $\mathcal{F}$, and so Lemma 11 implies $v \in \mathcal{K}^\circ(\mathcal{F}) = \text{int}(\mathcal{K}(\mathcal{F}))$. □

Proposition 4 has the following important corollary:

Corollary 12 *Let $\mathcal{F}$ be a sequence of pairwise-inflatable balls in $\mathbb{R}^d$. If $\mathcal{K}(\mathcal{F})$ has empty interior then it is a point.*

4 Pinning Number of Pairwise-Inflatable Balls

A family $\mathcal{F}$ of objects *pins* a line ℓ if ℓ is an isolated transversal to $\mathcal{F}$. The *pinning number* of a class $\mathcal{C}$ of families of objects is defined as the smallest integer k such that the following holds: if a family $\mathcal{F} \in \mathcal{C}$ pins a line ℓ then some subfamily $\mathcal{F}' \subset \mathcal{F}$ of size at most k already pins ℓ. A key ingredient in Hadwiger's original proof of his theorem [14] is the fact that the pinning number of disjoint planar convex sets is 3. In

this section we show a similar result for pairwise-inflatable balls in $\mathbb{R}^d$. Note that the pinning number k is simply the Helly number for the property of "not being pinned": if a line transversal to a family $\mathcal{F}$ is not pinned by any subfamily of size k then it is not pinned by $\mathcal{F}$.

Proposition 13 *The pinning number of pairwise-inflatable balls in $\mathbb{R}^d$ is at most $2d-1$.*

Our proof is based on Lemma 14, which shows that sets of compatible transversals are contractible and therefore homology cells, and Lemma 15, which applies the Topological Helly Theorem to these sets of lines and obtains a weak version of our Theorem 1. We state the next lemma using the notion of "compatible" transversal introduced in Sect. 2:

Lemma 14 *Let $\mathcal{F}$ be a sequence of pairwise-inflatable balls in $\mathbb{R}^d$ and $\mathcal{F}'$ be a subsequence of $\mathcal{F}$. Then the set L of line transversals to $\mathcal{F}'$ compatible with $\mathcal{F}$ is a contractible subset of $\mathbb{R}^{2d-2}$.*

Note the restriction on the direction of lines in L: there may be strict order-respecting line transversals to $\mathcal{F}'$ that are not compatible with $\mathcal{F}$.

Proof Given a line $\ell \in L$, let v_ℓ be its direction. A transversal ℓ to $\mathcal{F}'$ is *barycentric* if it goes through the center of mass of the intersection of $P_{v_\ell}(\mathcal{F}')$. For any direction v in $\mathcal{K}(\mathcal{F}')$ there is a unique barycentric transversal to $\mathcal{F}'$, which we denote $b_{\mathcal{F}'}(v)$.

Let L^* denote the set of barycentric transversals to $\mathcal{F}'$ with directions in $\mathcal{D}_{\mathcal{F}}$. The projection of a ball changes continuously with the direction of projection, so $b_{\mathcal{F}'}$ is continuous. Since the direction of a line changes continuously with the line, $b_{\mathcal{F}'}^{-1}$ is also continuous. Thus, $b_{\mathcal{F}'}$ defines a homeomorphism between L^* and $\mathcal{K}(\mathcal{F}') \cap \mathcal{D}_{\mathcal{F}}$.

By Lemma 4, $\mathcal{K}(\mathcal{F}')$ is convex and so is $\mathcal{D}_{\mathcal{F}}$. Thus, $\mathcal{K}(\mathcal{F}') \cap \mathcal{D}_{\mathcal{F}}$ is convex and hence contractible. It follows that L^* is also contractible. The map

$$\begin{cases} L \times [0,1] \to L, \\ (\ell, t) \mapsto \ell + t(b_{\mathcal{F}'}(v_\ell) - \ell), \end{cases}$$

is continuous and shows that L^* is a deformation retract of L. Since L^* is contractible, so is L. □

We can now apply the Topological Helly Theorem to obtain a "weak" Hadwiger-type result.

Lemma 15 *Let $\mathcal{F}$ be a sequence of at least $2d-1$ pairwise-inflatable balls in $\mathbb{R}^d$. If every subfamily $\mathcal{F}' \subset \mathcal{F}$ of $2d-1$ balls admits a strict line transversal with a direction in $\mathcal{D}_{\mathcal{F}}$, then $\mathcal{F}$ admits a strict line transversal.*

Proof We apply Theorem 3 on $\mathcal{L}(\mathcal{D}_{\mathcal{F}})$. With the parametrization discussed above, $\mathcal{L}(\mathcal{D}_{\mathcal{F}}) \subset \mathbb{R}^{2d-2}$. For $S \in \mathcal{F}$ let X_S be the subset of $\mathcal{L}(\mathcal{D}_{\mathcal{F}})$ of lines intersecting the interior of ball S. Clearly, X_S is an open set in $\mathbb{R}^{2d-2}$. Consider now the intersection

$Y := X_{S_1} \cap \ldots \cap X_{S_r}$ of r such sets. The set Y consists of exactly those lines in $\mathcal{L}(\mathcal{D}_\mathcal{F})$ that are strict transversals of $S_1, \ldots, S_r$. The assumption of the lemma implies that $Y \neq \emptyset$ for $r \leq 5$. By Lemma 14, Y is contractible and hence a homology cell. Theorem 3 now implies that $\bigcap_{S \in \mathcal{F}} X_S \neq \emptyset$, and so there is an order-respecting strict line transversal for $\mathcal{F}$. □

In principle, Lemma 15 is the Hadwiger-type result we are looking for. Its drawback is that it requires a subfamily of balls to have not only an order-respecting transversal, but one that, in a sense, respects the order on the *entire* family of balls. This is nonetheless enough to prove the desired result on the pinning number of pairwise-inflatable balls:

Proof of Proposition 13 Let $\mathcal{F}$ be a family of at least $2d$ pairwise-inflatable balls in $\mathbb{R}^d$ admitting an isolated line transversal ℓ. Let $\prec$ be the order on $\mathcal{F}$ induced by ℓ. Lemma 14 implies that the set of line transversals to $\mathcal{F}$ respecting $\prec$ is connected, and so ℓ is the only order-respecting line transversal to $\mathcal{F}$.

Since ℓ is not a strict transversal, $\mathcal{F}$ has no strict order-respecting transversal. By Lemma 15, there is a subfamily $\mathcal{F}' \subset \mathcal{F}$ of $2d-1$ balls that has no strict order-respecting transversal with direction in $\mathcal{D}_\mathcal{F}$, that is $\mathcal{K}^\circ(\mathcal{F}') \cap \mathcal{D}_\mathcal{F} = \emptyset$. However, $\mathcal{K}(\mathcal{F}') \cap \mathcal{D}_\mathcal{F} \neq \emptyset$ since it contains the direction of ℓ. Since $\mathcal{K}(\mathcal{F}')$ is convex, by Lemma 4, and $\mathcal{D}_\mathcal{F}$ is open, it follows that $\mathcal{K}^\circ(\mathcal{F}') = \emptyset$ and $\mathcal{F}'$ has no strict order-respecting transversal at all. Now, $\mathcal{K}(\mathcal{F}')$ is non-empty but has empty interior, so, by Corollary 12, $\mathcal{K}(\mathcal{F}')$ is a single direction v. Since $\mathcal{K}(\mathcal{F}') = \{v\}$, the balls $P_v(\mathcal{F}')$ intersect in a unique point and ℓ is the only order-respecting line transversal of $\mathcal{F}'$, and is thus isolated. □

5 Hadwiger and Helly-Type Theorems

We can now prove the main results of this paper.

A Hadwiger-Type Theorem Propositions 12 and 13 are all we need to reproduce Hadwiger's original proof of the 2-dimensional case.

Proof of Theorem 1 We simultaneously shrink all the balls and continue shrinking as long as every subset of size $2d$ has a transversal. If all the centers are aligned then the theorem trivially holds. Otherwise, at some point in the shrinking process a subfamily $\mathcal{F}'$ of size $2d$ stops having a transversal. The cone $\mathcal{K}(\mathcal{F}')$ changes continuously during the shrinking and must have empty interior before disappearing. Thus, by Corollary 12, at that moment the sequence $\mathcal{F}'$ has a unique transversal ℓ.

Now, by Proposition 13, there is then a subfamily $\mathcal{F}'' \subset \mathcal{F}'$ of at most $2d-1$ balls such that ℓ is the unique transversal of $\mathcal{F}''$. For any ball $X \in \mathcal{F} \setminus \mathcal{F}''$, the set $\mathcal{F}'' \cup \{X\}$ has a line transversal ℓ_X. Since the only line transversal of $\mathcal{F}''$ is ℓ, we must have $\ell_X = \ell$, and ℓ intersects X. It follows that ℓ is an unordered line transversal for $\mathcal{F}$.

Similarly, if any subfamily of size $2d+1$ admits a line transversal there exists a subfamily $\mathcal{F}'$ of $2d-1$ balls having a unique line transversal ℓ. For any $X, Y \in \mathcal{F}$ with $X \prec Y$, the subfamily $\mathcal{F}' \cup \{X, Y\}$ admits a line transversal that must be ℓ, and ℓ intersects X before Y. It follows that ℓ is an (order-respecting) line transversal of $\mathcal{F}$. □

Removing the Ordering Assumption We now generalize Theorem 1 by removing the restriction on the ordering. However, we restrict ourselves to the case of disjoint unit balls in $\mathbb{R}^d$ as we build on the following result by Cheong et al. [5].

Theorem 16 [5] *Let $\mathcal{F}$ be a family of at least nine disjoint unit balls in $\mathbb{R}^d$. Then $\mathcal{F}$ admits at most two distinct geometric permutations, which differ only in the swapping of two adjacent balls.*

Proof of Theorem 2 We first shrink the balls simultaneously until some subfamily $\mathcal{F}_{4d-1}$ of $4d-1$ balls is about to lose its last unordered transversal.

If $\mathcal{F}_{4d-1}$ admits more than one (unordered) line transversal (all of which vanish if the balls are shrunk any further), each transversal must realize a different geometric permutation. Theorem 16 then implies that $\mathcal{F}_{4d-1}$ has exactly two line transversals, ℓ_1 and ℓ_2, with two distinct geometric permutations. By Proposition 13, for each ℓ_i there are $2d-1$ balls in $\mathcal{F}_{4d-1}$ for which ℓ_i is the only line transversal respecting the ordering induced by ℓ_i. There is thus a subfamily $\mathcal{F}'$ of $\mathcal{F}_{4d-1}$ of exactly $4d-2$ balls (we can complete $\mathcal{F}'$ using balls from $\mathcal{F}_{4d-1}$ if needed) for which ℓ_1 and ℓ_2 are the only line transversals respecting their respective orders. By Theorem 16, $\mathcal{F}'$ admits at most two geometric permutations, and so ℓ_1 and ℓ_2 are its only line transversals. Since any subfamily of $4d-1$ balls has a line transversal, any ball of $\mathcal{F} \setminus \mathcal{F}'$ must intersect ℓ_1 or ℓ_2. If all the balls intersect both lines then the theorem is proved. Otherwise, there exists a ball A that intersects, say, ℓ_1 but not ℓ_2. Then $\mathcal{F}' \cup \{A\}$ is a family of $4d-1$ balls with a *unique* transversal. We are left with a set $\mathcal{F}_{4d-1}$ of $4d-1$ balls that has a unique transversal ℓ.

Let $\prec_\ell$ be the order on $\mathcal{F}_{4d-1}$ induced by ℓ. By Proposition 13, there is a subfamily $\mathcal{F}_{2d-1} \subset \mathcal{F}_{4d-1}$ such that ℓ is the unique transversal of $\mathcal{F}_{2d-1}$ respecting $\prec_\ell$. For each $Z \in \mathcal{F}_{4d-1} \setminus \mathcal{F}_{2d-1}$, let $\mathcal{F}_Z$ denote the set $\mathcal{F}_{4d-1} \setminus \{Z\}$. If one of the subsets $\mathcal{F}_Z$ has no other transversal than ℓ then every other ball of $\mathcal{F}$ intersects ℓ and the proof is complete.

We now assume that every $\mathcal{F}_Z$ has some transversal ℓ_Z distinct from ℓ and obtain a contradiction. Since $\mathcal{F}_Z$ contains $\mathcal{F}_{2d-1}$, ℓ_Z realizes a geometric permutation different from that of ℓ. By Theorem 16, the order induced by ℓ_Z on $\mathcal{F}_{4d-1}$ differs from $\prec_\ell$ by the swapping of two adjacent balls X, Y. Since ℓ_Z realizes a geometric permutation of $\mathcal{F}_{2d-1}$ different from ℓ, we must have $X, Y \in \mathcal{F}_{2d-1}$. Let $Z_1, Z_2 \in \mathcal{F}_{4d-1} \setminus \mathcal{F}_{2d-1}$, and consider the set $\mathcal{F}_{4d-1} \setminus \{Z_1, Z_2\}$. It admits the transversals ℓ, ℓ_{Z_1}, and ℓ_{Z_2} but, by Theorem 16, at most two geometric permutations. Since ℓ is the unique transversal respecting $\prec_\ell$, ℓ_{Z_1} and ℓ_{Z_2} must realize the same geometric permutation on $\mathcal{F}_{4d-1} \setminus \{Z_1, Z_2\}$. Thus the balls $X, Y \in \mathcal{F}$ do not depend on the choice of Z. Let $\prec$ be the order on $\mathcal{F}_{4d-1}$ obtained from $\prec_\ell$ by swapping X and Y. For any $Z \in \mathcal{F}_{4d-1} \setminus \mathcal{F}_{2d-1}$ the subfamily $\mathcal{F}_Z$ admits a line transversal respecting $\prec$. On the other hand, $\mathcal{F}_{4d-1}$ does not admit such a transversal as ℓ is its only transversal. By (the second half of) Theorem 1, there is a subset $\mathcal{F}_{2d+1} \subset \mathcal{F}_{4d-1}$ of at most $2d+1$ balls that does not admit a transversal respecting $\prec$. We must have $X, Y \in \mathcal{F}_{2d+1}$, as without both X and Y, $\prec_\ell$ and $\prec$ are equivalent. This implies that $|\mathcal{F}_{2d-1} \cup \mathcal{F}_{2d+1}| \leq 4d-2$. There is therefore a $Z \in \mathcal{F}_{4d-1} \setminus \mathcal{F}_{2d-1}$ such that $\mathcal{F}_{2d-1} \cup \mathcal{F}_{2d+1} \subseteq \mathcal{F}_Z$. However, ℓ_Z cannot be a line transversal to $\mathcal{F}_{2d+1}$, a contradiction. □

6 Conclusion and Open Problems

We conclude this paper with a few comments on our results followed by open problems they suggest.

- Weaker versions of Theorems 1 and 2 (with constants quadratic in d) can be obtained more easily, using only Lemma 4 and the reasoning of Holmsen et al. [17].
- In the plane, if three disjoint convex sets $\{C_1, \ldots, C_3\}$ pin a line ℓ then they are all tangent to ℓ and alternate: the first and the third are on the same side of ℓ, the second is on the other side. Thus, if ℓ does not intersect a fourth convex set C_4 some triple $\{C_x, C_y, C_4\}$ has no line transversal at all. This explains why, in Hadwiger's original proof the "Hadwiger number" is the same as the pinning number. A way to reduce the bound in Theorem 1 to $2d-1$ could be to prove a similar statement: given a sequence of pairwise inflatable balls $\mathcal{F}$ that pins a line ℓ and a ball C not intersecting ℓ, there is a subsequence $\mathcal{F}' \subset \mathcal{F}$ of size $|\mathcal{F}|-1$ such that $\mathcal{F}' \cup \{C\}$ has no transversal respecting the ordering on $\mathcal{F}'$. We have no idea whether such a statement actually holds.
- To apply the Topological Helly Theorem, we did not actually need that $\mathcal{K}(\mathcal{F})$ is convex, only that it is contractible. This may be important for further generalization.
- For general convex sets, even smooth ones, the pinning number is at least 6 as the following example using six unit-radius cylinders in $\mathbb{R}^3$, due to Günter Rote, shows: the first three cylinders are parallel to the x-axis and their axes go through the points $(0, 1, 0)$, $(0, -1, 1)$ and $(0, 1, 2)$ respectively. The last three cylinders are parallel to the y-axis and their axes go through the points $(1, 0, 10)$, $(-1, 0, 11)$ and $(1, 0, 12)$ respectively. The six cylinders have only one transversal—the z-axis—but any five have an infinite number of transversals.
- Lemmas 5 and 6 imply that two disjoint balls $A, B \subset \mathbb{R}^d$ are pairwise-inflatable if and only if they can be expressed as sections of two disjoint congruent balls in some higher-dimensional space. Generalizing this, let us call a set $\mathcal{F}$ of balls in $\mathbb{R}^d$ *inflatable* if $\mathcal{F}$ can be expressed as the intersection of a higher-dimensional set of disjoint congruent balls with a d-dimensional affine subspace. Batog recently showed that it is NP-hard to decide whether a given collection of balls is inflatable [2].

Problem 1 *What is the maximum number of geometric permutations of pairwise-inflatable balls in* $\mathbb{R}^d$*?*

To generalize Theorem 2 to pairwise-inflatable balls, one would need to extend Theorem 16 to those families. It is known that the number of geometric permutations of n disjoint balls in $\mathbb{R}^d$ is at most 3 if the balls have equal radii and $\Theta(n^{d-1})$ if the ratio

$$\frac{\text{largest radius}}{\text{smallest radius}}$$

is not bounded independently of n [23].

Problem 2 *For which classes of objects is the cone of directions* $\mathcal{K}(A_1, \ldots, A_n)$ *convex, or at least contractible?*

Our proof of convexity for the cone of directions of balls collapses for balls that are not pairwise-inflatable. In fact, the set Q^F_{AB} is not necessarily convex if B is much smaller than A but very close to it. Note that this problem was recently solved by Borcea et al. [3] for disjoint balls in arbitrary dimension.

Problem 3 *For which classes of objects is the set of order-respecting line transversals always connected?*

Our proof of Theorem 1 follows from (i) a bounded pinning number and (ii) the fact that as the set of order-respecting line transversals to a sequence disappears it first reduces to a single line. For strictly convex objects, property (ii) follows from the connectivity of the set of order-respecting transversals. Surprisingly, it is an open question whether this set is connected for even 4 disjoint balls in $\mathbb{R}^3$, whereas it is known to be connected for any triple of disjoint convex objects [8, Lemma 74]. We conjecture that general convex sets in $\mathbb{R}^d$ have a bounded pinning number. Thus, understanding how general this connectivity property is would provide insight in how general the example of Holmsen and Matousek [16], convex sets whose translates do not admit a Hadwiger theorem, actually is. Of course, a positive answer to Problem 2 for a particular family of convex sets implies a positive answer to Problem 3 for that family as well.

Problem 4 *Given a collection of disjoint unit balls, assume that any subset of size* $2d-1$ *admits a line transversal. Does any subset of size* $2d-1$ *admit a* compatible *line transversal?*

In other words, can our "weak Hadwiger theorem" (Lemma 15) be strengthened into a Hadwiger theorem with a better constant than Theorem 1?

Problem 5 *Is the pinning number of disjoint unit balls in* $\mathbb{R}^d$ *equal to* $2d-1$*?*

Surprisingly, the only known lower bound on the Helly number is the construction done by Hadwiger fifty years ago. Note that the bound in our Hadwiger theorem has to be higher than the pinning number of the corresponding family and one can therefore look for a lower bound on the pinning number. Intuitively, considerations on the dimension suggest that the pinning number in dimension d cannot be less than $2d-1$, the dimension of the underlying line space being $2d-2$.

Acknowledgements We thank Gregory Ginot for helpful discussions and suggesting the proof of Lemma 14, Günter Rote for the lower bound construction with cylinders mentioned in the conclusion, and Guillaume Batog for helpful discussions on inflatability.

Appendix Proof of Lemma 7

Proof Let F be the frame (O, x, y, z, w). We first observe that a translation of F along the x- or y-axis leaves Q^F_{AB} unchanged, while a translation of F along the z- or w-axis causes an equivalent translation of Q^F_{AB}. Rotating the x- and y-axes while leaving the z- and w-axes fixed causes a translation of Q^F_{AB} along the α-axis. Finally,

scaling F causes Q_{AB}^F to be stretched along the z- and w-axes. Since convexity is invariant under affine transformations, we can therefore assume that A and B are *unit-radius* balls with centers at $(0,0,0,-b)$ and $(e,0,0,b)$, where $b>0$, $e>0$. The disjointness of A and B implies that $e^2+4b^2-4>0$. Let D denote the lune-shaped region in the (z,w) plane that corresponds to the intersection of the two unit discs with centers $(0,-b)$ and $(0,b)$. If $(z,w)\notin D$ then $H(z,w)$ does not intersect both A and B. If $b>1$ then D is empty. If $b=1$ then D is reduced to $z=w=0$, $H(0,0)$ intersects both A and B in a point, and so Q_{AB}^F is a point. In the following we can therefore assume $b<1$.

Let

$$R(z,w)=\sqrt{1-z^2-w^2},$$

and let $R_+=R(z,w+b)$ and $R_-=R(z,w-b)$. If $(z,w)\in D$ then $H(z,w)\cap A$ is the disc with center $(0,0)$ and radius R_+, while $H(z,w)\cap B$ is the disc with center $(0,e)$ and radius R_-. Now, let

$$f(z,w)=\frac{R_++R_-}{e}.$$

Since A and B are disjoint, the discs $H(z,w)\cap A$ and $H(z,w)\cap B$ are disjoint, implying that $R_++R_-<e$, and so $0\le f(z,w)<1$. Consider

$$G(z,w)=\arcsin(f(z,w)).$$

Since $(z,w,\alpha)\in Q_{AB}^F$ if and only if $(z,w)\in D$ and $-G(z,w)\le\alpha\le G(z,w)$, it suffices to show that G is a concave function. A sufficient condition for this is that its Hessian $\mathcal{H}(G)$ be negative definite, which we endeavor to prove now. By symmetry with respect to the z- and w-axes, we need to prove negative definiteness only for $z,\ w\ge 0$.

In what follows, subscripts are used to denote partial derivatives. Also, reference to z,w as arguments of functions is dropped when no confusion can arise.

The Hessian of G is

$$\mathcal{H}(G)=\begin{pmatrix}G_{zz}&G_{zw}\\G_{zw}&G_{ww}\end{pmatrix}=\frac{(1-f^2)\mathcal{H}(f)+f(\nabla f)(\nabla f)^T}{(1-f^2)^{3/2}},$$

where $\mathcal{H}(f)$ is the Hessian of f and $\nabla f=(f_z,f_w)^T$ is its gradient. The Hessian of G is negative definite if and only if

$$\text{(i)}\quad G_{zz}<0\quad\text{and}\quad\text{(ii)}\quad\det\mathcal{H}(G)=G_{zz}G_{ww}-G_{zw}^2>0.$$

We prove these two inequalities in turn. For this, we need the following derivatives:

$$R_z=\frac{-z}{R},\qquad R_w=\frac{-w}{R},\qquad R_{zz}=\frac{w^2-1}{R^3},\qquad R_{zw}=\frac{-zw}{R^3},$$
$$R_{ww}=\frac{z^2-1}{R^3},\qquad R_{zzz}=\frac{3(w^2-1)z}{R^5}$$

(i) The first inequality is simple to check. We have

$$G_{zz} = \frac{(1-f^2)f_{zz} + f\,f_z^2}{(1-f^2)^{3/2}}.$$

Since the denominator is strictly positive for all z and w, the sign of G_{zz} is determined by its numerator which we denote by $g(z,w)$. The derivative of g with respect to z is:

$$g_z = (1-f^2)f_{zzz} + f_z^3.$$

For $z>0$, we have $R_z<0$ and $R_{zzz}<0$, so $f_z<0$ and $f_{zzz}<0$ implying that $g_z<0$. It follows that the function $z\mapsto g(z,w)$ is decreasing for $z>0$. Since $g(0,w)<0$ it follows that $g(z,w)<0$ for $z,w\geq 0$, so $G_{zz}<0$.

(ii) The second inequality is considerably more challenging. Let us introduce the following notations:

$$\gamma_+ = R_+^2, \qquad \gamma_- = R_-^2, \qquad \gamma = 1-z^2-w^2+b^2,$$
$$P = \gamma_+\gamma_-, \qquad S = \gamma_+ + \gamma_-.$$

γ_+, γ_- and γ satisfy the following constraints:

$$0<\gamma_+\leq 1-b^2<1, \qquad 0<\gamma_-\leq 4b(1-b)<1 \quad \text{and}$$
$$0<2b^2\leq\gamma<1+b^2<2.$$

Expanding $\det\mathcal{H}(G)$ gives $\det\mathcal{H}(G)=(1-f^2)\Delta$, where

$$\Delta = (1-f^2)\Delta_1 + f\Delta_2,$$
$$\Delta_1 = \det\mathcal{H}(f) = f_{zz}f_{ww} - f_{zw}^2, \qquad \Delta_2 = f_w^2 f_{zz} + f_z^2 f_{ww} - 2f_z f_w f_{zw}.$$

We first find that

$$\Delta_1 = \frac{1}{e^2P^2}(\mu_1+\mu_2\sqrt{P}),$$

where

$$\mu_1 = S^2-2P = \gamma_-^2+\gamma_+^2>0 \quad \text{and} \quad \mu_2 = P+\gamma(2-\gamma)>0.$$

Also,

$$\Delta_2 = \frac{1}{e^3P^{\frac{3}{2}}}(\lambda_-\sqrt{\gamma_-}+\lambda_+\sqrt{\gamma_+}),$$

where

$$\lambda_- = \gamma(\gamma-2)+2\gamma_+(\gamma-1)+P \quad \text{and} \quad \lambda_+ = \gamma(\gamma-2)+2\gamma_-(\gamma-1)+P.$$

Note that since $\lambda_-(z,0)=\lambda_+(z,0)=4z^2(z^2-1)\leq 0$, we can't conclude yet and have to go further along.

Putting everything together, we get

$$\Delta = \frac{\chi}{e^4 P^2},$$

where

$$\begin{aligned}
\chi &= \chi_1 + \chi_2\sqrt{P},\\
\chi_1 &= \mu_1(e^2 - S) + P(\lambda_+ + \lambda_- - 2\mu_2),\\
\chi_2 &= \mu_2(e^2 - S) - 2\mu_1 + \lambda_-\gamma_- + \lambda_+\gamma_+.
\end{aligned}$$

We want to prove that $\chi > 0$, implying $\Delta > 0$. Let $\delta = e^2 + 4b^2 - 4$. Noting that $S + 4 - 2\gamma = \gamma_+ + \gamma_- + 4 - 2\gamma = 4 - 4b^2$, we get that $e^2 - S = \delta + 4 - 2\gamma$. So we have:

$$\chi_1 = \mu_1\delta + \chi_1^*, \qquad \chi_2 = \mu_2\delta + \chi_2^*,$$

where

$$\begin{aligned}
\chi_1^* &= 2\mu_1(2 - \gamma) + P(\lambda_+ + \lambda_- - 2\mu_2),\\
\chi_2^* &= -2\mu_1 + 2\mu_2(2 - \gamma) + \lambda_-\gamma_- + \lambda_+\gamma_+.
\end{aligned}$$

Let $\chi^* = \chi_1^* + \chi_2^*\sqrt{P}$. Then

$$\chi = (\mu_1 + \mu_2\sqrt{P})\delta + \chi^* > \chi^*,$$

since $\mu_1 > 0, \mu_2 > 0, \delta > 0$.

Let us prove that $\chi^* \geq 0$. Let

$$\theta_1 = 2S^2 - 4P - SP - 2P\gamma, \qquad \theta_2 = 2(2 - \gamma) - S.$$

We can rewrite χ_1^* and χ_2^* in terms of θ_1 and θ_2:

$$\chi_1^* = (2 - \gamma)\theta_1 - P\gamma\theta_2, \qquad \chi_2^* = -\theta_1 + \gamma(2 - \gamma)\theta_2.$$

Now observe that χ^* factors:

$$\chi^* = \chi_1^* + \chi_2^*\sqrt{P} = (2 - \gamma - \sqrt{P})(\theta_1 + \theta_2\gamma\sqrt{P}).$$

Noting that $\theta_2 = 4(w^2 + z^2) \geq 0$ and

$$\theta_1 = 2S^2 - 8P + P(2(2 - \gamma) - S) = 2(\gamma_+ - \gamma_-)^2 + P\theta_2 \geq 0,$$

we see that the second factor of χ^* is positive. It remains to observe that $2 - \gamma + \sqrt{P} > 0$ and that

$$(2 - \gamma)^2 - P = 4(z^2(1 - b^2) + w^2) \geq 0,$$

to conclude that $2 - \gamma - \sqrt{P} \geq 0$ and $\chi^* \geq 0$. Overall, $\chi > 0$, $\Delta > 0$ and $\det \mathcal{H}(G) > 0$, which concludes the proof. □

References

1. Ambrus, G., Bezdek, A., Fodor, F.: A Helly-type transversal theorem for n-dimensional unit balls. Arch. Math. **86**(5), 470–480 (2006)
2. Batog, G., Goaoc, X.: Inflating balls is NP-hard (2006, manuscript)
3. Borcea, C., Goaoc, X., Petitjean, S.: Line transversals to disjoint balls. Discrete Comput. Geom. (2007, in press), doi: 10.1007/s00454-007-9016-z
4. Cheong, O., Goaoc, X., Holmsen, A.: Hadwiger and Helly-type theorems for disjoint unit spheres in $\mathbb{R}^3$. In: Proc. 20th Ann. Symp. on Computational Geometry, pp. 10–15, 2005
5. Cheong, O., Goaoc, X., Na, H.-S.: Geometric permutations of disjoint unit spheres. Comput. Geom. Theory Appl. **30**, 253–270 (2005)
6. Danzer, L.: Über ein Problem aus der kombinatorischen Geometrie. Arch. der Math. (1957)
7. Debrunner, H.: Helly type theorems derived from basic singular homology. Amer. Math. Mon. **77**, 375–380 (1970)
8. Goaoc, X.: Structures de visibilité globales: tailles, calculs et dégénérescences. Thèse d'université, Université Nancy 2 (May 2004)
9. Grünbaum, B.: On common transversals. Arch. Math. **IX**, 465–469 (1958)
10. Grünbaum, B.: Common transversals for families of sets. J. Lond. Math. Soc. **35**, 408–416 (1960)
11. Hadwiger, H.: Ungelöste Probleme, No. 7. Elem. Math. (1955)
12. Hadwiger, H.: Problem 107. Nieuw Arch. Wisk. **4**(3), 57 (1956)
13. Hadwiger, H.: Solution. Wisk. Opg. **20**, 27–29 (1957)
14. Hadwiger, H.: Über Eibereiche mit gemeinsamer Treffgeraden. Port. Math. **6**, 23–29 (1957)
15. Helly, E.: Über Systeme von abgeschlossenen Mengen mit gemeinschaftlichen Punkten. Monaths. Math. Phys. **37**, 281–302 (1930)
16. Holmsen, A., Matoušek, J.: No Helly theorem for stabbing translates by lines in $\mathbb{R}^d$. Discrete Comput. Geom. **31**, 405–410 (2004)
17. Holmsen, A., Katchalski, M., Lewis, T.: A Helly-type theorem for line transversals to disjoint unit balls. Discrete Comput. Geom. **29**, 595–602 (2003)
18. Katchalski, M.: A conjecture of Grünbaum on common transversals. Math. Scand. **59**(2), 192–198 (1986)
19. Matoušek, J.: Using the Borsuk-Ulam Theorem. Springer, Berlin (2003)
20. Megyesi, G., Sottile, F.: The envelope of lines meeting a fixed line and tangent to two spheres. Discrete Comput. Geom. **33**(4), 617–644 (2005)
21. Tverberg, H.: Proof of Grünbaum's conjecture on common transversals for translates. Discrete Comput. Geom. **4**(3), 191–203 (1989)
22. Wenger, R.: Helly-type theorems and geometric transversals. In: Goodman, J.E., O'Rourke, J. (eds.) Handbook of Discrete and Computational Geometry, 2nd edn, pp. 73–96. CRC Press, Boca Raton (2004), Chap. 4
23. Zhou, Y., Suri, S.: Geometric permutations of balls with bounded size disparity. Comput. Geom. Theory Appl. **26**, 3–20 (2003)

Grid Vertex-Unfolding Orthogonal Polyhedra

Mirela Damian · Robin Flatland · Joseph O'Rourke

Abstract An *edge-unfolding* of a polyhedron is produced by cutting along edges and flattening the faces to a *net*, a connected planar piece with no overlaps. A *grid unfolding* allows additional cuts along grid edges induced by coordinate planes passing through every vertex. A *vertex-unfolding* allows faces in the net to be connected at single vertices, not necessarily along edges. We show that any orthogonal polyhedra of genus zero has a grid vertex-unfolding. (There are orthogonal polyhedra that cannot be vertex-unfolded, so some type of "gridding" of the faces is necessary.) For any orthogonal polyhedron P with n vertices, we describe an algorithm that vertex-unfolds P in $O(n^2)$ time. Enroute to explaining this algorithm, we present a simpler vertex-unfolding algorithm that requires a $3 \times 1 \times 1$ refinement of the vertex grid.

Keywords Vertex-unfolding · Grid unfolding · Orthogonal polyhedra · Genus-zero

This is a significant revision of the preliminary version that appeared in [2].

J. O'Rourke's research was supported by NSF award DUE-0123154.

M. Damian (✉)
Dept. Comput. Sci., Villanova Univ., Villanova, PA 19085, USA
e-mail: mirela.damian@villanova.edu

R. Flatland
Dept. Comput. Sci., Siena College, Loudonville, NY 12211, USA
e-mail: flatland@siena.edu

J. O'Rourke
Dept. Comput. Sci., Smith College, Northampton, MA 01063, USA
e-mail: orourke@cs.smith.edu

1 Introduction

Two unfolding problems have remained unsolved for many years [3, 5]: (1) Can every convex polyhedron be edge-unfolded? (2) Can every polyhedron be unfolded? An *unfolding* of a 3D object is an isometric mapping of its surface to a single, connected planar piece, the "net" for the object, that avoids overlap. An *edge-unfolding* achieves the unfolding by cutting edges of a polyhedron, whereas a *general-unfolding* places no restriction on the cuts. A net representation of a polyhedron finds use in a variety of applications [8]—from flattening monkey brains [10] to manufacturing, from sheet metal [12] to low-distortion texture mapping [11].

It is known that some nonconvex polyhedra cannot be unfolded without overlap with cuts along edges. However, no example is known of a nonconvex polyhedron that cannot be unfolded with unrestricted cuts. Advances on these difficult problems have been made by specializing the class of polyhedra, or easing the stringency of the unfolding criteria. On one hand, it was established in [1] that certain subclasses of *orthogonal polyhedra*—those whose faces meet at right angles and whose edges are parallel to coordinate axes—that are multiples of 90°—have an unfolding. In particular, the class of *orthostacks*, stacks of extruded orthogonal polygons, was proven to have an unfolding (but not an edge-unfolding). On the other hand, loosening the criteria of what constitutes a net to permit connection through points/vertices, the so-called *vertex-unfoldings*, led to an algorithm to vertex-unfold any triangulated manifold [6] (and indeed, any simplicial manifold in higher dimensions). A vertex unfolding maps the surface to a single, connected piece K in the plane, but K may have "cut vertices" whose removal disconnects K.

A second loosening of the criteria is the notion of grid unfoldings, which are especially natural for orthogonal polyhedra. A *grid unfolding* adds edges to the surface by intersecting the polyhedron with planes parallel to Cartesian coordinate planes through every vertex. The two approaches were recently married in [7], which established that any orthostack may be grid vertex-unfolded. For orthogonal polyhedra, a grid unfolding is a natural median between edge-unfoldings and unrestricted unfoldings.

Our main result is that any orthogonal polyhedron, without shape restriction except that its surface be homeomorphic to a sphere, has a grid vertex-unfolding. We present an algorithm that grid vertex-unfolds any orthogonal polyhedron with n vertices in $O(n^2)$ time. We also present, along the way, a simpler algorithm for $3 \times 1 \times 1$ *refinement* unfolding, a weakening of grid unfolding that we define in the following. We believe that the techniques in our algorithms may help show that all orthogonal polyhedra can be grid edge-unfolded.

2 Definitions

We distinguish between a *strict net*, in which the net boundary does not self-touch, and a *net* for which the boundary may touch but no interior points overlap. The latter corresponds to the physical model of cutting out the net from a sheet of paper, with perhaps some cuts representing *edge overlap*, and this is the model we use in this paper. We also insist as part of the definition of a vertex-unfolding, again keeping

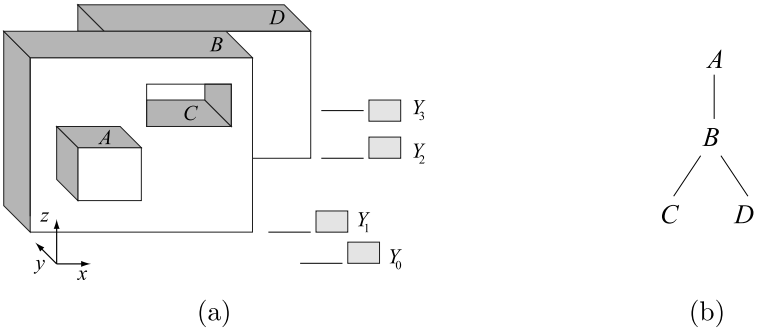


Fig. 1 Definitions. **a** Shaded connected pieces are bands; A, B and D are protrusions; C is a dent. **b** An unfolding tree captures band adjacency structure and determines the algorithm's recursive calls

in spirit with the physical model, that the unfolding "path" never self-crosses on the surface in the following sense. If (A, B, C, D) are four gridfaces incident in that cyclic order to a common vertex v, then the net does not include both the connections AvC and BvD.[1]

We use the following notation to describe the six type of faces of an orthogonal polyhedron, depending on the direction in which the outward normal points: *front*: $-y$; *back*: $+y$; *left*: $-x$; *right*: $+x$; *bottom*: $-z$; *top*: $+z$. We take the z-axis to define the vertical direction; *vertical* faces are parallel to the xz-plane or the yz plane. Directions clockwise and counterclockwise are defined from the perspective of a viewer positioned at $y = -\infty$. We distinguish between an original vertex of the polyhedron, which we call a *corner vertex* or just a *vertex*, and a *gridpoint*, a vertex of the grid (which might be an original vertex). A *gridedge* (*gridface*) is an edge (face) of the grid that lies on the surface of the polyhedron.

A $k_1 \times k_2 \times k_3$ *refinement* of a surface [4] starts with a grid unfolding and further partitions each gridface into a grid of edges. Positive integers k_1, k_2, and k_3 are associated with the amount of refinement in the x, y, and z dimensions, respectively; e.g., z-perpendicular gridfaces are refined into a $k_1 \times k_2$ grid, and similarly x-perpendicular (y-perpendicular) gridfaces are refined into a $k_2 \times k_3$ ($k_1 \times k_3$) grid. We will consider refinements of grid unfoldings, with the convention that a $1 \times 1 \times 1$ refinement is an unrefined grid unfolding.

Let O be a solid orthogonal polyhedron with the surface homeomorphic to a sphere (i.e., genus zero). Let Y_i be the plane $y = y_i$ orthogonal to the y-axis. Let $Y_0, Y_1, \ldots, Y_i, \ldots$ be the finite sequence of parallel planes passing through every vertex of O, with $y_0 < y_1 < \cdots < y_i < \cdots$. We define *layer* i to be the portion of O between planes Y_i and Y_{i+1}. Observe that a layer may include a collection of disjoint connected components; we call each such component a *slab*. The *band* of a slab is the connected surface piece composed of gridfaces parallel to the y axis that surround the slab. Referring to Fig. 1a, layer 0, 1, and 2 each contain one slab (with outer bands A, B, and D, respectively). Note that each slab is bounded by an outer (surface) band, but it may also contain inner bands, bounding holes. Outer bands are

[1] This was not part of the original definition in [6] but was achieved by those unfoldings.

called *protrusions* and inner bands are called *dents* (C in Fig. 1a). In other words, band A is a *protrusion* if a traversal of the rim of A in Y_i, counterclockwise from the viewpoint of $y = -\infty$, has the interior of O to the left of A, and a *dent* if this traversal has the interior of O to the right.

For each i, define $P_i = \partial O \cap Y_i$ as the portion of the surface of O lying in plane Y_i. P_i^+ is the portion of P_i with normal in direction $+y$ (composed of back faces), and P_i^- the portion with normal in direction $-y$ (composed of front faces). By convention, band points in P_i that are not incident to either front or back faces (e.g., when one band aligns with another), belong to both P_i^+ and P_i^-. Thus $P_i = P_i^+ \cup P_i^-$.

3 Dents vs. Protrusions

We observe that dents may be treated exactly the same as protrusions with respect to unfolding, because an unfolding of a 2-manifold to another surface (in our case, a plane) depends only on the intrinsic geometry of the surface, and not on how it is embedded in $\mathbb{R}^3$. Note that we are concerned only with the final unfolded "flat state" [3, 5], and not with possible intersections during a continuous sequence of partially unfolded intermediate states. Our unfolding algorithm relies solely on the amount of surface material surrounding each point: the cyclic ordering of the gridfaces incident to a vertex, and the pair of gridfaces sharing a gridedge. All these local relationships remain unchanged if we conceptually "pop-out" dents to become protrusions, i.e., a "Flatland" creature living in the surface could not tell the difference; nor can our algorithm. We note that the popping-out is conceptual only, for it could produce self-intersecting objects. Also dents are gridded independently of the rest of the object so as to avoid unnecessary surface cuts that would correspond to y-planes containing dent vertices only. From the point of view of unfolding, it does not matter whether dents are popped out or not.

Although the dent/protrusion distinction is irrelevant to the unfolding, the interrelationships between dents and protrusions touching a particular Y_i do depend on this distinction. To cite just the simplest example, there cannot be two nested protrusions to the same side of Y_i, but a protrusion could have a dent in it to the same side of Y_i (e.g., protrusion B encloses dent C to the same side of Y_1 in Fig. 1a). These relationships are crucial to the connectivity of the band graph G_b, discussed in Appendix.

4 Overview

The two algorithms we present share a common central structure, with the second achieving a stronger result; both are vertex-unfoldings that use orthogonal cuts only. We note that it is the restriction to orthogonal cuts that makes the vertex-unfolding problem difficult: if arbitrary cuts are allowed, then a general vertex-unfolding can be obtained by simply triangulating each face and applying the algorithm from [6].

The $(3 \times 1 \times 1)$-algorithm unfolds any genus-0 orthogonal polyhedron that has been refined in one direction 3-fold. The bands themselves are never split (unlike in [1]). The algorithm is simple. The $(1 \times 1 \times 1)$-algorithm also unfolds any genus-0 orthogonal polyhedron, but this time achieving a grid vertex-unfolding, i.e., without

refinement. This algorithm is more delicate, with several cases not present in the $(3 \times 1 \times 1)$-algorithm that need careful detailing. Clearly this latter algorithm is stronger, and we vary the detail of presentation to favor it. The overall structure of the two algorithms is the same:

1. A band "unfolding tree" T_u is constructed by shooting rays vertically from the top of bands. The root of T_u is a *frontmost* band (of smallest y-coordinate), with ties broken arbitrarily.
2. A forward and return *connecting* path of vertical front/back gridfaces is identified, each of which connects a parent band to a child band in T_u.
3. Each band is unfolded horizontally as a unit, but interrupted when a connecting path to a child is encountered. The parent band unfolding is suspended at that point, and the child band is unfolded recursively.
4. The vertical front and back faces of each slab are partitioned according to an illumination model, with variations for the more complex $(1 \times 1 \times 1)$-algorithm. Front/back gridfaces are attached below and above appropriate horizontal sections of the band unfolding.

The final unfolding lays out all bands horizontally, with the front and back gridfaces hanging below and above the bands. Nonoverlap is guaranteed by this strict two-direction structure.

Although our result is a broadening of that in [7] from orthostacks to all orthogonal polyhedra, we found it necessary to employ techniques different from those used in that work. The main reason is that, in an orthostack, the adjacency structure of bands yields a path, which allows the unfolding to proceed from one band to the next along this path, never needing to return. In an orthogonal polyhedron, the adjacency structure of bands is generally not linear. Thus in our algorithm, unfolding band-by-band leads to a tree traversal (e.g., Fig. 1b), which requires traversing each arc in both directions. It is this aspect which we consider our main novelty, and which leads us to hope for an extension to edge-unfoldings as well.

5 $(3 \times 1 \times 1)$-Algorithm

5.1 Computing the Unfolding Tree T_u

Define a *z-beam* to be a front or back rectangle on the surface of O whose top and bottom edges are gridedges on two bands. In the degenerate case, a z-beam has height zero and connects two rims along a section where they coincide. We say that two bands, b_i and b_j, are *z-visible* if there is a z-beam connecting a gridedge of b_i to a gridedge of b_j. There can be many z-beams connecting two bands, so for each pair of bands we select a representative z-beam of minimal (vertical) height. Let G be the graph that contains a node for each band of O and an arc for each pair (b_i, b_j) of z-visible bands such that $i \neq j$.

Lemma 1 *G is connected.*

Proof First observe that every gridface of O is either part of a band or part of a z-beam (possibly a z-beam connecting a band to itself). Now consider making vertical

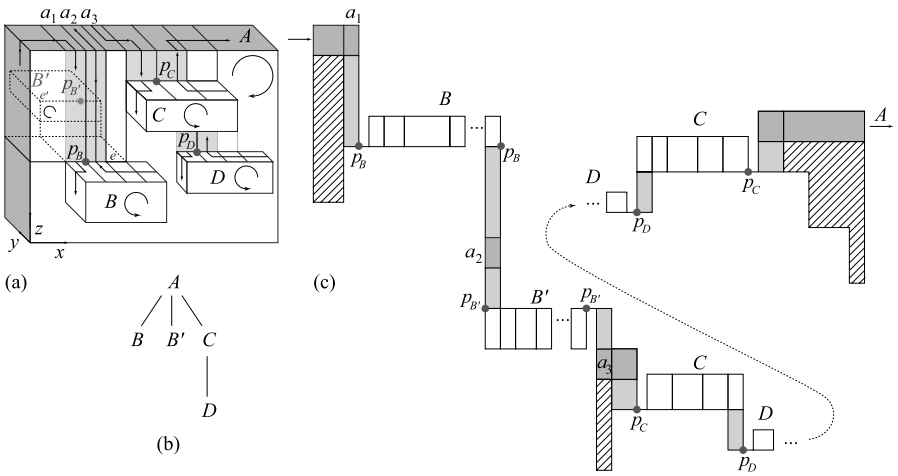


Fig. 2 **a** Orthogonal polyhedron. **b** Unfolding tree T_u. **c** Unfolding of bands and front (hachured) grid-face pieces connecting to A. Vertex connection through the pivots points p_B, $p_{B'}$, p_C, p_D is shown exaggerated for clarity

cuts on the surface of O along the extent of the left and right sides of each z-beam. Since O is connected and only vertical cuts are made, the resulting structure remains connected and can be viewed as a multigraph, where bands are nodes and z-beams are edges. Since G is the subset of this multigraph obtained by removing self-loops and duplicate edges, G is also connected. □

Let the unfolding tree T_u be any spanning tree of G, with the root selected arbitrarily from among all bands adjacent to Y_0. We apply the $3 \times 1 \times 1$ refinement procedure to partition each front, back, top, and bottom gridface of O into three congruent subfaces, by adding two new gridedges orthogonal to the x-axis. This partitions the top and bottom edges of each z-beam into three refined gridedges and divides the beam itself into three vertical columns of refined gridfaces. See Fig. 2a. Let A be an arbitrary band, let B be one of its children in T_u, and let e be the gridedge on B's rim where the z-beam from A attaches. We define the *pivot point* p_B for band B to be the $\frac{1}{3}$-point of e (or, in circumstances to be explained later, the $\frac{2}{3}$-point), and so it coincides with a point of the $3 \times 1 \times 1$-refined grid. The unfolding of O will follow the connecting vertical ray that extends from p_B on B to A. Note that if e belongs to both A and B, then the ray connecting A and B degenerates to a point. To either side of a connecting ray we have two *connecting paths* of gridfaces, the *forward* and *return* path. In Fig. 2a, these connecting paths are the shaded strips on the front face of A.

5.2 Unfolding Bands into a Net

Starting at a frontmost *root band*, each band is unfolded as a conceptual unit, but interrupted by the connecting rays incident to it from its front and back faces. In Fig. 2, band A is unfolded as a rectangle, but interrupted at the rays connecting to

(front children) B, C and (back child) B'. At each such ray the parent band unfolding is suspended, the unfolding follows the forward connecting path to the child, the child band is recursively unfolded, then the unfolding returns along the return connecting path back to the parent, resuming the parent band unfolding from the point it left off.

Figure 2 illustrates this unfolding algorithm. The clockwise unfolding of A, laid out horizontal to the right, is interrupted to traverse the forward path down to B, and B is then unfolded as a rectangle (composed of its contiguous gridfaces). The base p_B of the connecting ray is called a *pivot point* because the counterclockwise unfolding of B is rotated 180° counterclockwise about p_B so that the unfolding of B is also to the right. It is only here that we use point-connections that render the unfolding a vertex-unfolding. The unfolding of B proceeds counterclockwise back to p_B, crosses over A to unfold B', then a clockwise rotation by 180° around the second image of pivot $p_{B'}$ orients the return path to A so that the unfolding of A continues to the right. Note that the unfolding of C is itself interrupted to unfold child D. Also note that there is edge overlap in the unfolding at each of the pivot points, and this overlap could not be eliminated without violating the condition that all surface pieces face the same way (up, in our case).

The reason for the $3 \times 1 \times 1$ refinement is that the upper edge e' of the back child band B' has the same (x, z)-coordinates as the upper edge e of B on the front face. In this case, the gridfaces of band A induced by the connecting paths to B would be "overutilized" if there were only two. Let a_1, a_2, a_3 be the three faces of A induced by the $3 \times 1 \times 1$ refinement of the connecting path to B, as in Fig. 2. Then the unfolding path winds around A to a_1, follows the forward connecting path to B, returns along the return connecting path to a_2, crosses over A and unfolds B' on the back face, with the return path now joining to a_3, at which point the unfolding of A resumes. In this case, the pivot point $p_{B'}$ for B' is the $\frac{2}{3}$-point of e'. Other such conflicts are resolved similarly. It is now easy to see that the resulting net has the general form illustrated in Fig. 2c:

1. The faces of each band fall within a horizontal rectangle whose height is the band width.
2. These band rectangles are joined by front/back connecting paths on either side, connecting through pivot points.
3. The strip of the plane above and below each band face that is not incident to a connecting path, is empty.
4. The net is therefore an orthogonal polygon monotone with respect to the horizontal.

5.3 Attaching Front and Back Faces to the Net

Finally, we "hang" front and back faces from the bands as follows. The front face of each band A is partitioned by imagining A to illuminate downward lightrays from the rim in the front face. The pieces that are illuminated are then hung vertically downward from the horizontal unfolding of the A band. The portions unilluminated will be attached to the obscuring bands.

In the example in Fig. 2, this illumination model partitions the front face of A into three pieces (the striped pieces in Fig. 2c). These three pieces are attached under A;

the portions of the front face obscured by B but illuminated downward by B are hung beneath the unfolding of B (not shown in the figure), and so on. Because the vertical illumination model produces vertical strips, and because the strips above and below the band unfoldings are empty, there is always room to hang the partitioned front face. Thus, any orthogonal polygon may be vertex-unfolded with a $3 \times 1 \times 1$ refinement of the vertex grid.

Although we believe this algorithm can be improved to $2 \times 1 \times 1$ refinement, the complications needed to achieve this are similar to what is needed to avoid refinement entirely, so we instead turn directly to $1 \times 1 \times 1$ refinement.

6 (1 × 1 × 1)-Algorithm

Although the $(1 \times 1 \times 1)$-algorithm follows the same general outline as the $(3 \times 1 \times 1)$-algorithm, there are significant complications, which we outline before going into detail. First, without the refinement of z-beams into three strips to allow avoidance of conflicts on opposite sides of a slab (e.g., B and B' in Fig. 2a), we found it necessary to replace the z-beams by a pair of z-rays that are in some sense the boundary edges of a z-beam. Selecting two rays per band permits a 2-coloring algorithm (Theorem 4) to identify rays that avoid conflicts. Generating the ray-pairs (Sect. 6.1.1) requires care to ensure that the band graph G_b is connected (Appendix). This graph, and the 2-coloring, lead to an unfolding tree T_u (Sect. 6.2). From here on, there are fewer significant differences compared to the $(3 \times 1 \times 1)$-algorithm. Without the luxury of refinement, there is more need to share vertical paths on the front or back face of a slab (Fig. 11). Finally, the connecting paths obscure the illumination of some grid faces, which must be attached to the connecting paths. We now present the details, in this order:

1. Determine Conflict-Free Pivot Points (Sect. 6.1) via
    a. Ray-Pair Generation (Sect. 6.1.1)
    b. Ray Graph (Sect. 6.1.2)
2. Construct T_u (Sect. 6.2)
3. Select Connecting Paths (Sect. 6.2.1)
4. Determine Unfolding Directions (Sect. 6.2.2)
5. Recurse:
    a. Unfold Bands into a Net (Sect. 6.3)
    b. Attach Front and Back Faces to the Net (Sect. 6.4)

6.1 Determining Conflict-Free Pivot Points

The pivot p_A for a band A is the gridpoint of A where the unfolding of A starts and ends. The y-edge of A incident to p_A is the first edge of A that is cut to unfold A.

Let A be an arbitrary band delimited by planes Y_i and Y_{i+1}. Say that two gridpoints $u \in Y_i$ and $v \in Y_{i+1}$ are *in conflict* if the upward rays emerging from u and v hit first the endpoints of the same y-edge of A; otherwise, u and v are *conflict-free*. If u lies either on a vertical edge, or on a vertically extreme horizontal edge, then the ray at u degenerates to u itself.

Our goal is to select conflict-free pivots for all bands in T_u, which will help us later avoid competition over the use of certain gridfaces in the unfolding, an issue that will become clear in Sect. 6.3. Selecting these pivots is the most delicate aspect of the $(1 \times 1 \times 1)$-algorithm. Ultimately, we represent pivoting conflicts in the form of a graph G_r (Sect. 6.1.2), from which T_u will be derived.

6.1.1 Ray-Pair Generation

In order to avoid pivoting conflicts, for each band we will need two choices for its connecting ray. Thus the algorithm generates the rays in pairs. Because there is no refinement, the two rays originate at grid points on the same band, but they may terminate on different bands. A simple example is shown in Fig. 3a, where the ray pair originating on band D hits two different bands, B and C. This example also suggests that one cannot consider ray pairs connecting pairs of bands, as in the $(3 \times 1 \times 1)$-algorithm (which would connect D to A in this example), but instead we focus on shooting pairs of rays upward from strategic locations on the boundary of each band, and then selecting a subset of these rays so that the conflicts can be resolved and T_u is connected. To ensure connectedness of all bands, several ray-pairs must be issued upward from each band. Figure 3b shows an example: no pair of rays can emanate upward from the top of $B \cap P_i^-$ or $C \cap P_i^-$; one pair of rays shoots upward from the top of each component of $A \cap P_i^-$: (r_1, r_2) connects A to B and (r_3, r_4) connects A to C; finally, one pair of rays (r_5, r_6) issues from the top of $A \cap P_i^+$, which connects A to D. So, overall, three pairs of rays are generated for band A. We now turn to describing in detail the method for generating ray-pairs.

Let band A intersect plane Y_i. The algorithm is a for-loop over all A. We identify *chunks*, $A_1, A_2, \ldots, A_m$, of the rim $A \cap Y_i$, where each chunk A_j is a connected component of either $A \cap P_i^-$ or $A \cap P_i^+$ that contains at least one horizontal gridedge. (Note that these chunks do not necessarily cover $A \cap Y_i$.) We define $S(A_j)$ as the set of all vertical segments $s = (a, b)$, with $a \in A_j$, such that

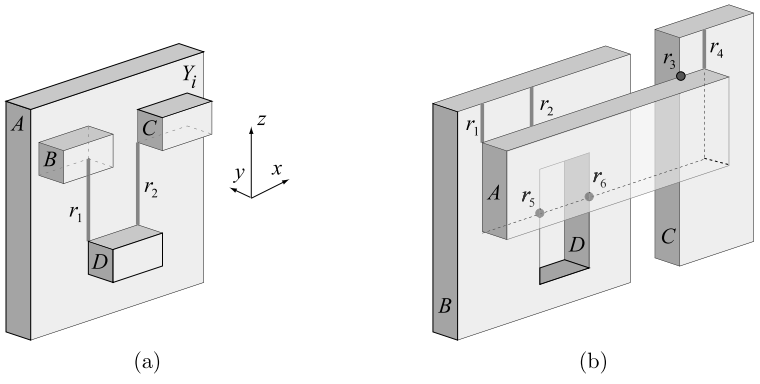


Fig. 3 **a** The ray pair (r_1, r_2) connects band D to two different bands B and C. **b** To ensure connectivity, three pairs of rays must be issued for A: (r_1, r_2), (r_3, r_4), and (r_5, r_6)

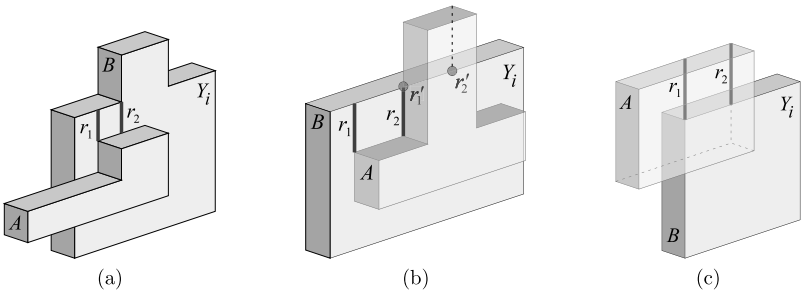


Fig. 4 Generating ray-pairs: **a** (r_1, r_2) for A; $S(B) = \emptyset$. **b** (r_1, r_2) for A (note that r_2 runs along source band A); degenerate ray-pair (r_1', r_2') for B. **c** $S(A) = \emptyset$; (r_1, r_2) for B

1. s is either a point, with $b = a$, or a front/back segment, with a below b.
2. $b \in B$ for some band $B \neq A$.
3. The open segment $s \setminus \{a, b\}$ may contain points of A (see r_2 in Fig. 4b), but no points of other bands.

For each band A, for each chunk $A_j \subseteq A$, if $S(A_j)$ contains at least two rays connecting A to the same band B, we select one ray pair (r_1, r_2) that satisfies two restrictions: (i) among the segments in $S(A_j)$ incident to a highest x-gridedge in A_j, r_1 is the leftmost one, and (ii) r_2 is the segment one x-gridedge to the right of r_1. Figure 4 shows a few examples. As mentioned earlier, several ray pairs could be generated for any one band, and indeed several pairs could connect two bands (e.g., see Fig. 4b where bands A and B are connected by two ray pairs).

Let G_b be the *band graph* whose nodes are bands. Two bands are connected by an arc in G_b if the ray-pair algorithm generates a ray connecting them. We call a collection of bands in G_b *ray-connected* if they are in the same connected component of G_b. We establish that G_b is a connected graph, i.e., all bands are ray-connected to one another, even if only one ray per pair is employed:

Lemma 2 *G_b is connected. Furthermore, the subgraph of G_b induced by exactly one ray per ray-pair (arbitrarily selected) is connected.*

Whereas the connectedness of bands by z-beams in the $(3 \times 1 \times 1)$-algorithm is straightforward, the complex possible relationships between bands makes connectedness via rays more subtle. We relegate the proof to the Appendix (Appendix) in order to not interrupt the main flow of the algorithm.

The over-generation of ray-pairs noted above is designed to ensure connectedness. Eventually many rays will be discarded by the time T_u is constructed in Sect. 6.2.

6.1.2 Ray Graph G_r

One pair of rays per pair of bands suffices to ensure that all bands are ray-connected. If multiple pairs of rays exist for a pair of bands, pick one pair arbitrarily and discard the rest. Then define a ray graph G_r as follows. The nodes of G_r are vertical rays, perhaps degenerating to points, connecting gridpoints between two bands that both intersect a common Y_i plane. The arcs of G_r record two types of potential pivoting conflicts:

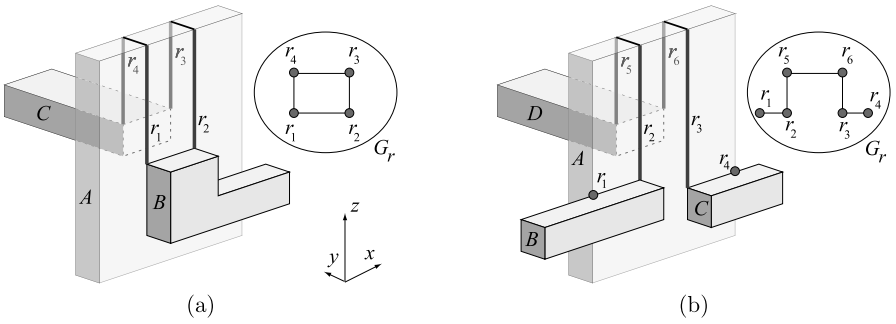


Fig. 5 Building G_r. **a** G_r is a 4-cycle; $\{r_1, r_2\}$ and $\{r_3, r_4\}$ are x-arcs and the others are y-arcs. **b** G_r is a path; $\{r_2, r_5\}$ and $\{r_3, r_6\}$ are y-arcs and the others are x-arcs

(i) The nodes for each pair of rays issuing from the top of a band B are adjacent in G_r. Call such arcs *x-arcs*; geometrically they can be viewed as parallel to the x-axis.
(ii) The nodes for two rays incident to opposite sides of the rim of a band A, connected by a y-segment on the band, are adjacent in G_r. Call such arcs *y-arcs*; geometrically they can be viewed as parallel to the y-axis.

Figure 5 shows two simple examples of G_r involving nodes on opposite sides of one band A. Before proceeding, we list the consequences of the two types of arcs in G_r. Assuming that we can 2-color G_r {red, blue}, and we select the base of (say) the red rays as pivots, then: (i) exactly one pivot is selected for each band, and (ii) no two pivot rays are in conflict across a band. So our goal now is to show that G_r is 2-colorable. Because a graph is 2-colorable if and only if it is bipartite, and a graph is bipartite if and only if every cycle is of even length, we aim to prove that every cycle in G_r is of even length. We start by listing a few relevant properties of G_r:

1. Every node $r \in G_r$ has exactly one incident x-arc. The rays are generated in pairs, and the pairs are connected by an x-arc. As no such ray is shared between two bands, at most one x-arc is incident to any r.
2. Nodes have at most degree 3, with the following structure: degree-1 nodes have an incident x-arc; degree-2 nodes have both an incident x- and y-arc; and degree-3 nodes have an incident x-arc and two incident y-arcs.
3. Each x-arc spans exactly one pair of adjacent y-gridlines, and each y-arc spans exactly one band rim-to-rim. The former is by the definition of ray pairs, which issue from adjacent gridpoints, and the latter follows from the grid partitioning of the object into bands.

Our next step requires embedding G_r in an xy-plane Π. Toward that end, we coordinatize the nodes and arcs of G_r as follows. A node $r \in G_r$ is a z-ray, and is assigned the (x, y) coordinates of the ray. Note that this means collinear rays get mapped to the same point; however, we treat them as distinct. The x-arcs are then parallel to the x-axis, and the y-arcs are parallel to the y-axis. In essence, this coordinatization is a view from $z = +\infty$.

Figure 6 shows a more complex example illustrating this viewpoint. The object is composed of 7 bands B_i, one of which (B_3) is a dent. There are 12 ray nodes,

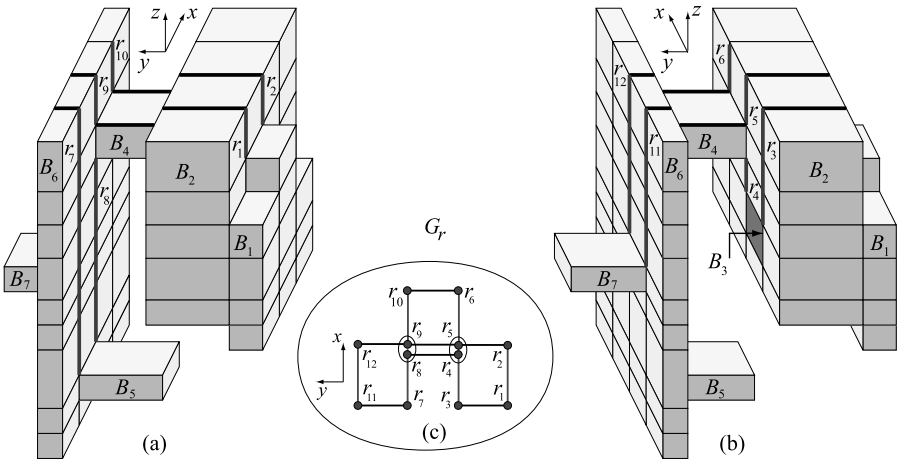


Fig. 6 **a**, **b** Two side views of an object; z-rays and y-arcs are marked with thick lines. **c** G_r coordinatized into xy-plane Π; (r_5, r_6, r_{10}, r_9) is a 4-cycle; $(r_1, r_3, r_4, r_8, r_7, r_{11}, r_{12}, r_9, r_5, r_2)$ is a 10-cycle

two pairs of which lie on the same z-vertical line, namely (r_4, r_5) and (r_8, r_9). Note that there are y-arcs crossing both the top of and the bottom[2] of B_4. The graph G_r has a 4-cycle and a 10-cycle, both detailed in the caption (as well as a 12-cycle not detailed).

Lemma 3 *Every cycle in G_r is of even length.*

Proof Let C be a cycle in G_r. The coordinatization described above maps C to a (perhaps self-crossing) closed path in the xy-plane Π, a path which may visit the same (x, y) point more than once, and/or traverse the same edge in Π more than once. Any such closed path on a grid must have even length, for the following reason.

First, by Property (3) above, each edge of the path in Π connects adjacent grid lines: an edge never "jumps over" one or more grid lines. Second, any such closed lattice path changes parity with each step, in the following sense. Number the x- and y-gridlines with integers $0, 1, 2, \ldots$ left to right and bottom to top, respectively. Define the parity of a gridpoint of Π to be the sum of its x- and y-gridline coordinates, mod 2. Then each step of the path, necessarily in one of the four compass directions, changes parity, as it changes only one of x or y. Returning to the start point to close the path must return to the starting coordinates, and so to the same parity. Thus, there must be an even number of parity changes along any closed path. Therefore, C has an even number of edges. □

We have now established this:

Theorem 4 *G_r is 2-colorable.*

[2] A dent is included in this example precisely to introduce such a bottom y-arc into G_r.

Note that nowhere in the above proof do we assume genus zero, so this theorem holds for polyhedra of arbitrary genus.

Band Pivoting We are finally ready to specify the pivot points. By Theorem 4, we can 2-color the nodes of G_r {red, blue}. We choose all red ray-nodes of G_r to be pivoting rays, in that their base points become pivot points. As remarked before, this selection guarantees that each band is pivoted, and no two pivots are in conflict. For the root band we choose a pivot point—the point at which the unfolding starts and ends—to be a grid point on the front rim connected by a y-segment to a blue ray. Because the rays are generated in pairs, there must be a blue ray incident to the root band. This choice guarantees that the root pivot is not in conflict with any other (necessarily red) pivot.

6.2 Unfolding Tree T_u

The next task is to define a band spanning tree T_u, based on the band graph G_b. Define G'_b, to retain just the arcs of G_b corresponding to the red ray nodes (in the above 2-coloring) in G_r. This maintains the connectivity by Lemma 2. Then take T_u to be any spanning tree of G'_b rooted at a frontmost band. The arcs in T_u and their associated rays thus determine a pivot point for each band.

With T_u finally in hand, the remainder of the $(1 \times 1 \times 1)$-algorithm follows the overall structure of the $3 \times 1 \times 1$ algorithm, with variations as mentioned before, as detailed below.

6.2.1 Selecting Connecting Paths

Having established a pivot point for each band, we are now ready to define the *forward* and *return* connecting paths for a child band in T_u. A "path" here refers to a connected sequence of gridfaces that the unfolding follows to get from one band to another. Let B be an arbitrary child of band A. If the pivot point p_B of B is at the intersection of B and A, then both forward and return connection paths for B reduce to point p_B (see Fig. 7). If B does not intersect A, then a ray r connects p_B to A (Figs. 8a and 10a). The connecting paths are the two vertical paths separated by r composed of the gridfaces sharing an edge with r (paths k_1 and k_2 in Figs. 8a and 10a). The path first encountered in the unfolding of A is used as a forward connecting path; the other path is used as a return connecting path.

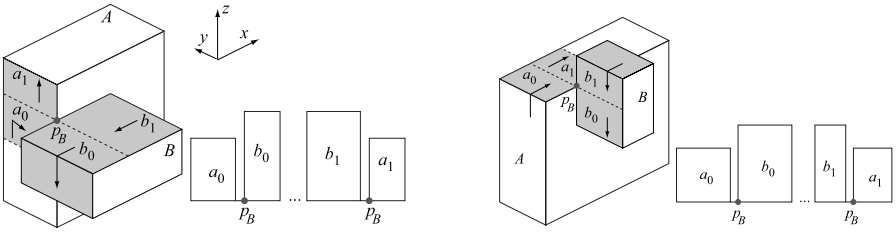


Fig. 7 Unfolding B when the ray connecting B to A degenerates to p_B

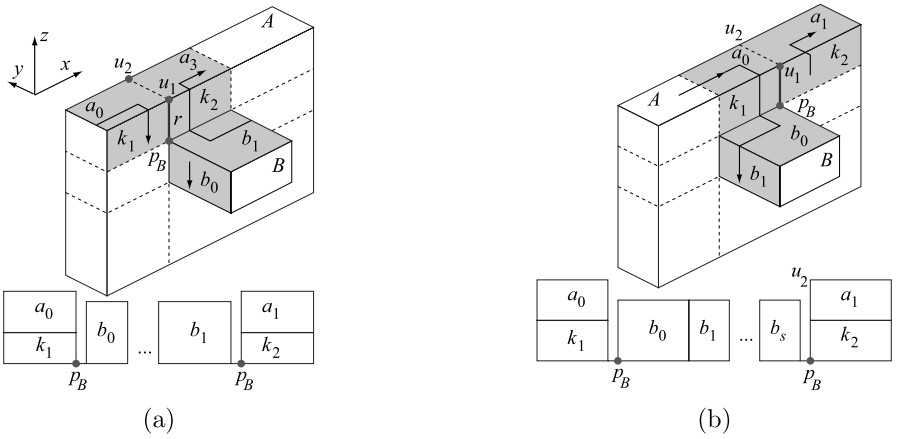


Fig. 8 Unfolding B: u_1 is not a corner vertex of A **a** p_B incident to a left gridface of B **b** p_B incident to a top gridface of b

6.2.2 Determining Unfolding Directions

A top-down traversal of T_u assigns an unfolding direction to each band in T_u as follows. The root band in T_u may unfold either clockwise or counterclockwise, but for definiteness we set the unfolding direction to clockwise. Let B be the band in T_u currently visited and let A be the parent of B. If the upward ray r incident to p_B connects B to a bottom gridpoint of A, then B unfolds in the same clockwise/counterclockwise direction as A. Otherwise, r connects B to a top or a side (for degenerate rays) gridpoint of A; in this case, B unfolds in the direction opposite to that of A. In other words, A and B unfold in the same direction if B "hangs below" A, and in opposite direction otherwise.

6.3 Unfolding Bands into a Net

Let A be a band to unfold, initially the root band. The unfolding of A starts at its pivot point p_A and proceeds in the unfolding direction (clockwise or counterclockwise) of A. Henceforth we assume without loss of generality that the unfolding of A proceeds clockwise (with respect to a viewpoint at $y = -\infty$); the counterclockwise unfolding of A is a vertical reflection of the clockwise unfolding of A. In the following we describe our method to unfold every child B of A recursively. As mentioned earlier, each band unfolds horizontally, from left to right, with recursive interruptions to unfold its children.

Without loss of generality, we assume that A and B are both protrusions (cf. Sect. 3). The possible unfoldings for a child B fall naturally into three cases. Case 1 handles the situation when B's pivot is at the intersection of A and B. Cases 2 and 3 handle situations when B's pivot is connected by a ray to A; Case 2 deals with situations in which B's connecting paths do not overlap any other connecting paths, and Case 3 addresses overlapping paths.

Case 1: Pivot $p_B \in A \cap B$. Then, whenever the unfolding of A reaches p_B, we unfold B as in Fig. 7. The unfolding uses the two band gridfaces of A incident to p_B

(a_0 and a_1 in Fig. 7). Let b_0 be the first gridface of B in counterclockwise order about p_B. In the unfolding, we rotate b_0 around p_B so that the counterclockwise unfolding of B extends horizontally to the right. The unfolding of B proceeds counterclockwise back to p_B, then the band gridface a_1 incident to p_B is oriented about p_B so that the unfolding of A continues to the right.

Note that, because the pivots of any two children of A are conflict-free, there is no competition over the use of a_0 and a_1 in the unfolding. Note also that the unfolding path does not self-cross. For example, the cyclic order of the gridfaces incident to p_B in Fig. 7a is $(a_0, A_{\text{front}}, b_0, b_1, A_{\text{back}}, a_1)$, and the unfolding path follows $(a_0, b_0, \ldots, b_1, a_1)$.

Case 2: Pivot $p_B \notin A \cap B$ and the (forward, return) connecting paths for B do not overlap other connecting paths (except at their boundaries); we will later see that connecting paths may overlap. Let us settle some notation first (cf. Fig. 8a): r is the ray connecting B to A; k_1 and k_2 are forward and return connecting paths for B (one to either side of r); u_1 is the endpoint of r that lies on A; and u_2 is the other endpoint of the y-edge of A incident to u_1. We discuss three situations:

Case 2a: u_1 is neither a reflex corner nor a bottom corner of A. In this case, whenever the unfolding of A reaches k_1, the unfolding of B proceeds according to one of three subcases, depending on the position of p_B. If p_B touches a left gridface of B, the unfolding proceeds as in Fig. 8a, and if it touches a right gridface, the unfolding proceeds as in Fig. 8b. In both cases, b_0, the first gridface of B in counterclockwise order around p_B, is rotated so that the unfolding of B extends to the right, B is recursively unfolded, and the return path k_2 is rotated about p_B so that the unfolding of A continues to the right. The final subcase occurs when p_B touches only top gridfaces of B. Then the unfolding is identical to that in Fig. 8b but with b_s a top gridface.

Case 2b: u_1 is a reflex corner of A. In this case, the unfolding of B proceeds as in Fig. 9a,b. It is the existence of the vertical strip incident to u_1 (marked t in Fig. 9) that

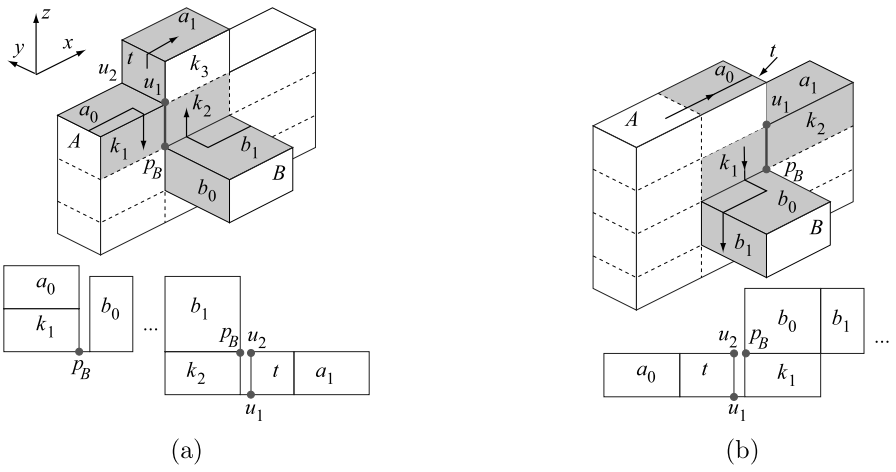


Fig. 9 Unfolding B: u_1 is a corner vertex of A. **a** t is a left strip **b** t is a right strip

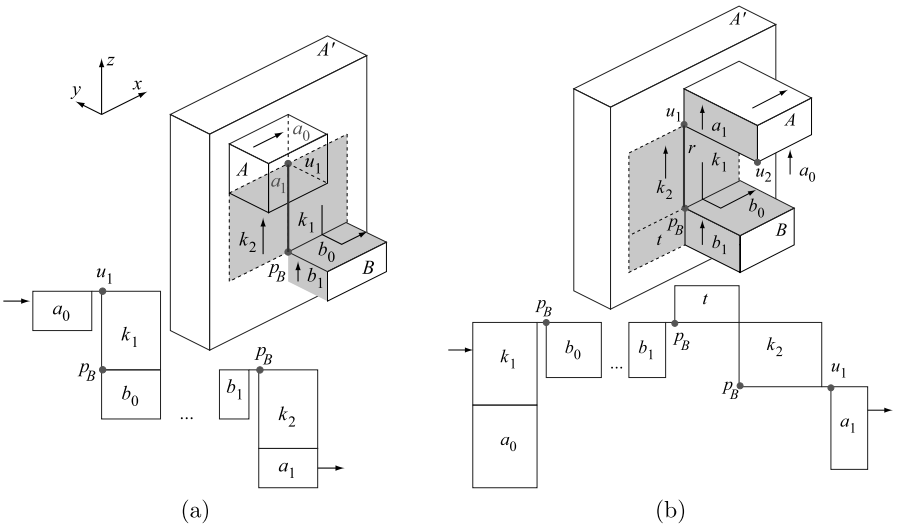


Fig. 10 Unfolding B: u_1 is a bottom corner of A **a** rightmost, and **b** leftmost gridface of A vertically aligned with leftmost gridface of B

makes handling this case different from Case 2a. Note, however, that the existence of t implies the existence of at least two gridfaces on either the return path or the forward path for B, depending on whether t is a left (Fig. 9a) or a right (Fig. 9b) strip of gridfaces. In the former case the unfolding starts as in Case 2a (Fig. 9a), and once the unfolding of B returns to p_B, it continues along the return path k_2 up to u_1, then unfolds t and orients it about u_1 so that the unfolding of A continues to the right. The gridface(s) that cover the gap above k_2 (marked k_3 in Fig. 9a) will be attached below the adjacent top gridface of A (a_1 in Fig. 9a) in the last phase of the unfolding algorithm (Sect. 6.4).

If t is a strip of right gridfaces, then we unfold t before descending along the forward path down to B, as in Fig. 9b (note the vertical symmetry with the unfolding in Fig. 9a); the unfolding of B then proceeds as in Case 2a (Fig. 8b).

Case 2c: u_1 is a bottom corner of A. In this case, the unfolding proceeds as in Fig. 10a or 10b, depending on whether u_1 is a right or a left bottom corner of A. The unfolding illustrated in Fig. 10a follows the familiar unfolding pattern: orient the first gridface of B in counterclockwise order around p_B so that the unfolding of B extends to the right; once the unfolding of B returns to p_B, follow the return path back to A and unfold the gridface of A clockwise to the right of u_1 (a_1 in Fig. 10a) so that the unfolding of A continues to the right. A similar pattern applies to the case illustrated in Fig. 10b, with one subtle difference meant to aid in unfolding front and back faces (discussed in Sect. 6.4): in unfolding bands, we aim at maintaining the vertical position of the (forward, return) connecting paths in the unfolding, so that vertical strips hanging below these connecting paths in 3D could also hang vertically in the 2D unfolding. More on this in Sect. 6.4. Observe that the z-vertical edges of k_1 and k_2 from Fig. 10a hang remain vertical in the unfolding. However, the z-vertical edges of k_2 from Fig. 10b must unfold as horizontal edges, otherwise it would not

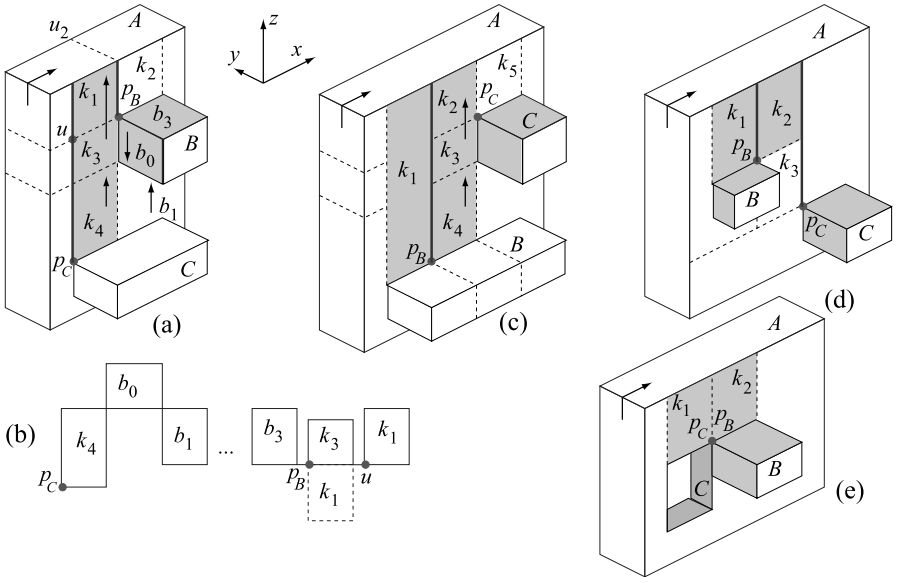


Fig. 11 **a** Return path for C includes k_4, k_3, k_1; forward path for B is k_1. **b** Unfolding for (**a**). **c** Return path for B includes k_5, k_4, k_2; forward path for C is k_2. **d** Return path for B is k_2; forward path for C includes k_2, k_3. **e** Forward (return) paths are identical for B and C

be possible to orient a_1 around u_1 so as to continue unfolding A to the right of k_2. This is the reason for employing the gridface strip marked t in the unfolding, so that z-vertical sides of t remain vertical in the unfolding, and any gridface strip hanging below t could be attached to t vertically in the unfolding.

We note that Fig. 10 illustrates only the situation in which p_B is incident to a left gridface of B, but it should not be difficult to observe that the same idea applies to any top pivot of B; the pivot position only affects the start and end unfolding position of B, and everything else remains the same.

Case 3: Pivot $p_B \notin A \cap B$ and a connecting path for B overlaps a connecting path for another descendant C of A. This case is slightly more complex, because it involves conflicts over the use of the connecting paths for B. The following three situations are possible.

Case 3a: The forward path k_1 for B overlaps the return path for another descendant C of A. This situation is illustrated in Fig. 11a. In this case, the unfolding of B starts as soon as the unfolding along the return path from C to A meets a gridface of B incident to p_B (gridface b_0 in Fig. 11a). At this point we recursively unfold B as before (see Fig. 11b), then the unfolding continues along the return path for C back to A. Figure 11b shows gridface k_1 in two positions: we let k_1 hang down only if the next gridface to unfold is a right gridface of a child of A (see also the transition from k_7 to c_5 in Fig. 12); otherwise, use k_1 in the upward position, a freedom permitted to us by rotating about vertex u.

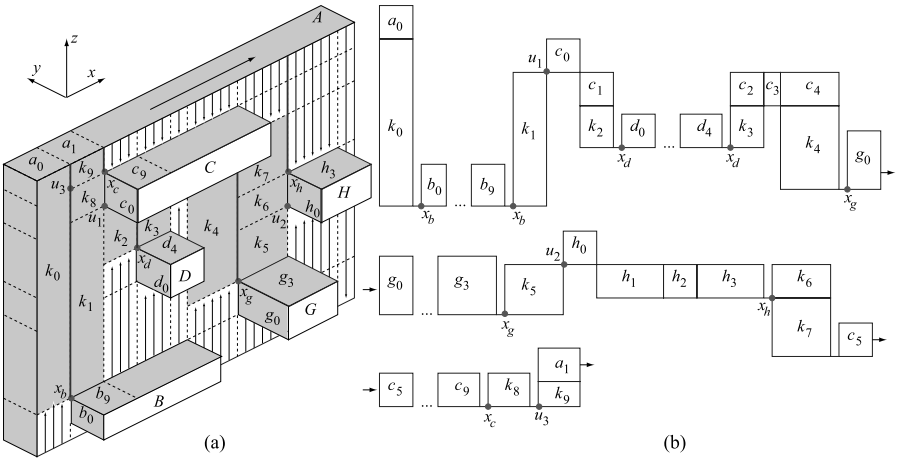


Fig. 12 **a** An example. **b** The vertex-unfolding

Case 3b: The return path for B overlaps the forward path for another descendant C of A. This situation is illustrated in Figs. 11c and 11d. The case depicted in Fig. 11c is similar to the one in Fig. 11a and is handled in the same manner. For the case depicted in Fig. 11d, notice that k_2 is on both the forward path for C and the return path for B. However, no conflict occurs here: from k_2 the unfolding continues downward along the forward path to C and unfolds C next.

Case 3c: The forward path k_1 for B overlaps the forward path for another descendant C of A. This situation occurs when either B or another band C incident to B is a dent, as illustrated in Fig. 11e. Again, no conflict occurs here: the recursive unfolding of C, which returns to $p_C = p_B$, is followed by the recursive unfolding of B, which returns to p_B, then the unfolding continues along the return path for B (C) back to A. We note that the forward paths for B and C overlap if and only if their reverse paths overlap, so this case also handles the situation in which the reverse paths overlap.

Figure 12 shows a more complex example that emphasizes these subtle unfolding issues. Note that the return path k_1, k_8, k_9 for B overlaps the forward path k_9 for C; and the return path k_5, k_6 and k_7 for G overlaps the forward path for H, which includes k_7. The unfolding produced by the method described in this section is depicted in Fig. 12b.

6.4 Attaching Front and Back Faces to the Net

Front and back faces of a slab are "hung" from bands following the basic idea of the illumination model discussed in Sect. 5.3. There are three differences, however, caused by the employment of some front and back gridfaces for the connecting paths, which can block illumination from the bands.

1. We illuminate both upward and downward from each band: each x-edge illuminates the vertical front/back face it attaches to. This alone already suffices to han-

dle the example in Fig. 12: all front and back faces are illuminated downward from the top of A, upward from the bottom of A, and upward from the top of B.

2. Some gridfaces still might not be illuminated by any bands, because they are obscured both above and below by paths in connecting gridfaces. Therefore we incorporate the connecting gridfaces into the band for the purposes of illumination. For example, in Fig. 10a, k_2 illuminates downward and k_1 illuminates upward. The reason this strategy works is that, with one exception, each vertical connecting strip remains vertical in the unfolding, and so illuminated strips can be hung safely without overlap. Note that although k_2 illuminates downward, it is rotated about p_B so that what was down in 3D becomes up in the unfolding. So the faces illuminated downward from k_2 get "hung upward."
3. The one exception is the return connecting path k_2 in Fig. 10b. This paths unfolds "on its side," i.e., what is vertical in 3D becomes horizontal in 2D. Note, however, that the gridface t below such a path (a gridface always present), is oriented vertically. We thus consider t to be part of the connecting path for illumination purposes, permitting the strip below to be hung under t.

Because our cases are exhaustive, all gridfaces of (say) the front face of A are either illuminated by A, or by some descendant of A on the front face, augmented by the connecting paths as just described. (In fact every gridface is illuminated twice, from above and below.) Hanging the strips then completes the unfolding.

6.5 Algorithm Complexity

Because there are so few unfolding algorithms, that there is *some* algorithm for a class of objects is more important than the speed of the algorithm. Nevertheless, we offer an analysis of the complexity of our algorithm. Let n be the number of corner vertices of the polyhedron, and $N = O(n^2)$ be the number of gridpoints. The vertex grid can be easily constructed in $O(N)$ time, leaving a planar surface map consisting of $O(N)$ gridpoints, gridedges, and gridfaces. The computation of connecting rays (Sect. 6.2) requires determining the components of $A \cap P_i^+$ and $A \cap P_i^-$, for each band A and incident plane Y_i. These can be easily read from the planar map by running through the n vertices of each of the $O(n)$ bands and determining, for each vertex, whether it belongs to P_i^+ or P_i^-. Each of the $O(n)$ band components shoots a vertical ray from one corner vertex, in a 2D environment (the plane Y_i) of n noncrossing orthogonal segments. Determining which band a ray hits involves a ray-shooting query. Although an implementation would employ an efficient data structure, perhaps BSP trees [9], for complexity purposes the naive $O(n)$ query cost suffices to lead to $O(n^2)$ time to construct G_r. Selecting pivots (Sect. 6.1) involves 2-coloring G_r in $O(n)$ time, and computing the unfolding tree T_u in a breadth-first traversal of G_r, which takes $O(n)$ time. Unfolding bands (Sect. 6.3) involves a depth-first traversal of T_u in $O(n)$ time, and laying out the $O(N)$ gridfaces in $O(N)$ time. Thus, the algorithm can be implemented to run in $O(N) = O(n^2)$ time.

7 Further Work

Extending these algorithms to arbitrary genus orthogonal polyhedra remains an interesting open problem. Holes that extend only in the x and z directions within a slab

seem unproblematic, as they simply disconnect the slab into several components. Holes that penetrate several slabs (i.e., extend in the y direction) present new challenges, as they may obstruct vertical band visibility necessary to establish that the band graph is connected. One idea to handle such holes is to place a virtual xz-face midway through the hole, and treat each half-hole as a dent (protrusion).

Acknowledgements We thank the anonymous referees for their careful reading and insightful comments.

Appendix: Proof of Lemma 2 (Connectedness of G_b)

For a band A, let $r_i(A)$ be the closed region of Y_i whose boundary is the rim of A, i.e., $A \cap Y_i$. Two subsets of $P_i = \partial O \cap Y_i \subset Y_i$ are *path-connected*, or just *connected*, if there are points in each that are connected by a path that lies in P_i. We first develop notation to describe the relevant portions of $r_i(A)$ that are connected to each band A. Recall from Sect. 2 that P_i^+ is composed of back faces and P_i^- of front faces.

We decompose the set of points in P_i into sets $c_i(A)$ for all bands A that meet P. The sets $c_i(A)$ will have disjoint interiors, overlapping only on their boundaries. Initially assign $c_i(A) = A \cap P_i$; we now augment these sets. Let p be an arbitrary point in P_i. We consider four cases, which ultimately reduce to a single case. First let p be on a front face, i.e., on P_i^-. Then p is either on a protrusion that lies behind Y_i (Fig. 13a), or on a dent in front of Y_i (Fig. 13b). Symmetric cases occur when p is on a back face (on P_i^+), either on a protrusion in front of Y_i (Fig. 13c), or a dent behind Y_i (Fig. 13d). Let p be on a front face of a band A (encompassing the first two cases). If p is path-connected to A, we add p to $c_i(A)$. Otherwise, p must be in $r_i(B)$ of a unique dent band B, which is itself in a protrusion B', both in front of Y_i. In this case, we add p to $c_i(B)$. For example, in Fig. 17a, p lies on the front face of A and is path-connected to A, and therefore $p \in c_i(A)$ (even though it is also path-connected to the surrounding dent B). In Fig. 18a, however, p lies on the front face of A' but is not path-connected to A', and therefore p is instead in the set $c_i(B)$ for the surrounding dent B. Figure 16b illustrates the symmetric case where p is on the back face of protrusion B', and because p is path-connected to B', $p \in c_i(B')$ (even though p is also path-connected to B).

The above definition of $c_i(A)$ ensures that $\cup c_i(A) = P_i$, where the union is over all A that meet P_i. Moreover the c_i sets have disjoint interiors. We now concentrate

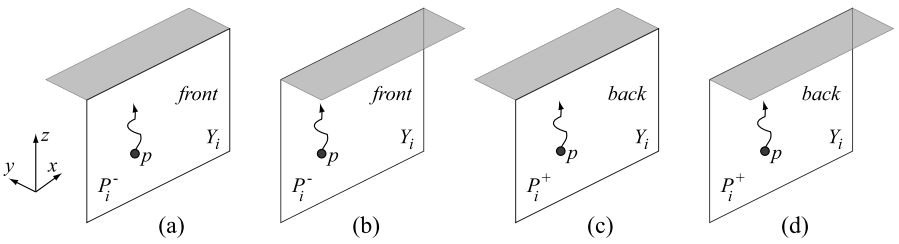


Fig. 13 Four cases: p is located on: **a** front face of protrusion behind Y_i, **b** front face of dent in front of Y_i, **c** back face of protrusion in front of Y_i, **d** back face of dent behind Y_i

on the boundaries of the c_i sets, and raise the observation we need to a lemma for later reference:

Lemma 5 *For protrusion A and dent B on opposite sides of Y_i such that $c_i(A) \cap c_i(B)$ is nonempty, it must be that $A \cap B$ is nonempty, i.e., the band rims share one or more points.*

Proof Suppose to the contrary that $A \cap B$ is empty. Then either $r_i(A)$ and $r_i(B)$ are disjoint, in which case $c_i(A) \cap c_i(B)$ is empty, a contradiction, or $r_i(A) \supset r_i(B)$, in which case B is a cavity in object O, violating our genus-zero assumption. □

This lemma justifies the following definition:

$$c_i(A, B) = \begin{cases} A \cap B, & \text{if } A \cap B \neq \emptyset, \text{ and at least one of } A \text{ and } B \text{ is a dent,} \\ c_i(A) \cap c_i(B) & \text{otherwise.} \end{cases}$$

This definition is intended to identify gridpoints on either A or B from which rays are issued by the ray-pair generation algorithm (Sect. 6.1.1). The reason for treating intersecting dents and protrusions differently is a subtle one, and is captured by Fig. 16b: B is a dent behind Y_i and B' is a protrusion in front of Y_i; $c_i(B')$ is the piece of the back face of B' enclosed by B; u is a highest gridpoint in $B \cap B'$, while w is a highest gridpoint in $c_i(B) \cap c_i(B')$; u is a potential ray basepoint, while w is not. The above definition eliminates points such as w from the set $c_i(A, B)$.

Our connectivity proof for G_b proceeds as follows. Let $P_i^1, P_i^2, \ldots$ denote the connected components of P_i, with $P_i = P_i^1 \cup P_i^2 \cup \cdots$. The bands incident to each of these are connected by rays (as discussed in Sect. 6.1.2) that lie in planes other than Y_i (see Fig. 14 for an example). We first argue that, to prove that G_b is ray-connected, it suffices to prove that each P_i^m is ray-connected. Remove from O all the slabs $S_1, S_2, \ldots$ incident to Y_0. Establish that the bands in the resulting object O' are ray-connected, via induction. The inductive hypothesis implies that the bands in each connected component of O' are ray-connected. Now put back the slabs. Each S_m corresponds to a component P_i^m. We will prove that all bands incident to P_i^m are ray-connected to one another. This along with the fact that O itself is connected implies that all bands are ray-connected. Henceforth we concentrate on one such connected component P_i^m, call it $Q \subset Y_i$ for succinctness. Let χ be the collection of all bands that intersect Q. Then $\bigcup_{A \in \chi} c_i(A) = Q$. The idea of the connectedness

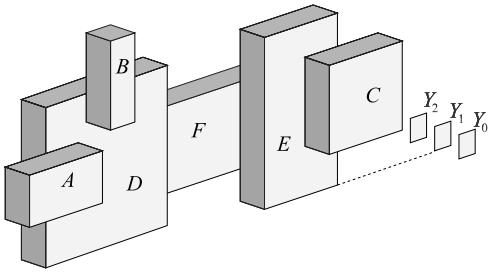


Fig. 14 P_1 contains two connected components, one incident to A, B, D, and one incident to C, E; pairs of bands incident to different components are connected by rays that lie in Y_2

proof is that the bands get connected in upward chains, and ultimately to each other through "common ancestor" higher bands. We choose to prove it by contradiction, arguing that a highest disconnected component cannot exist.

Lemma 6 *All bands in χ are ray-connected. Furthermore, if one arbitrary ray in each ray-pair is discarded, χ remains ray-connected.*

Proof For the purpose of contradiction, assume that not all bands in χ are ray-connected. Let $\chi_1, \chi_2, \ldots$ be the maximal subsets of χ that are ray-connected. Let $Q_j = \bigcup_{A \in \chi_j} c_i(A)$. Then $Q = \bigcup_j Q_j$. Since Q is connected, the subsets Q_j are not disjoint, in that for every Q_j there is an Q_k such that $Q_j \cap Q_k$ is nonempty. This along with Lemma 5 implies that

$$Q_{jk} = \bigcup_{A \in \chi_j} \bigcup_{B \in \chi_k} c_i(A, B)$$

is also nonempty. Let j and k be such that Q_{jk} contains a *highest* x-gridedge (gridpoint, if Q_{jk} contains only isolated points) among all Q_{jk}. Let u be the leftmost highest gridpoint in Q_{jk}. Let $A \in \chi_j$ and $B \in \chi_k$ be such that $u \in c_i(A, B)$.

We have thus identified two bands A and B, ray-disconnected because they lie in different components of Q, which contribute this highest gridpoint u in the "highest" intersection Q_{jk}. We now examine in turn the four protrusion/dent possibilities for these two bands.

Case 1. A and B are both protrusions on opposite sides of Y_i. Assume without loss of generality that A is behind Y_i, B is in front of Y_i, and u is on B (as depicted in Fig. 15). We discuss two subcases:

a. u is on a top edge of A or B; choose B without loss of generality (Figs. 15a, b). Then our ray-pair algorithm generates a ray-pair (r, r'), with r incident to u and r' incident to the gridpoint u' clockwise from u. Consider r (the analysis is similar for r'). If r hits A, then in fact A and B are ray-connected, contradicting the fact that A and B belong to different ray-connected components of χ. So let us assume

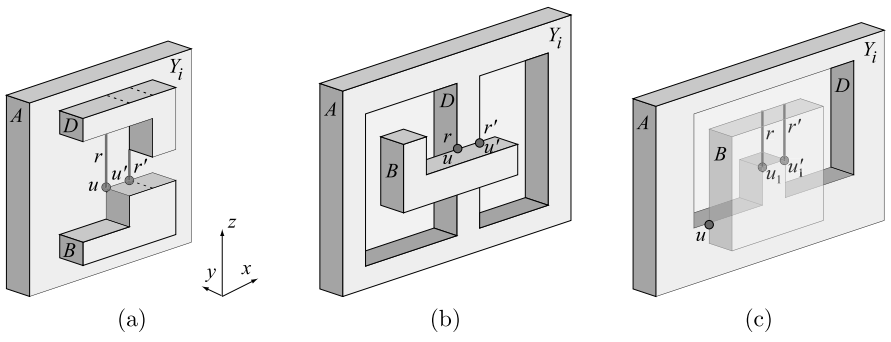


Fig. 15 Case 1: A and B are both protrusions on opposite sides of Y_i **a** D is a protrusion **b** D is a dent with a vertical side incident to u **c** D is a dent with a bottom edge incident to u

that r hits another band $D \in \chi_\ell$. Figure 15a, b illustrates the situation when D is a protrusion (dent). If $\ell = j$, then D and A are ray-connected in χ_j, and since B and D are ray-connected, it follows that B and A are ray-connected, a contradiction. On the other hand, if $\ell \neq j$, then $c_i(A, D)$ (and implicitly $Q_{j\ell}$) has a gridpoint higher than u, contradicting our choice of j, k and u.

b. u is not on a top edge of A or B, and so must be on a vertical (left, right) edge of A or B; again we choose B without loss of generality (Fig. 15c). Then u must be at the intersection between a dent D in protrusion A, and B. Because $A \cap D$ is empty, we fall into the second case of the definition of $c_i(A, D)$, which is therefore $c_i(A) \cap c_i(D)$. In this case, the same arguments as in Case a show that D and A are ray-connected, meaning that $D \in \chi_j$. Let u_1 be the leftmost among the highest gridpoints of $D \cap c_i(B)$. Then our ray-pair algorithm generates a ray-pair (r, r') from u_1 and its right neighbor u'_1. Consider r (the analysis is similar for r'). If r hits B, then B is ray-connected to D, which is ray-connected to A, a contradiction. If r hits a band E other than D, then it must be that $E \in \chi_k$, the same component containing B. Otherwise B and E would yield an intersection point higher than u, contradicting our choice of A and B. This means that B is ray-connected to E, which is ray-connected to D, which is ray-connected to A, a contradiction.

Case 2. A is a protrusion and B is a dent, both on a same side of Y_i. The case when A and B are both in front of Y_i (illustrated in Fig. 16a) is identical to Case 1 above, once one conceptually pops out B into a protrusion. We now discuss the case when A and B are both behind Y_i.

Assume first that $c_i(A, B)$ contains no top edges of B, as depicted in Fig. 16b. Let B' be a protrusion in front of Y_i covering the top of B. Then $c_i(A, B')$ and $c_i(B', B)$ each contains a gridpoint higher than u (see point $w' \in c_i(A, B')$ and $w \in c_i(B, B')$ in Fig. 16). The following two contradictory observations settle this case:

a. It must be that $B' \notin \chi_k$; otherwise Q_{jk} would contain a gridpoint in $c_i(A, B')$ higher than u.
b. If $B' \in \chi_\ell$, then it must be that $\ell = k$; otherwise $Q_{\ell k}$ would contain a gridpoint in $c_i(B', B)$ higher than u.

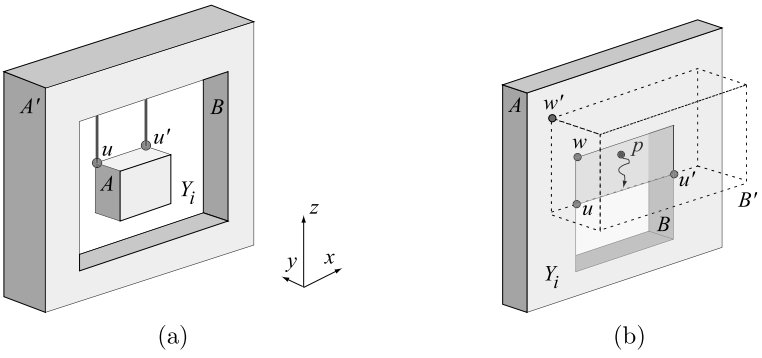


Fig. 16 Case 2: A is a protrusion and B is a dent **a** in front of Y_i **b** behind Y_i

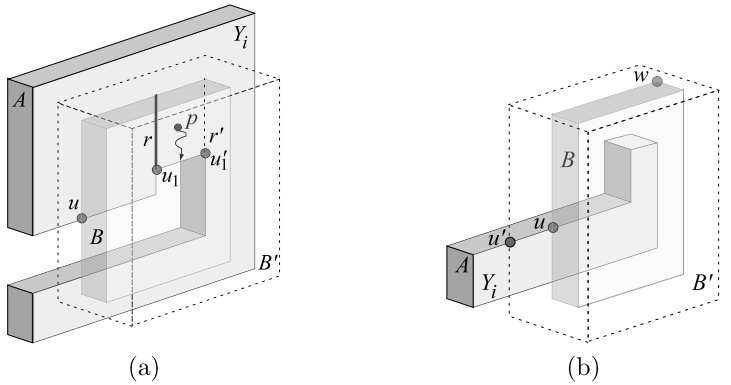


Fig. 17 Case 3: A is a protrusion behind Y_i; B is a dent in B', both in front of Y_i

If $c_i(A, B)$ contains at least one top gridedge of B, then arguments similar to the ones used for the case illustrated in Fig. 15a (conceptually popping B to become a protrusion) settle this case as well.

Case 3. A is a protrusion and B is a dent on opposite sides of Y_i (see Fig. 17). Let B' be the protrusion in front of Y_i enclosing B. We discuss three subcases:

a. $c_i(A)$ contains a top edge of B (see Fig. 17a). This means that $c_i(A) \cap r_i(B)$ is nonempty, and the ray-pair algorithm shoots a ray-pair (r, r') upward from the endpoints of a highest gridedge $\{u_1, u_1'\}$ of $A \cap r_i(B)$. Consider ray r (the analysis is similar for r'). If r hits B, then A and B are in fact ray-connected, a contradiction. If r hits a band D other than B, then arguments similar to the ones for the case illustrated in Fig. 15a (Case 1) lead to a contradiction.
b. $c_i(A)$ contains a bottom edge of B. This case is symmetrical to the one above in that a ray upward from a gridpoint of $B \cap r_i(A)$ hits A, thus ray-connecting A and B.
c. $c_i(A)$ contains neither a top nor a bottom edge of B (see Fig. 17b). Arguments similar to the ones used in Case 1 (protrusions on opposite sides of Y_i) show that A and B' are ray-connected. That B and B' are ray-connected follows immediately from the fact that $c_i(B, B')$ has a gridpoint higher than u (w in Fig. 17b). These together imply that A and B are ray-connected, a contradiction.

Case 4. A and B are both dents: A is a dent behind Y_i enclosed within protrusion A', and B is a dent in front of Y_i enclosed within protrusion B' (see Fig. 18). The genus-zero assumption implies that $r_i(A) \cap r_i(B)$ is a polygonal region of positive area. Since $u \in c_i(A) \cap c_i(B)$, we have that $u \in r_i(A) \cap r_i(B)$. Let β be the boundary segment of $r_i(A) \cap r_i(B)$ incident to u. We discuss two subcases:

a. $\beta \subset P_i^-$, meaning that $\beta \subset A$ (see Fig. 18a). An analysis similar to the one for the case illustrated in Fig. 17a (Case 3) shows that A and B are ray-connected, a contradiction.

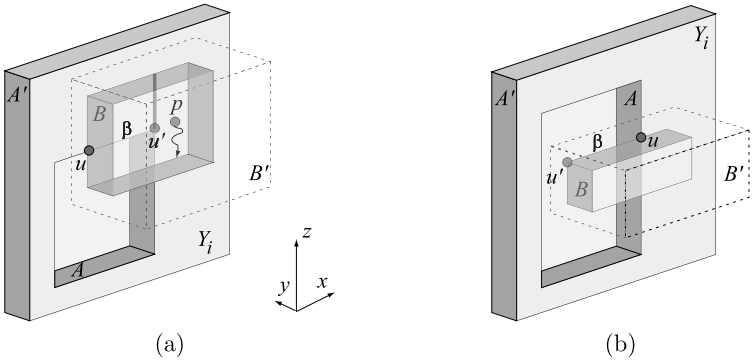


Fig. 18 Case 4: A is a dent behind Y_i, enclosed within protrusion A'. B is a dent in front of Y_i, enclosed within protrusion B'

b. $\beta \subset P_i^+$, meaning that $\beta \subset B$ (see Fig. 18b). We show that A and A' are ray-connected, B and B' are ray-connected, and A' and B' are ray-connected. This implies that A and B are ray-connected, a contradiction. First note that the ray-pair algorithm shoots a ray-pair (r, r') upward from a highest gridedge on β. An analysis similar to the one for the case illustrated in Fig. 15a (conceptually popping B to become a protrusion) shows that r and r' must hit B', thus ray-connecting B and B'. That A and A' are ray-connected follows immediately from the fact that $c_i(A, A')$ has a gridpoint higher than u, and similarly for A' and B'.

Having exhausted all possible cases, the connectivity claim of the lemma is established. Because the proof for each of these cases goes through by considering either the first or second ray of a ray-pair, retaining either ray suffices to preserve connectivity. Thus the second claim of the lemma is established as well. □

References

1. Biedl, T., Demaine, E., Demaine, M., Lubiw, A., O'Rourke, J., Overmars, M., Robbins, S., Whitesides, S.: Unfolding some classes of orthogonal polyhedra. In: Proc. 10th Canad. Conf. Comput. Geom., pp. 70–71, 1998
2. Damian, M., Flatland, R., O'Rourke, J.: Grid vertex-unfolding orthogonal polyhedra. In: 23rd Symp. Theoretical Aspects Comput. Sci., 2006. Lecture Notes Comput. Sci., vol. 3884, pp. 264–276. Springer, Berlin (2006)
3. Demaine, E.D., O'Rourke, J.: A survey of folding and unfolding in computational geometry. In: Goodman, J.E., Pach, J., Welzl, E. (eds.) Combinatorial and Computational Geometry. Cambridge University Press, Cambridge (2005)
4. Demaine, E.D., O'Rourke, J.: Open problems from CCCG 2004. In: Proc. 17th Canad. Conf. on Comput. Geom., pp. 303–306, 2005
5. Demaine, E.D., O'Rourke, J.: Geometric Folding Algorithms: Linkages, Origami, Polyhedra. Cambridge University Press, Cambridge (2007). http://www.gfalop.org
6. Demaine, E.D., Eppstein, D., Erickson, J., Hart, G.W., O'Rourke, J.: Vertex-unfoldings of simplicial manifolds. In: Bezdek, A. (ed.) Discrete Geometry, pp. 215–228. Dekker, New York (2003)
7. Demaine, E.D., Iacono, J., Langerman, S.: Grid vertex-unfolding of orthostacks. In: Japan Conf. Discrete Comput. Geom. 2004. Lecture Notes Comput. Sci., vol. 3742, pp. 76–82. Springer, Berlin (2005). Int. J. Comput. Geom. Appl. (to appear)

8. O'Rourke, J.: Folding and unfolding in computational geometry. In: Discrete Comput. Geom., Japan Conf. Discrete Comput. Geom., 1998. Lecture Notes Comput. Sci., vol. 1763, pp. 258–266. Springer, Berlin (2000).
9. Paterson, M.S., Yao, F.F.: Optimal binary space partitions for orthogonal objects. J. Algorithms **13**, 99–113 (1992)
10. Schwartz, E.L., Shaw, A., Wolfson, E.: A numerical solution to the generalized map-maker's problem: flattening nonconvex polyhedral surfaces. IEEE Trans. Pattern Anal. Mach. Intell. **11**(9), 1005–1008 (1989)
11. Tarini, M., Hormann, K., Cignoni, P., Montani, C.: Polycube-maps. ACM Trans. Graph. **23**(3), 853–860 (2004)
12. Wang, C.-H.: Manufacturability-driven decomposition of sheet metal products. PhD thesis, Carnegie Mellon University, The Robotics Institute (1997)

Empty Convex Hexagons in Planar Point Sets

Tobias Gerken

Abstract Erdős asked whether every sufficiently large set of points in general position in the plane contains six points that form a convex hexagon without any points from the set in its interior. Such a configuration is called an empty convex hexagon. In this paper, we answer the question in the affirmative. We show that every set that contains the vertex set of a convex 9-gon also contains an empty convex hexagon.

Keywords Erdős-Szekeres problem · Ramsey theory · Convex polygons and polyhedra · Empty hexagon problem

1 Introduction

In 1935, Erdős and Szekeres [5] proved that for each positive integer n there exists a smallest positive integer $g(n)$ such that every planar set of at least $g(n)$ points in general position contains n points that are the vertices of a convex n-gon. Here, general position means that no three points are collinear.

The best known bounds for $g(n)$ are $2^{n-2} + 1 \leq g(n) \leq \binom{2n-5}{n-2} + 1$. The lower bound is due to Erdős and Szekeres [6] and the upper bound was established recently by Tóth and Valtr [11]. The lower bound is sharp for $n \leq 5$ and is conjectured to be sharp for all n by Erdős and Szekeres [5, 6]. For a survey of results related to the Erdős–Szekeres theorem, see [1, 2, 9, 11].

In 1978, Erdős [3, 4] posed the problem of determining the smallest positive integer $h(n)$, if it exists, such that any set X of at least $h(n)$ points in general position in the plane contains n points that are the vertices of an *empty* convex polygon; that is, a convex n-gon whose interior does not contain any point of X. Trivially, $h(n) = n$

T. Gerken (✉)
Zentrum Mathematik, Technische Universität München, Boltzmannstr. 3,
85747 Garching, Germany
e-mail: gerken@ma.tum.de

for $n \leq 3$. It is easy to see that $h(4) = 5$. In 1978, Harborth [7] proved that $h(5) = 10$, while Horton [8] showed in 1983 that for all $n \geq 7$, $h(n)$ does not exist. The problem of determining the existence of $h(6)$ has since been open. Based on computer experiments, Overmars [10] showed that $h(6) \geq 30$ (if it exists). In this paper, we prove the following theorem which implies that *every sufficiently large planar point set in general position contains the vertex set of an empty convex hexagon*.

Theorem 1 $h(6) \leq g(9)$.

The above bounds yield $129 \leq g(9) \leq 1717$. Note that there exist sets of points without empty convex hexagons that have eight points on the convex hull [10].

2 Overview of the Proof

Proof In the following, let X be a finite planar set of points in general position that contains the vertex set of a convex 9-gon. By the Erdős–Szekeres theorem [5] this is always the case if $|X| \geq g(9)$. Let $H \subseteq X$ be the vertex set of a convex 9-gon in X with the *minimum* $|X \cap \text{conv}(H)|$, where $\text{conv}(M)$ denotes the convex hull of the set M. Let $I := \text{conv}(H) \cap (X \setminus H)$ be the set of points of X inside the convex hull of H. Note that $\text{conv}(I)$ is a convex polygon and denote by ∂I its vertex set. If $|I| > 2$, let $J := \text{conv}(I) \cap (X \setminus \partial I)$ be the set of points of X inside the convex hull of ∂I. Note that $\text{conv}(J)$ is again a convex polygon and denote by ∂J its vertex set; see Fig. 1. Let $i := |\partial I|$ and $j := |\partial J|$. Note that $0 \leq i, j \leq 8$ as otherwise there would be a 9-gon H' with smaller $|X \cap \text{conv}(H')|$. This leaves the 57 cases $0 \leq i \leq 2$ and $(i, j) \in \{3, \ldots, 8\} \times \{0, \ldots, 8\}$. We argue that in each case either an empty convex u-gon can be found ($u \geq 6$) or a convex 9-gon H' with smaller $|X \cap \text{conv}(H')|$ is present which contradicts the minimality condition imposed on H. (More precisely, the vertex set of an empty convex u-gon can be found. In the following, we do not make this distinction when the meaning is clear from the context.)

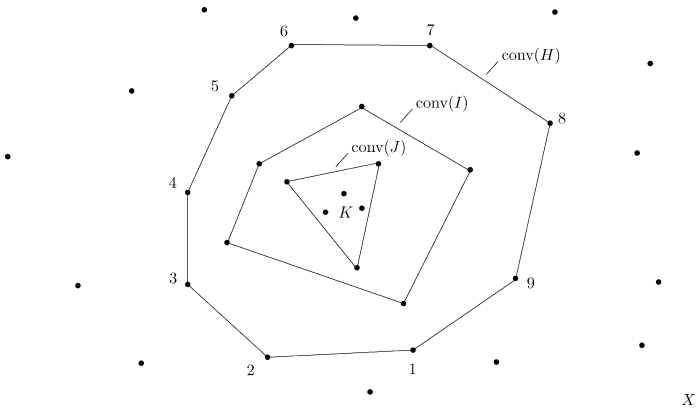


Fig. 1 Basic notation

Table 1 Overview of the proof structure: for example, the proof for the case $(8, 5)$ is given in Sects. 6 (special case) and 10 (general case)

i/j	0	1	2	3	4	5	6	7	8
0	3	–	–	–	–	–	–	–	–
1	3	–	–	–	–	–	–	–	–
2	3	–	–	–	–	–	–	–	–
3	4	4	4	4	4	4	4	4	4
4	5	5	5	5	5	5	5	5	5
5	6	7.1	8	8	8	8	8	8	8
6	3	7.2	7.3	4/5	9	9	9	9	9
7	3	7.3	3	4/5	5	6/10	3/10	3/10	3/10
8	3	3	3	4	5	6/10	3/10	3/10	3/10

2.1 Notation

We use (i, j) to denote a specific case, where i and j are defined as above. Sometimes we use the notation (i, j, k), where k refers to the number of points of X inside the convex hull of J; that is, the cardinality of $K := \text{conv}(J) \cap (X \setminus \partial J)$. The notation $\geq x$ indicates that x is a lower bound for i, j or k. Refer to Table 1 for locating the proof of a specific case.

2.2 Definitions

Given three points in general position, P, Q, R, define the halfplane $H_{PQ}(R)$ as the open halfplane defined by the line $\overline{PQ}$ that contains R. A *convex chain* is a set of consecutive vertices of a convex polygon. Given a convex chain of three points, $\overline{ABC}$, the *3-sector* specified by this chain is defined as

$$(ABC) := [H_{AB}(C) \cap H_{BC}(A)] \setminus \text{conv}(\{A, B, C\}).$$

Note that three points in general position, S, T, U, lying in (ABC) can be used to construct a convex hexagon if $A, B, C \in (STU)$; see Fig. 2a.

Given a convex chain of four points, $\overline{ABCD}$, the corresponding *4-sector* is defined as

$$(ABCD) := [(ABC) \cap (BCD)] \setminus \text{conv}(\{A, B, C, D\}).$$

Note that two points, S, T, lying in $(ABCD)$ can be used to construct a convex hexagon if the line $\overline{ST}$ does not intersect $\text{conv}(\{A, B, C, D\})$; see Fig. 2b. This means that by construction, given an edge PQ of $\text{conv}(I)$ (respectively $\text{conv}(J)$), at most three vertices of $\text{conv}(H)$ (respectively $\text{conv}(I)$) can lie in an open halfplane that is defined by the line $\overline{PQ}$ and does not include any other point of I (respectively J) if no empty convex hexagon is to occur. In the following figures, we use the notation (PQ) to hint to this fact; see Fig. 2c.

Finally, given a convex chain of five points, $\overline{ABCDE}$, the corresponding *5-sector* is defined as

$$(ABCDE) := [(ABCD) \cap (BCDE)] \setminus \text{conv}(\{A, B, C, D, E\}).$$

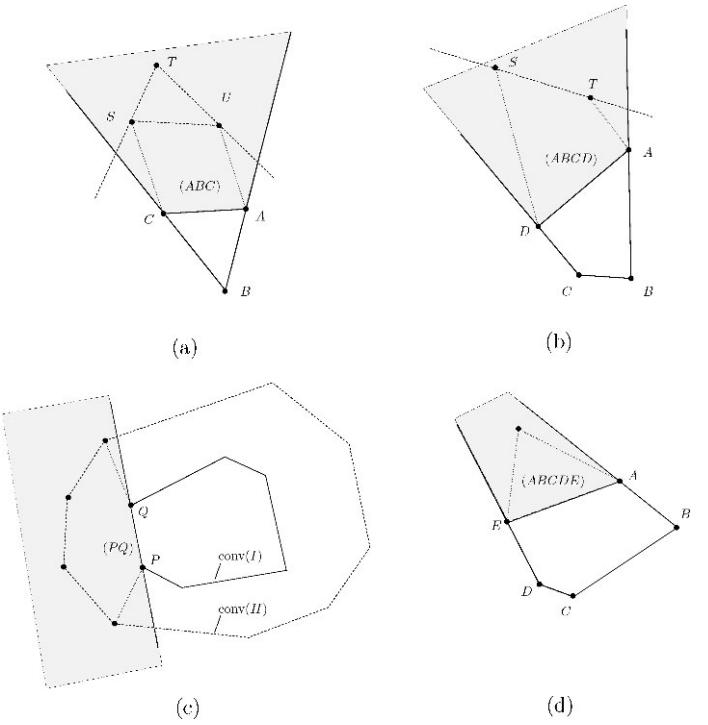


Fig. 2 Definition: sector

Note that a single point lying in $(ABCDE)$ can be used to construct a convex hexagon; see Fig. 2d.

3 Elementary Cases

Note that the cases **(0, 0)**, **(≥ 6, 0)** and **(≥ 3, ≥ 6, 0)** are trivial as an empty convex hexagon is present. The cases **(1, 0)** and **(8, 1)** can be dealt with by considering a line through the single interior point and one of the vertices of the convex 9- respectively 8-gon. Due to the general position, on one side of this line a convex chain of four vertices must be present which together with the two preselected points can be used to construct an empty convex hexagon. A similar argument settles the cases **(2, 0)**, **(8, 2)** and **(7, 2)**.

4 The Cases (3, ≥ 0) and (≥ 6, 3)

We approach the cases $(3, \geq 0)$ and $(\geq 6, 3)$ in two batches:

4.1 The Cases $(3, \geq 0)$ and $(8, 3)$

Follow the notation as indicated in Fig. 3. The variables stand for the number of vertices of the convex 9- respectively 8-gon in each sector. Assume that no empty

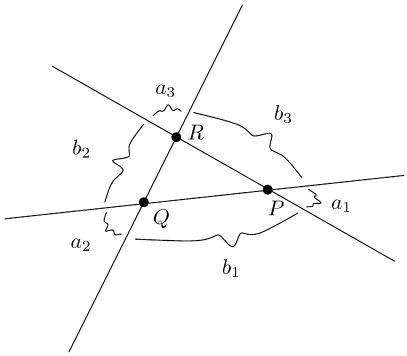


Fig. 3 Notation for the cases $(3, \geq 0)$ and $(\geq 6, 3)$

convex hexagon is present. Note that

$$1 \leq a_1 + b_1 + a_2 \leq 3, \tag{4.1}$$

$$1 \leq a_2 + b_2 + a_3 \leq 3, \tag{4.2}$$

$$1 \leq a_3 + b_3 + a_1 \leq 3 \tag{4.3}$$

by construction and as otherwise a convex chain of four vertices together with two vertices of the triangle could be used to form an empty convex hexagon. Also,

$$0 \leq b_i \leq 2 \quad (1 \leq i \leq 3) \tag{4.4}$$

as otherwise a convex chain of three points together with two vertices of the triangle and either the third vertex of the triangle or (if existent) one of its interior points can be used to form an empty convex hexagon. Summing up the upper bounds in (4.1–4.4) yields

$$2 \cdot \sum_{i=1}^{3} (a_i + b_i) \leq 15. \tag{4.5}$$

Therefore, at most seven vertices can be placed around the triangle and in the two cases at hand an empty convex hexagon is present.

4.2 The Cases $(6, 3)$ and $(7, 3)$

The cases **(6, 3, 0)** and **(7, 3, 0)** can be settled by a careful investigation of the $(a_1, b_1, a_2, b_2, a_3, b_3)$-tuples that are feasible for the set of constraints (4.1–4.4). Note that tuples (a_i, b_i, a_{i+1}) with $a_i = a_{i+1} = 0$ are not feasible, as a convex 9-gon H' with smaller $|X \cap \mathrm{conv}(H')|$ could be constructed; see Fig. 4. In Fig. 4a, replace the vertices of the 9-gon lying in the union of sectors (AQB) and (BRC) (at least one by construction and at most four in total if no empty convex hexagon is present) by points from the convex chain $\overline{AQRC}$ of length four. In Fig. 4b, accordingly replace the at most four vertices of the 9-gon lying in the union of sectors (AQB_1), (B_1QPRB_2) and (B_2RC).

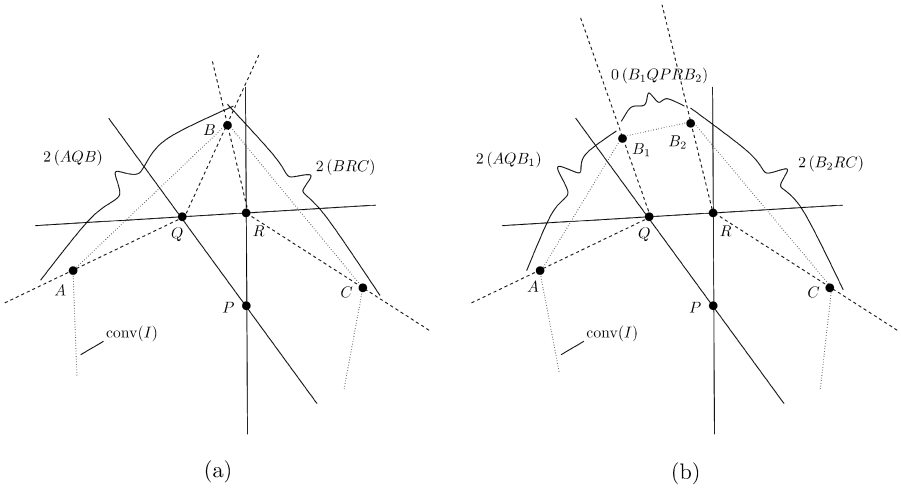


Fig. 4 $(x, 3)$: degenerate cases with $(a_i, b_i, a_{i+1}) = (0, 1, 0)$ and $(a_i, b_i, a_{i+1}) = (0, 2, 0)$ respectively. Numbers indicate the number of vertices of the 9-gon that can lie in each sector without forming an empty convex hexagon

Table 2 Cases $(6, 3, 0)$ and $(7, 3, 0)$: Combinatorial subcases under the assumptions (i) $a_1 \geq a_2 \geq a_3$ and (ii) $b_2 \geq b_3$ for fixed (a_1, b_1, a_2)

a_1	a_2	a_3	b_1	b_2	b_3	Solution
3	0	0	*	*	*	Infeasible
2	1	1	0	(≤ 1)	0	$\sum_{i=1}^{3}(a_i + b_i) \leq 5$
2	1	0	0	(≤ 2)	(≤ 1)	$(2, 0, 1, 2, 0, 1)$
2	0	0	*	*	*	Infeasible
1	1	1	(≤ 1)	(≤ 1)	(≤ 1)	$(1, 1, 1, 1, 1, 1)$
1	1	0	1	2	2	$(1, 1, 1, 2, 0, 2)$
1	1	0	1	2	1	$(1, 1, 1, 2, 0, 1)$
1	1	0	1	2	0	$\sum_{i=1}^{3}(a_i + b_i) = 5$
1	1	0	1	(≤ 1)	(≤ 1)	$\sum_{i=1}^{3}(a_i + b_i) \leq 5$
1	1	0	0	2	2	$(1, 0, 1, 2, 0, 2)$
1	1	0	0	(≤ 2)	(≤ 1)	$\sum_{i=1}^{3}(a_i + b_i) \leq 5$
1	0	0	*	*	*	Infeasible
0	0	0	*	*	*	Infeasible

*Marks an arbitrary entry

Note that constraint (4.1) implies $a_1 + a_2 \leq 3$

Now assume without loss of generality that (i) $a_1 \geq a_2 \geq a_3$ and (ii) $b_2 \geq b_3$ for fixed (a_1, b_1, a_2); see Fig. 3. Then the only solutions to the above set of constraints (modulo rotations and reflections) are $(2, 0, 1, 2, 0, 1)$, $(1, 1, 1, 1, 1, 1)$,

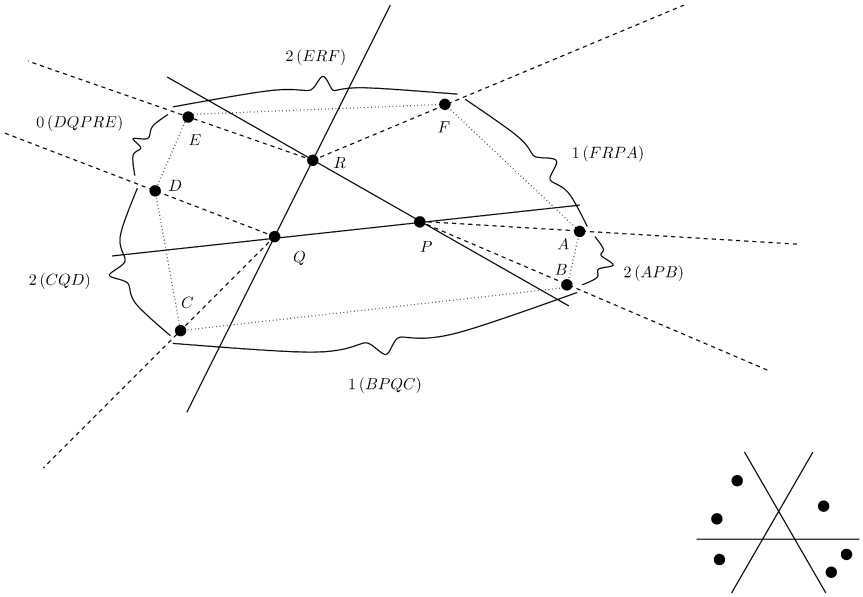


Fig. 5 The case (6, 3, 0) with (2, 0, 1, 2, 0, 1)

(1, 1, 1, 2, 0, 2), (1, 1, 1, 2, 0, 1) and (1, 0, 1, 2, 0, 2); see Table 2. These can be treated individually as follows:

- The subcase (2, 0, 1, 2, 0, 1) can be treated as indicated in Fig. 5. Here and in the following, numbers indicate the number of vertices of the outer polygon that can lie in each sector without forming an empty convex hexagon. As the union of sectors allows for at most eight points in convex position in the outmost layer, due to the presence of a convex 9-gon an empty convex hexagon must occur.
- Figure 6 indicates how to settle the subcase (1, 1, 1, 1, 1, 1), provided the vertex Q of triangle PQR lies inside the triangle BDF. In that case the quadrilateral $BQDC$ exists. Similarly we can treat the case that some other of the points P, Q lies inside the triangle BDF. If none of the points P, Q, R lies inside the triangle BDF, the empty convex hexagon $PBQDRF$ occurs.
- Figure 7 indicates how to settle the subcase (1, 1, 1, 2, 0, 2), provided that the point Q lies outside the triangle BCD. In that case the quadrilateral $CBQD$ exists. If Q lies inside the triangle BCD, the empty convex hexagon $BQDERP$ occurs.
- The subcase (1, 1, 1, 2, 0, 1) can be treated as indicated in Fig. 8.
- Figure 9 indicates how to settle the subcase (1, 0, 1, 2, 0, 2).

The proof for the cases **(6, 3, ≥ 1)** and **(7, 3, ≥ 1)** is given in the following Sect. 5.

5 The Cases (4, ≥ 0) and (≥ 7, 4)

The cases **(4, ≥ 0)** and **(≥ 7, 4)** can be dealt with simultaneously in three steps:

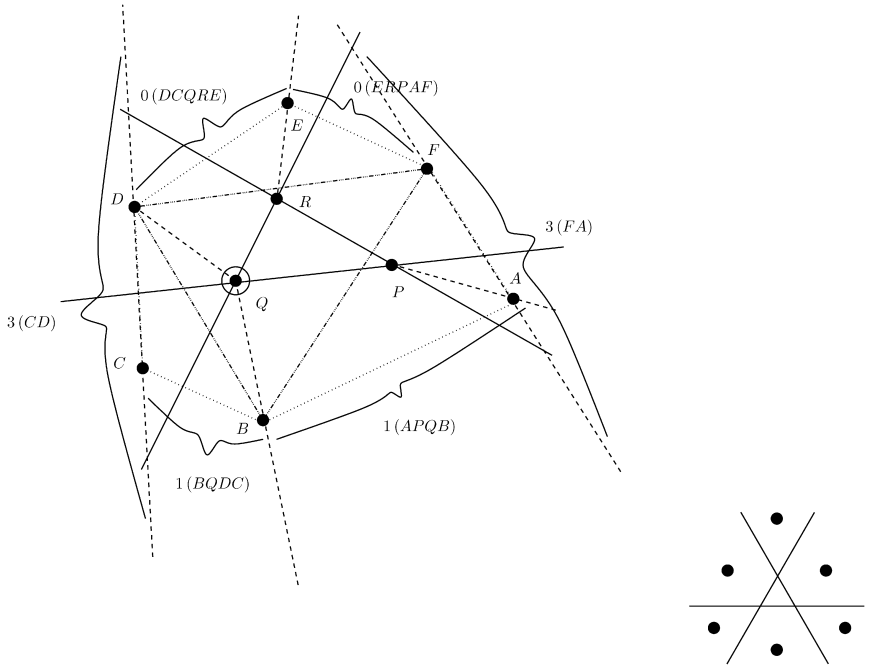


Fig. 6 The case (6, 3, 0) with (1, 1, 1, 1, 1, 1). It is assumed that $Q \in \triangle BDF$

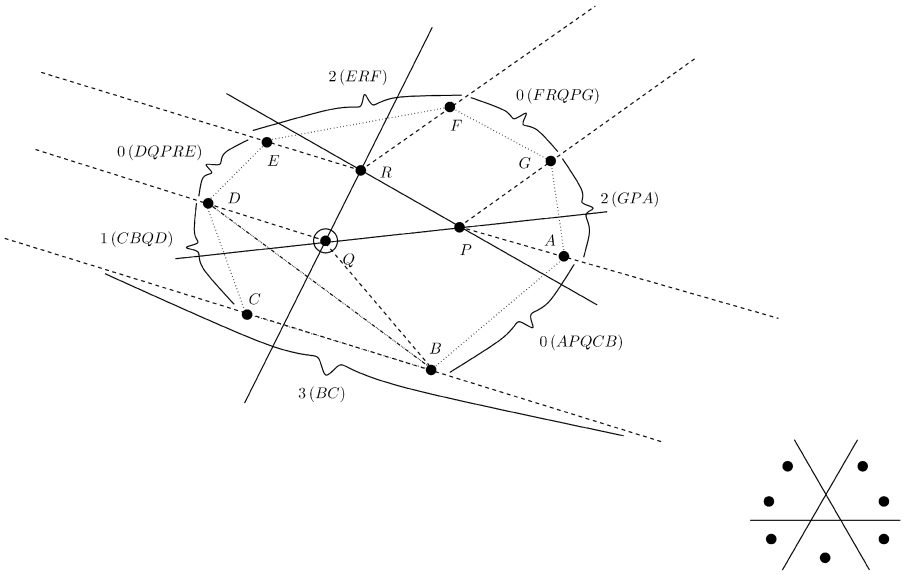


Fig. 7 The case (7, 3, 0) with (1, 1, 1, 2, 0, 2). It is assumed that $Q \notin \triangle BCD$

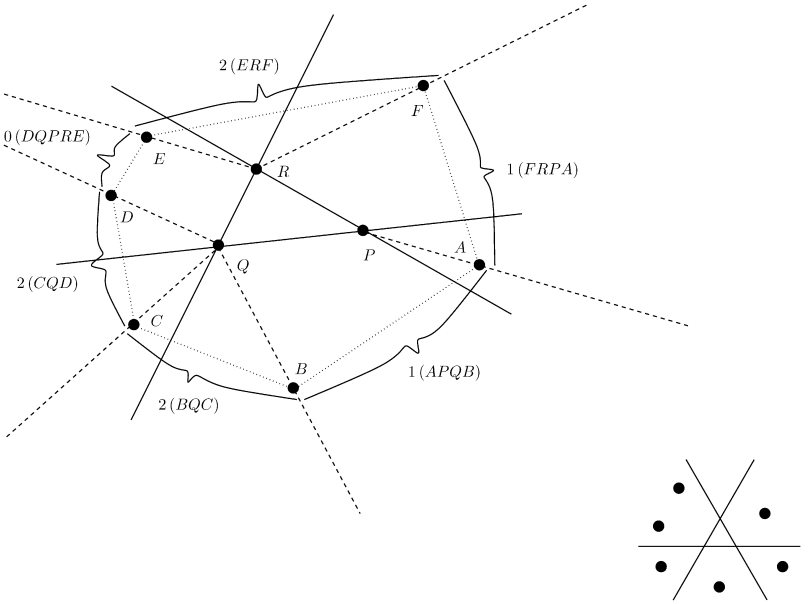


Fig. 8 The case (6, 3, 0) with (1, 1, 1, 2, 0, 1)

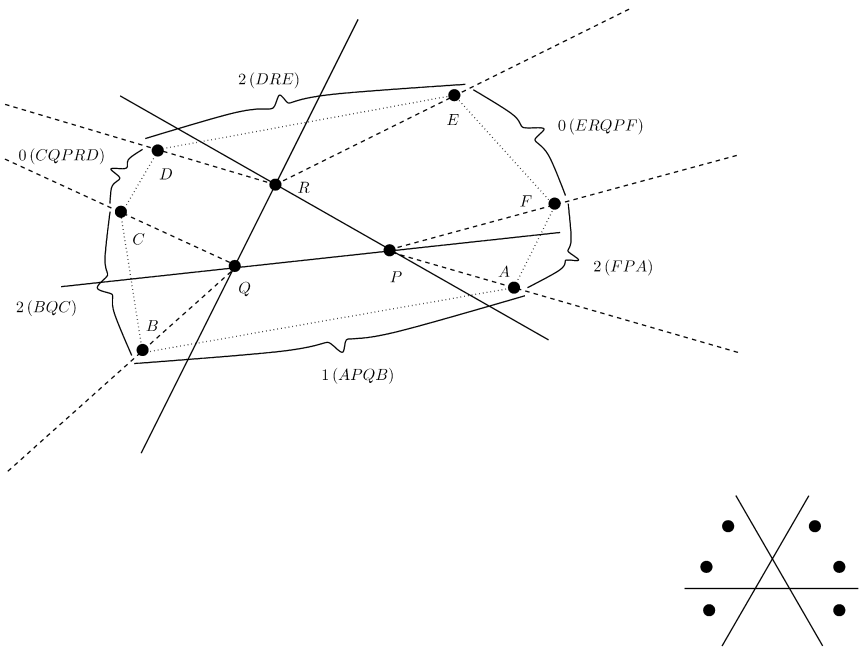


Fig. 9 The case (6, 3, 0) with (1, 0, 1, 2, 0, 2)

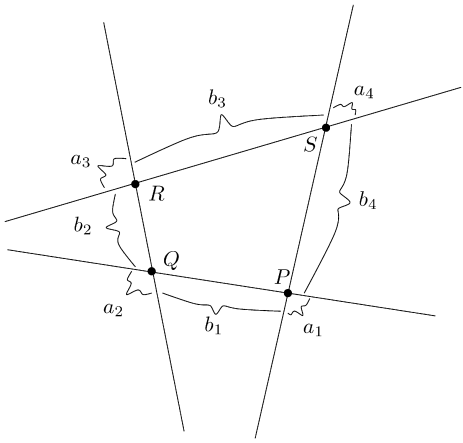


Fig. 10 Notation for the cases $(4, \geq 0)$ and $(\geq 7, 4)$

5.1 Step 1a

First, consider the cases **(4, 0)** and **(8, 4, 0)**. We use the same type of approach as in Sect. 4. Following the notation as indicated in Fig. 10, where variables again refer to the number of vertices of the 9- respectively 8-gon lying in each sector, we arrive at the set of inequalities

$$1 \leq a_1 + b_1 + a_2 \leq 3, \tag{5.1}$$

$$1 \leq a_3 + b_3 + a_4 \leq 3, \tag{5.2}$$

if no empty convex hexagon is to occur. (Vertices lying in more than one sector are assigned arbitrarily to one particular sector they lie in and therefore only counted once.) If no empty convex hexagon is to be present, the constraint

$$0 \leq b_2 + b_4 \leq 1 \tag{5.3}$$

must also hold. By summing up the upper bounds in (5.1–5.3), it follows that at most seven vertices can be placed around the 4-gon, a contradiction in these two cases.

5.2 Step 1b

We next consider the case **(7, 4, 0)** and evaluate the feasible solutions to the set of constraints (5.1–5.3). By symmetry, any feasible $(a_1, b_1, a_2, b_2, a_3, b_3, a_4, b_4)$-tuple must also satisfy the following set of inequalities:

$$1 \leq a_1 + b_4 + a_4 \leq 3, \tag{5.4}$$

$$1 \leq a_2 + b_2 + a_3 \leq 3, \tag{5.5}$$

$$0 \leq b_1 + b_3 \leq 1. \tag{5.6}$$

It follows directly from (5.3) and (5.6) that $\sum_{i=1}^{4} b_i \leq 2$. Furthermore, it follows from (5.1) and (5.2) (respectively (5.4) and (5.5)) that if $b_2 = b_4 = 0$ or $b_1 = b_3 = 0$,

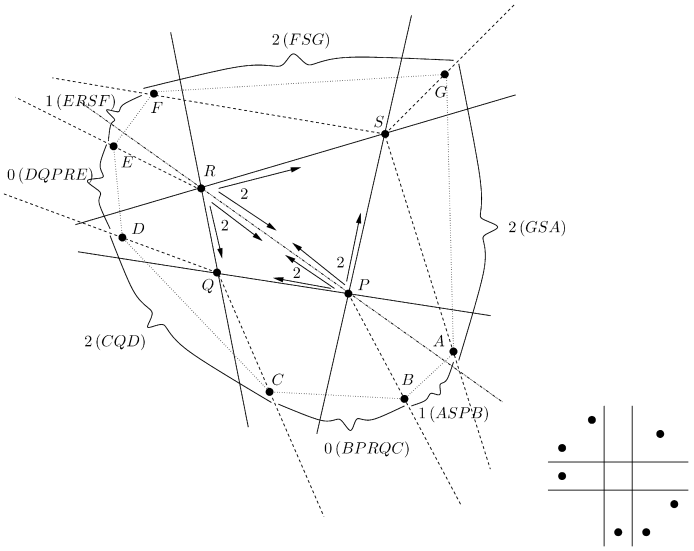


Fig. 11 The case (7, 4, 0) with (2, 1, 0, 1, 2, 0, 1, 0)

at most six vertices can be placed around the 4-gon. Therefore, $(b_1, b_2, b_3, b_4) = (1, 1, 0, 0)$ without loss of generality. By choosing $a_1 \in \{0, 1, 2\}$, it follows that only the following $(a_1, b_1, a_2, b_2, a_3, b_3, a_4, b_4)$-tuples are feasible: (2, 1, 0, 1, 2, 0, 1, 0), (1, 1, 1, 1, 1, 0, 2, 0) and (0, 1, 2, 1, 0, 0, 3, 0) (modulo rotations and reflections). These can be treated individually as follows:

- The subcase (2, 1, 0, 1, 2, 0, 1, 0) can be treated as indicated in Fig. 11. Note that at most two of the points D, E, F can lie in one of the sectors (QPR) and (RPS) without the occurrence of an empty convex hexagon. The same holds for A, B, C and the sectors (QRP) and (PRS). This is indicated by the arrows. Note that one 4- and one 5-sector arise.
- Figure 12 indicates how to settle the subcase (1, 1, 1, 1, 1, 0, 2, 0), provided that the vertex Q of the quadrilateral $PQRS$ lies outside the triangle BCD. In that case, the quadrilateral $BQDC$ exists. Note that if Q lies inside the triangle BCD, there exists an empty convex hexagon $BQDRSP$.
- The subcase (0, 1, 2, 1, 0, 0, 3, 0) can be treated as indicated in Fig. 13. Note that if B and C both lie in (PSQ) or both lie in (QSR), an empty convex hexagon occurs ($ABCQSP$ and $BCDRSQ$ respectively). Again, this is indicated by the arrows.

5.3 Step 2

Now we investigate the cases **(4, 1)** and **(8, 4, 1)**. Consider the sectors occurring when rays emanate from the single point in J (respectively K) through the vertices of the convex 4-gon. Each of the four sectors can only contain two vertices of the convex 9- respectively 8-gon as otherwise an empty convex hexagon could be constructed. Since $4 \cdot 2 < 9$, in the case of the 9-gon an empty convex hexagon must occur. The case of the 8-gon is settled with a similar sector argument on the next level as indicated in Fig. 14.

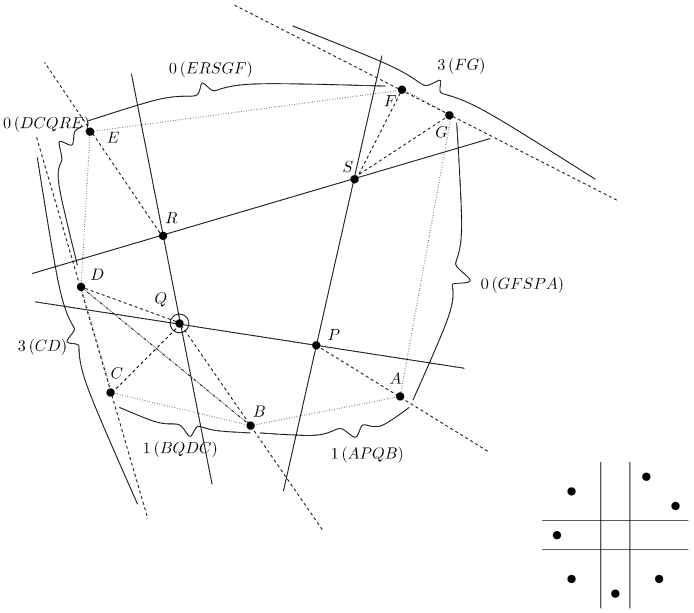


Fig. 12 The case $(7, 4, 0)$ with $(1, 1, 1, 1, 1, 0, 2, 0)$. It is assumed that $Q \notin \triangle BCD$

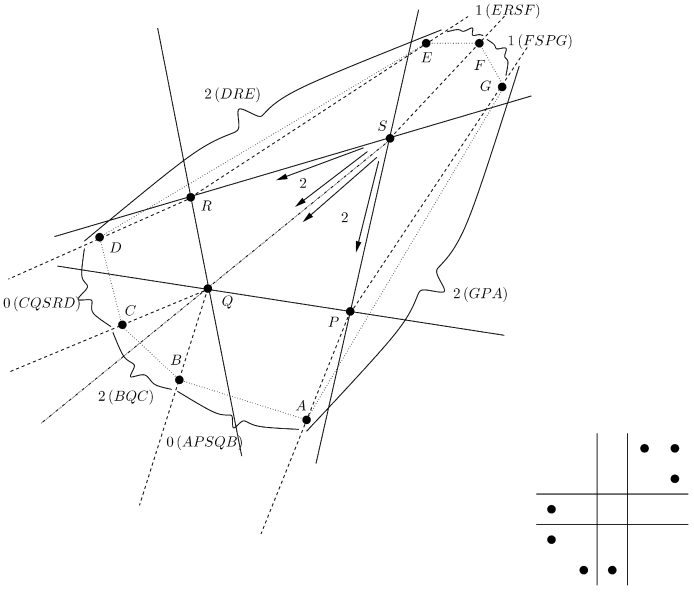


Fig. 13 The case $(7, 4, 0)$ with $(0, 1, 2, 1, 0, 0, 3, 0)$

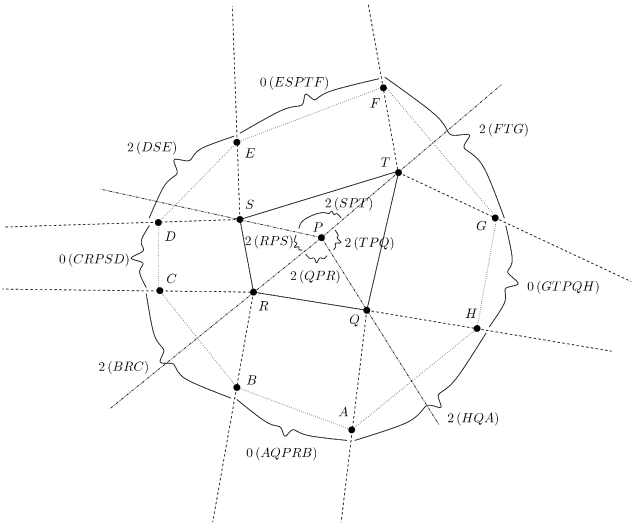


Fig. 14 The case (8, 4, 1)

5.4 Step 3

In dealing with the cases **(4, ≥ 2)**, fix a point $P \in J$, construct the sectors as in Step 2 and afterwards replace P with an appropriate point from J in each sector (if necessary). Now argue as in Step 2. Proceed accordingly in the cases **(8, 4, ≥ 2)** by choosing an arbitrary $P \in K$.

5.5 Remark

The approach of Sects. 5.3 and 5.4 also works straightforwardly in the cases **(6, 3, ≥ 1)** (as indicated in Fig. 15), **(7, 3, ≥ 1)** and **(7, 4, ≥ 1)** (as indicated in Fig. 16). Again, the idea is to fix a point $P \in K$ and to create sectors from rays emanating from P that pass through the vertices of the j-gon. Argue that each of these sectors can only contain at most two vertices of the i-gon without the occurrence of an empty convex hexagon. This remains true if other points of K should lie in some of the sectors. Now create another set of sectors such that their union covers the complete region outside of conv(I) as indicated in the figures. This approach is extended in Sect. 10 dealing with the cases $(\geq 7, \geq 5, \geq 1)$.

6 The Cases (5, 0) and (≥ 7, 5, 0)

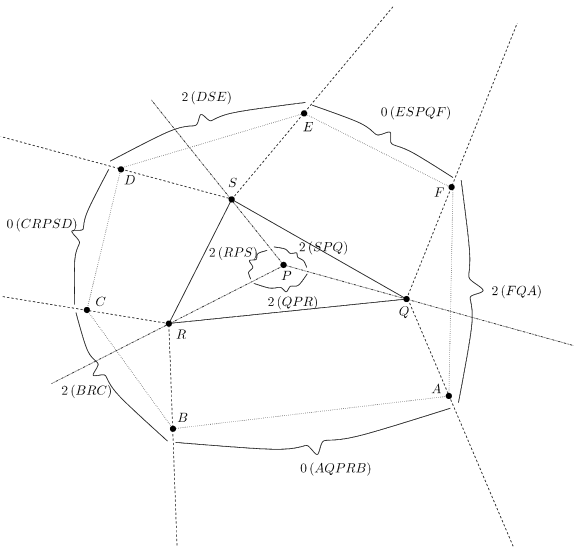
6.1 The cases (5, 0) and (8, 5, 0)

We use the same basic approach as in Sects. 4 and 5, extending the concept and notation of Figs. 3 and 10 in the natural way. We arrive at the set of inequalities

$$b_i = 0 \quad (1 \leq i \leq 5) \quad \text{and} \tag{6.1}$$

$$1 \leq a_i + a_{i+1} \leq 3 \quad (1 \leq i \leq 5, a_6 := a_1) \tag{6.2}$$

Fig. 15 The case (6, 3, 1)



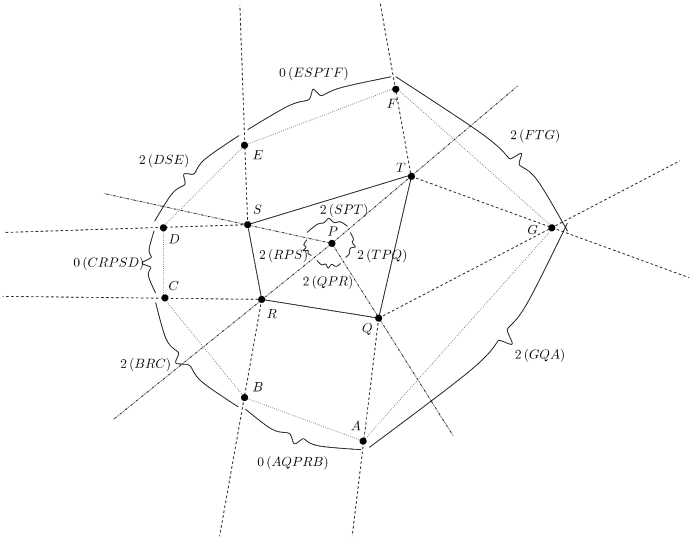


Fig. 16 The case (7, 4, 1)

if no empty convex hexagon is to be present (again counting vertices lying in more than one sector only once). This set of inequalities yields

$$2 \cdot \sum_{i=1}^{5} (a_i + b_i) \leq 15, \tag{6.3}$$

which implies the desired contradiction that an outer convex polygon with at most seven vertices can be present.

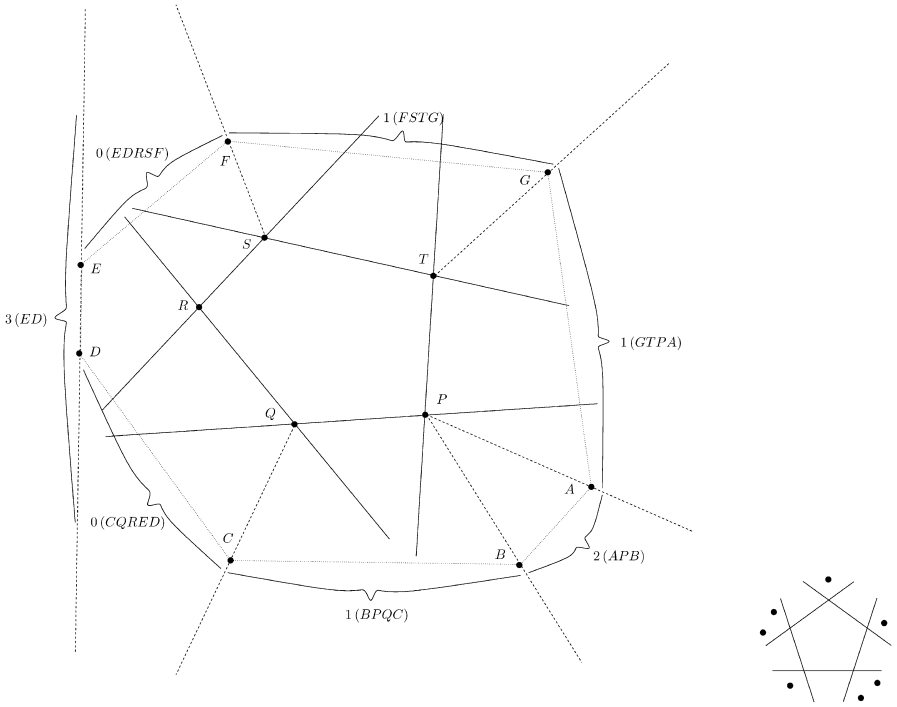


Fig. 17 The case (7, 5, 0)

6.2 The Case (7, 5, 0)

A closer investigation of the constraints (6.1–6.3) shows that in this case the only feasible $(a_1, b_1, a_2, b_2, a_3, b_3, a_4, b_4, a_5, b_5)$-tuple (modulo rotation) is (2, 0, 1, 0, 2, 0, 1, 0, 1, 0). This case can be settled as indicated in Fig. 17.

7 Individual Cases

7.1 The Case (5, 1)

This case can be dealt with as indicated in Fig. 18. Observe that P must lie in one of the triangles $\triangle ABD$, $\triangle BCE$, $\triangle CDA$, $\triangle DEB$ or $\triangle EAC$ (as these cover the convex 5-gon). Without loss of generality P is inside the triangle ABD (as in the figure). The line PD cuts the 5-gon into the two quadrilaterals $AEDP$ and $PDCB$ (and one triangle). It follows that $m_1 + m_2 \leq 1$ and $n_1 + n_2 \leq 1$ if no empty convex hexagon is to be present. (As in previous sections, variables refer to the number of vertices of the 9-gon lying in the corresponding sectors.) This leads to at most eight points that can be placed in convex position around the 5-gon without creating an empty convex hexagon.

Fig. 18 The case (5, 1)

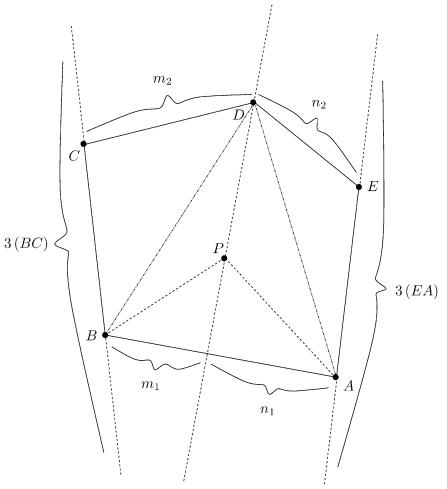
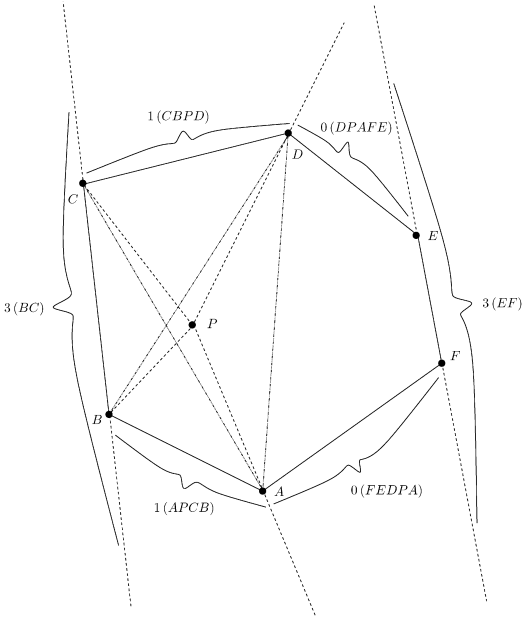


Fig. 19 The case (6, 1). See Sect. 7.2 for details



7.2 The Case (6, 1)

This case can be dealt with as indicated in Fig. 19. Note that P must lie in one of the 4-gons $ADEF$ or $ABCD$ (as in the figure). Note furthermore that if in the latter case, $P \in \triangle ABC$ or $P \in \triangle BCD$, an empty convex hexagon occurs ($APCDEF$ respectively $BPDEFA$). Therefore, assume that the convex 4-gons $APCB$ and $CBPD$ exist and argue as indicated in the figure.

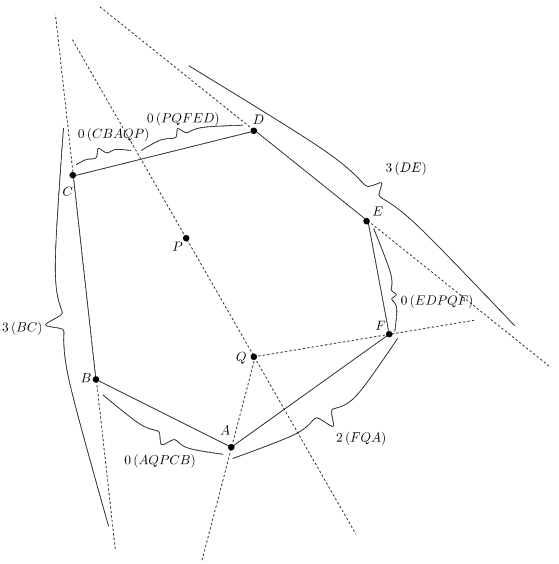


Fig. 20 The case (6, 2). See Sect. 7.3 for details

7.3 The Cases (6, 2) and (7, 1)

The case **(6, 2)** can be dealt with as indicated in Fig. 20. Note that if four vertices of the 6-gon lie on one side of the line $\overline{PQ}$, an empty convex hexagon can be constructed. The case **(7, 1)** is treated similarly. Here, one of the vertices of the convex 7-gon takes the role of P.

8 The Cases (5, ≥ 2)

8.1 A Key Observation

The following observation is needed in later sections.

Observation 1 Suppose that $j > 2$ and let $2 \leq t \leq \min\{i-1, j\}$. Consider a sequence of t consecutive vertices $V_1, V_2, \ldots, V_t$ of $\mathrm{conv}(J)$. Denote by T_n the set of vertices of the i-gon $\mathrm{conv}(I)$ lying in the halfplane that is defined by the line $\overline{V_n V_{n+1}}$ and that does not contain any other points of J. If $|\bigcup_{n=1}^{t-1} T_n| < t$, a 9-gon H' with smaller $|X \cap \mathrm{conv}(H')|$ can be constructed.

Proof We prove by induction over t. We use U_l ($l \in \mathbb{N}_0$) to denote vertices of $\mathrm{conv}(I)$. Note that $|T_n| > 0$ for all n by the definition of J.

Let $t = 2$. Assume that $T_1 =: \{U_1\}$; see Fig. 21. We claim that at most four vertices of the 9-gon can lie in the union of the 3-sectors $(U_0V_1U_1)$ and $(U_1V_2U_2)$, where U_0 and U_2 are the vertices of conv(I) preceding and succeeding U_1. (Note that $U_0 \neq U_2$ as we presume $t < i$.) The bound follows directly if no other point of J lies within the triangles $\triangle U_0V_1U_1$ respectively $\triangle U_1V_2U_2$. Otherwise replace V_1 (respectively V_2) by appropriate $V_1' \in J \cap \triangle U_0V_1U_1$ and $V_2' \in J \cap \triangle U_1V_2U_2$ to obtain new

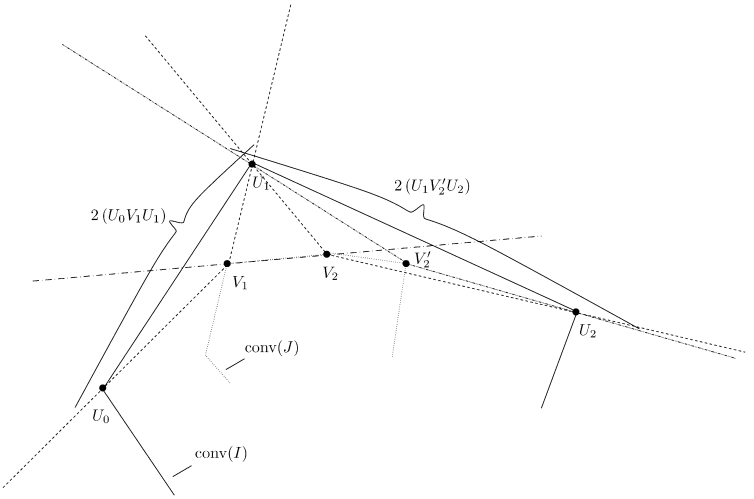


Fig. 21 Observation 1: $t = 2$

3-sectors $(U_0V_1'U_1)$ and $(U_1V_2'U_2)$ such that the corresponding triangles $\triangle U_0V_1'U_1$ and $\triangle U_1V_2'U_2$ do not contain any points of J. Note that these 3-sectors cover the region outside of $\mathrm{conv}(I)$ that was originally covered by $(U_0V_1U_1)$ and $(U_1V_2U_2)$. (In fact, they cover a larger region.) Each of them allows for at most two vertices of the 9-gon without the occurrence of an empty convex hexagon and the claim follows. Replacing these vertices by points from the convex chain $\overline{U_0V_1V_2U_1}$ of length four yields a 9-gon H' with smaller $|X \cap \mathrm{conv}(H')|$. (A similar argument was used in Sect. 4.2.)

Now let $t > 2$. We have to prove that if $|\bigcup_{n=1}^{t-1} T_n| < t$, a 9-gon H' with smaller $|X \cap \mathrm{conv}(H')|$ can be constructed. If $|\bigcup_{n=1}^{t-1} T_n| < t - 1$, we are done by the induction hypothesis as $|\bigcup_{n=1}^{t-2} T_n| \leq |\bigcup_{n=1}^{t-1} T_n|$. Therefore, assume that $|\bigcup_{n=1}^{t-1} T_n| = t - 1$. Label the consecutive vertices of $\mathrm{conv}(I)$ as U_l $(l \in \mathbb{N}_0)$ in such a way that $U_1 \in T_1$ and $U_0 \notin T_1$. By the induction hypothesis this implies $U_2 \in T_1$ as otherwise $|T_1| = 1$. Now construct sectors as follows: start with the 3-sector $(U_0V_1U_1)$ that can hold at most two vertices of the 9-gon without the occurrence of an empty convex hexagon. (As above, replace V_1 by V_1' if necessary.) Next construct the 4-sectors $(U_1V_1V_2U_2), (U_2V_2V_3U_3), \ldots, (U_{t-2}V_{t-2}V_{t-1}U_{t-1})$ that can hold at most one vertex of the 9-gon each if no empty convex hexagon is to occur; see Fig. 22.

Note that at each step the construction is well-defined by the induction hypothesis. We can construct the 4-sector $(U_1V_1V_2U_2)$ as $U_1, U_2 \in T_1$. Assume there exists a smallest $p \in \mathbb{N}$ such that $U_pV_pV_{p+1}U_{p+1}$ is not a convex quadrilateral. This means that $U_p \in \left(\bigcup_{m=1}^{p-1} T_m\right) \setminus T_p$ or $U_{p+1} \in \left(\bigcup_{m=p+1}^{t-1} T_m\right) \setminus T_p$. In the first case, this implies $|\bigcup_{m=p}^{t-1} T_m| \leq (t-1) - p$. In the second case, it follows that $|\bigcup_{m=1}^{p} T_m| \leq p$. In both cases, the induction hypothesis implies that a 9-gon H' with smaller $|X \cap \mathrm{conv}(H')|$ can be constructed.

Therefore, the 4-sectors can be constructed as described. Finally construct the 3-sector $(U_{t-1}V_tU_t)$ that can hold at most two vertices of the 9-gon without the

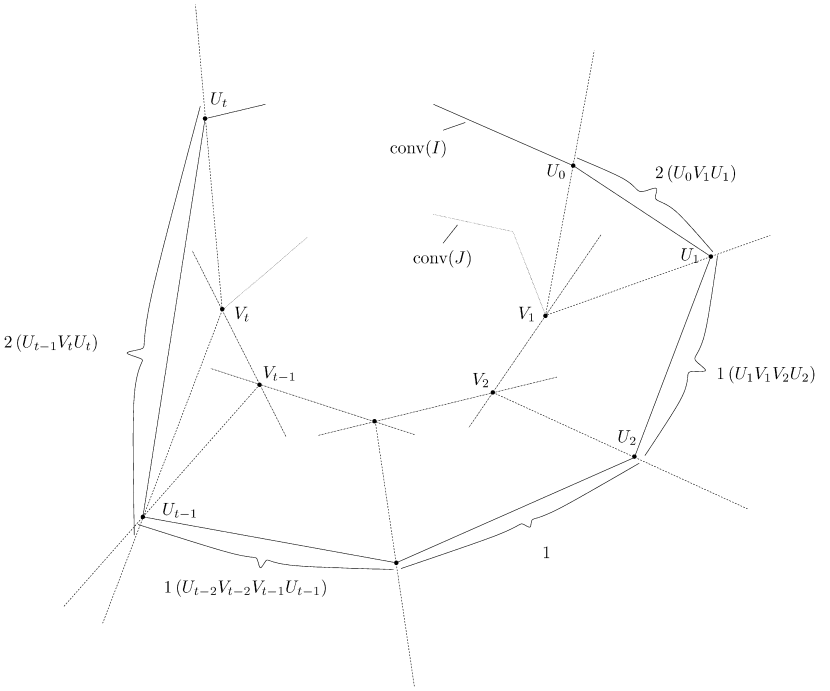


Fig. 22 Observation 1: $t > 2$

occurrence of an empty convex hexagon. (As above, replace V_t by V_t' if necessary.) Note that $U_t \notin T_{t-1}$ as we presume $|\bigcup_{n=1}^{t-1} T_n| = t-1$. It follows that at most $2 \cdot 2 + (t-2) \cdot 1 = t+2$ vertices of the 9-gon can lie in the union of sectors

$$(U_0V_1U_1) \cup \bigcup_{l=1}^{t-2} (U_lV_lV_{l+1}U_{l+1}) \cup (U_{t-1}V_tU_t).$$

Replacing these vertices by points from the convex chain $\overline{U_0V_1V_2\cdots V_tU_t}$ of length $t+2$ yields a 9-gon H' with smaller $|X \cap \text{conv}(H')|$. □

8.2 The Cases $(5, \geq 2)$

Consider the line through two consecutive vertices of conv(J), say P and Q, and let T_{PQ} be the set of vertices of the convex 5-gon lying in a halfplane that is defined by the line $\overline{PQ}$ and that does not contain any other points of J. (This halfplane is unique if $|J| > 2$.) Consider possible values for $|T_{PQ}|$:

- $|T_{PQ}| = 0$: This case is not possible by the definition of J.
- $|T_{PQ}| = 1$: In this case, a 9-gon H' with smaller $|X \cap \text{conv}(H')|$ can be constructed. Set $t = 2$ in Observation 1.
- $2 \leq |T_{PQ}| \leq 3$: This is the assumption for our subsequent considerations.
- $|T_{PQ}| > 3$: In this case, an empty convex hexagon can be constructed by using a convex chain of four vertices of the 5-gon together with P and Q.

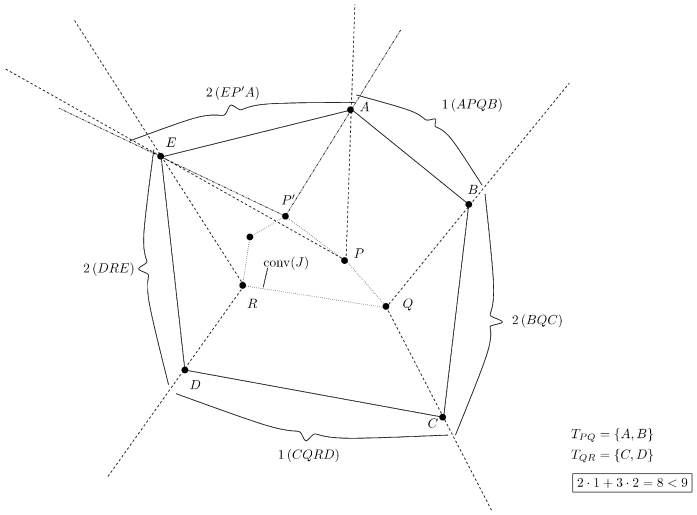


Fig. 23 The cases $(5, \geq 2)$: Example with $|T_{PQ} \cup T_{QR}| \geq 4$

Therefore, assume that

$$2 \leq |T_{PQ}| \leq 3. \tag{8.1}$$

Let R be the next vertex on the convex hull of J after passing through P and Q (if $|J| = 2$ then $R = P$). Define the set T_{QR} accordingly (take the other halfplane if $|J| = 2$). For the same reasons as above assume that

$$2 \leq |T_{QR}| \leq 3 \tag{8.2}$$

and consider the following three possibilities:

8.2.1 $|T_{PQ} \cup T_{QR}| \geq 4$

In this case we can choose consecutive vertices A, B, C, D of the 5-gon such that $A, B \in T_{PQ}$ and $C, D \in T_{QR}$. Label the remaining vertex of the 5-gon E. Construct the two 4-sectors $(APQB)$ and $(CQRD)$ that can hold at most one vertex of the 9-gon each without the occurrence of an empty convex hexagon. Next construct the 3-sector (BQC) that can hold at most two vertices of the 9-gon if no empty convex hexagon is to occur. Construct furthermore the two 3-sectors (DRE) and (EPA). Note that the union of these five sectors covers the complete region outside of conv(I); see also Fig. 23. Each of the two latter 3-sectors can hold at most two vertices of the 9-gon without the occurrence of an empty convex hexagon. (If necessary, replace R (respectively P) by appropriate $R' \in J \cap \triangle DRE$ and $P' \in J \cap \triangle EPA$ to obtain new 3-sectors $(DR'E)$ and $(EP'A)$ such that the corresponding triangles $\triangle DR'E$ and $\triangle EP'A$ do not contain any points of J as in the proof of Observation 1.) It follows that at most $2 \cdot 1 + 3 \cdot 2 = 8$ vertices of the 9-gon can be placed around the 5-gon without the occurrence of an empty convex hexagon. Note in particular that the case **(5, 2)** is covered by the argument in this subsection.

8.2.2 $|T_{PQ} \cup T_{QR}| = 3$

The case $|T_{PQ} \cup T_{QR}| = 3$ can be treated by the same approach as in the previous subsection. Choose consecutive vertices A, B, C of the 5-gon such that $A, B \in T_{PQ}$ and $B, C \in T_{QR}$. Label the remaining vertices of the 5-gon D and E such that the vertices C, D, E are consecutive. Construct the two 4-sectors $(APQB)$ and $(BQRC)$. Next construct the 3-sectors (CRD), (DRE) and (EPA). As above, replace the points R and P by appropriate points in J and modify the 3-sectors if necessary. Again, we arrive at the contradiction that at most $2 \cdot 1 + 3 \cdot 2 = 8$ vertices of the 9-gon can be placed around the 5-gon without the occurrence of an empty convex hexagon.

8.2.3 $|T_{PQ} \cup T_{QR}| \leq 2$

This case leaves the possibility of constructing a 9-gon H' with smaller $|X \cap \text{conv}(H')|$. Set $t = 3$ in Observation 1.

9 The Cases (6, ≥ 4)

The approach is similar to the one in Sect. 8. The key idea is to partition the region outside of $\text{conv}(I)$ into two 3-sectors and four 4-sectors. Each 3-sector is defined by two consecutive vertices of the 6-gon and one vertex of $\text{conv}(J)$. It can hold at most two vertices of the 9-gon if no empty convex hexagon is to occur. Each 4-sector is defined by two consecutive vertices of the 6-gon and two consecutive vertices of $\text{conv}(J)$. It can hold at most one vertex of the 9-gon without the occurrence of an empty convex hexagon. It follows that a total of $2 \cdot 2 + 4 \cdot 1 = 8$ vertices of the 9-gon can be placed around the 6-gon without the occurrence of an empty convex hexagon.

Consider a chain of consecutive vertices of $\text{conv}(J)$, $\overline{VWXYZ}$, where $V = Z$ if $j = 4$. Define the sets T_{VW}, T_{WX}, T_{XY} and T_{YZ} as in Sect. 8 (that is, T_{VW} is the set of vertices of the convex 6-gon lying in the halfplane defined by the line $\overline{VW}$ that does not contain any other points of J, etc.). As in Sect. 8, we assume that

$$2 \leq |T_{KL}| \leq 3 \quad ((K, L) \in \{(V, W), (W, X), (X, Y), (Y, Z)\}). \tag{9.1}$$

By setting $t = 3, 4, 5$ in Observation 1 (Sect. 8), it follows that we may also assume that

$$|T_{KL} \cup T_{LM}| \geq 3, \tag{9.2}$$

$$|T_{KL} \cup T_{LM} \cup T_{MN}| \geq 4, \tag{9.3}$$

$$|T_{VW} \cup T_{WX} \cup T_{XY} \cup T_{YZ}| \geq 5 \tag{9.4}$$

with $(K, L, M) \in \{(V, W, X), (W, X, Y), (X, Y, Z)\}$ (in (9.2)) and $(K, L, M, N) \in \{(V, W, X, Y), (W, X, Y, Z)\}$ (in (9.3)). Note that (9.4) also holds in the case **(6, 4)**, where Observation 1 does not apply (since $t > j$). Note furthermore that by construction it is not possible that there is a $P \in T_{KL} \cap T_{MN}$ with $P \notin T_{LM}$ $((K, L, M, N) \in \{(V, W, X, Y), (W, X, Y, Z)\})$. We now give an explicit construction for the two 3-sectors and the four 4-sectors. A concrete example can be found in Fig. 24. The combinatorial subcases are depicted in Fig. 25.

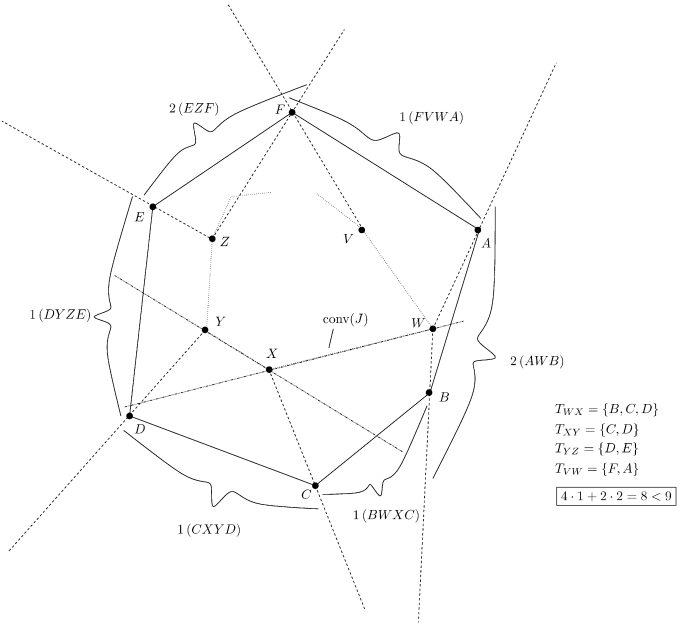


Fig. 24 The cases $(6, \geq 4)$: Example with $T_{WX} \cap T_{XY} \neq \emptyset$

9.1 $T_{WX} \cap T_{XY} \neq \emptyset$

Label the consecutive vertices of the 6-gon A, B, C, D, E, F such that $B \in T_{WX}, C \in T_{WX} \cap T_{XY}$ and $D \in T_{XY}$. Note that $F \notin T_{WX}$ and $F \notin T_{XY}$ as otherwise $|T_{WX}| > 3$ or $|T_{XY}| > 3$. Consider the following possibilities:

(1) $A \notin (T_{VW} \cup T_{WX})$; see Fig. 25a. It follows from (9.1) and (9.2) that $B, C \in T_{VW}$ and $D \in T_{WX}$. (9.3) implies $E \in T_{XY}$. Construct the three 4-sectors $(BVWC)$, $(CWXD)$ and $(DXYE)$. Next, construct the 3-sector (AVB). (Replace V by an appropriate $V' \in J \cap \triangle AVB$ if necessary.)

- If $E \in T_{YZ}$ then (9.4) implies $F \in T_{YZ}$. Construct the 4-sector $(EYZF)$ and the 3-sector (FZA). (Again, replace Z by Z' if necessary.)
- If $E \notin T_{YZ}$ then it follows from (9.1) that $A, F \in T_{YZ}$. ($F \notin T_{XY}$ implies in particular $F \notin T_{XY} \setminus T_{YZ}$.) In this case construct the 3-sector (EYF) together with the 4-sector $(FYZA)$.

In both cases we arrive at a set of four 4-sectors and two 3-sectors as claimed. *In the following cases, assume that* $A \in (T_{VW} \cup T_{WX})$.

(2) $E \notin (T_{XY} \cup T_{YZ})$. This case is symmetric to the previous one. Therefore, *in the following assume that* $E \in (T_{XY} \cup T_{YZ})$.

(3) $A \in T_{WX} \setminus T_{VW}$; see Fig. 25b. It follows from (9.1) that $E, F \in T_{VW}$. ($F \notin T_{WX}$ implies in particular that $F \notin T_{WX} \setminus T_{VW}$.) Construct the 3-sectors (FWA) and (BXC) together with the 4-sectors $(EVWF)$, $(AWXB)$ and $(CXYD)$. It follows that $D \in T_{YZ}$ as otherwise $|T_{YZ} \cup T_{VW}| = |\{E, F\}| < 3$. Note that $(E \in T_{VW}) \wedge (E \in (T_{XY} \cup T_{YZ}))$ implies $E \in T_{YZ}$. Therefore, we

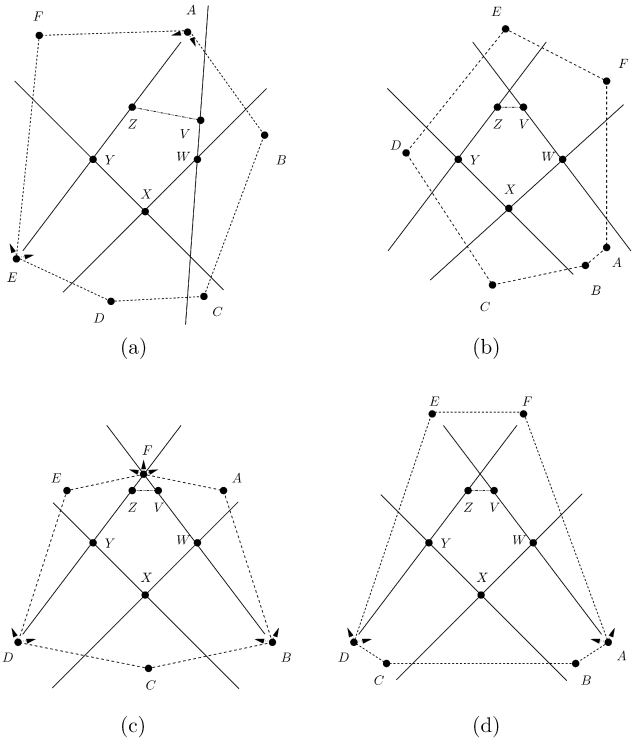


Fig. 25 The cases $(6, \geq 4)$: Combinatorial subcases. The assumption in (**a**)–(**c**) is that $B \in T_{WX}$, $C \in T_{WX} \cap T_{XY}$ and $D \in T_{XY}$. In (**a**), $A \notin (T_{VW} \cup T_{WX})$. In (**b**), $A \in T_{WX} \setminus T_{VW}$ and $E \in (T_{XY} \cup T_{YZ})$. In (**c**), $A \in T_{VW}$ and $E \in T_{YZ}$. In (**d**), it is assumed that $A, B \in T_{WX} \setminus T_{XY}$ and $C, D \in T_{XY} \setminus T_{WX}$. Only those point positions that are essential for the construction of the sectors are indicated

can construct the 4-sector $(DYZE)$. Again the six sectors can be constructed as claimed. *In the following assume that* $A \in T_{VW}$.

(4) $E \in T_{XY} \setminus T_{YZ}$. This case is symmetric to the previous one. Therefore, *in the following assume that* $E \in T_{YZ}$.

(5) $A \in T_{VW}$ and $E \in T_{YZ}$; see Fig. 25c. Construct the 4-sectors $(BWXC)$ and $(CXYD)$. Consider the following four possibilities:

- $B \in T_{VW} \wedge D \in T_{YZ}$. Construct the 4-sector $(AVWB)$ together with the 3-sector (FVA). (Replace V by V' if necessary.) Accordingly, construct the 4-sector $(DYZE)$ together with the 3-sector (EZF). (Replace Z by Z' if necessary.)
- $B \notin T_{VW} \wedge D \in T_{YZ}$. Construct the 4-sector $(DYZE)$ together with the 3-sector $(EZ'F)$ as in the previous subcase. If $B \notin T_{VW}$, it follows from (9.1) that $F \in T_{VW}$. In this case, construct the 3-sector (AWB) together with the 4-sector $(FVWA)$.
- $B \in T_{VW} \wedge D \notin T_{YZ}$. This subcase is symmetric to the previous one.
- $B \notin T_{VW} \wedge D \notin T_{YZ}$. It follows that $F \in T_{VW}$ and $F \in T_{YZ}$. Accordingly, construct the 4-sectors $(FVWA)$ and $(EYZF)$ together with the 3-sectors (AWB) and (DYE).

In each case, we arrive at a set of four 4-sectors and two 3-sectors that cover the complete region outside of conv(I) as claimed.

9.2 $T_{WX} \cap T_{XY} = \emptyset$

See Fig. 25d. Then by construction, there exist consecutive vertices A, B, C, D of conv(I) such that $A, B \in T_{WX} \setminus T_{XY}$ and $C, D \in T_{XY} \setminus T_{WX}$. Construct the 4-sectors $(AWXB)$ and $(CXYD)$ as well as the 3-sector (BXC). Label the remaining vertices of the 6-gon E, F such that D, E, F are consecutive. Now distinguish four possibilities:

- $D \in T_{YZ} \wedge A \in T_{VW}$. It follows that $E \in T_{YZ}$ as otherwise $|T_{XY} \cup T_{YZ}| < 3$. Accordingly, $F \in T_{VW}$ as otherwise $|T_{VW} \cup T_{WX}| < 3$. Construct the 4-sectors $(DYZE)$ and $(FVWA)$ together with the 3-sector (EZF). (Replace Z by an appropriate Z' if necessary.)
- $D \notin T_{YZ} \wedge A \in T_{VW}$. As in the previous case, construct the 4-sector $(FVWA)$. If $E \in T_{XY} \setminus T_{YZ}$ it follows that $|T_{YZ} \cup T_{VW} \cup T_{WX}| = |\{A, B, F\}| < 4$. Therefore, assume that $E \in T_{YZ}$. It follows that $F \in T_{YZ}$ as otherwise $|T_{YZ}| < 2$. Construct the 3-sector (DYE) and the 4-sector $(EYZF)$.
- $D \in T_{YZ} \wedge A \notin T_{VW}$. This case is symmetric to the previous one.
- $D \notin T_{YZ} \wedge A \notin T_{VW}$. Note that this case is not feasible as it would imply $|T_{YZ} \cup T_{VW}| = |\{E, F\}| < 3$.

In each feasible case, we can construct the six sectors as claimed above.

10 The Cases ($\geq 7, \geq 5, \geq 1$)

Up to this point, we have settled all cases except for **($\geq 7, \geq 5, \geq 1$)**. These cases, except for three special cases (see below), can all be settled via the same set of arguments. As above, let $K := \text{conv}(J) \cap (X \setminus \partial J)$. Fix a point $P \in K$. Consider rays emanating from P through each vertex of the convex j-gon conv(J). This divides the region outside the j-gon into j sectors and in each sector at most two vertices of conv(I) can lie without forming an empty convex hexagon. (To see this, construct 3-sectors and replace P by an appropriate $P' \in K$ where needed.) Consider all possible vertex distributions. (These are summarized in Table 3.) We want to partition the region outside the convex i-gon conv(I) into sectors and to show that in each case at most eight vertices of the 9-gon can be placed inside the union of these sectors without creating an empty convex hexagon. The following three simple rules are sufficient to prove this:

10.1 The First Rule

The first rule deals with two vertices of conv(I) lying in the same sector.

Rule 1 *Let* A_1, A_2 *denote two consecutive vertices of* conv(I) *lying in the same sector* (aPb), *where* a *and* b *are consecutive vertices of* conv(J). *Then no vertex of the* 9*-gon can lie in the sector* (A_1abA_2) *without the occurrence of an empty convex hexagon.*

Table 3 The cases ($\geq 7, \geq 5, \geq 1$): Combinatorial subcases. (Π indicates possible permutations)

(7, 5, ≥ 1)	**(8, 5, ≥ 1)**
$\Pi(2, 2, 2, 1, 0)$	$\Pi(2, 2, 2, 2, 0)$
$\Pi(2, 2, 1, 1, 1)$	$\Pi(2, 2, 2, 1, 1)$
(7, 6, ≥ 1)	**(8, 6, ≥ 1)**
$\Pi(2, 2, 2, 1, 0, 0)$	$\Pi(2, 2, 2, 2, 0, 0)$
$\Pi(2, 2, 1, 1, 1, 0)$	$\Pi(2, 2, 2, 1, 1, 0)$
$(2, 1, 1, 1, 1, 1)$	$\Pi(2, 2, 1, 1, 1, 1)$
(7, 7, ≥ 1)	**(8, 7, ≥ 1)**
$\Pi(2, 2, 2, 1, 0, 0, 0)$	$\Pi(2, 2, 2, 2, 0, 0, 0)$
$\Pi(2, 2, 1, 1, 1, 0, 0)$	$\Pi(2, 2, 2, 1, 1, 0, 0)$
$\Pi(2, 1, 1, 1, 1, 1, 0)$	$\Pi(2, 2, 1, 1, 1, 1, 0)$
$(1, 1, 1, 1, 1, 1, 1)$	$(2, 1, 1, 1, 1, 1, 1)$
(7, 8, ≥ 1)	**(8, 8, ≥ 1)**
$\Pi(2, 2, 2, 1, 0, 0, 0, 0)$	$\Pi(2, 2, 2, 2, 0, 0, 0, 0)$
$\Pi(2, 2, 1, 1, 1, 0, 0, 0)$	$\Pi(2, 2, 2, 1, 1, 0, 0, 0)$
$\Pi(2, 1, 1, 1, 1, 1, 0, 0)$	$\Pi(2, 2, 1, 1, 1, 1, 0, 0)$
$(1, 1, 1, 1, 1, 1, 1, 0)$	$\Pi(2, 1, 1, 1, 1, 1, 1, 0)$
	$(1, 1, 1, 1, 1, 1, 1, 1)$

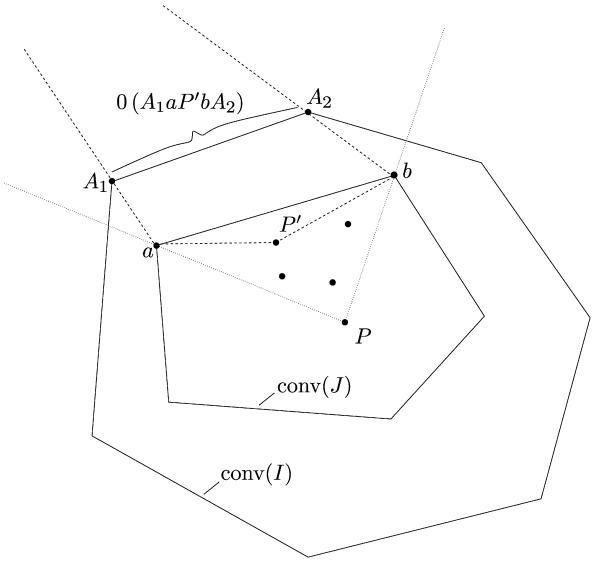
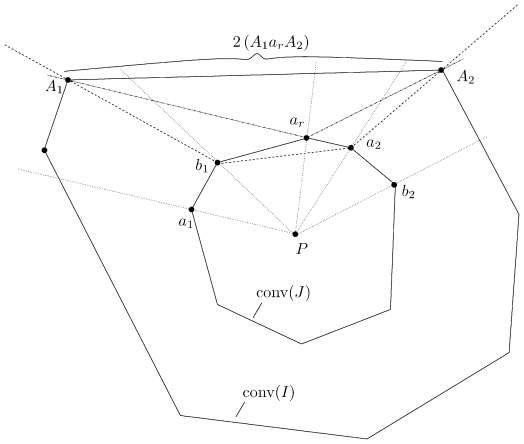


Fig. 26 Rule 1

Proof The claim follows directly from the presence of an empty convex 5-gon $A_1aP'bA_2$, where $P' \in J \cap \triangle aPb$ is chosen appropriately; see Fig. 26. □

Fig. 27 Rule 2



10.2 The Second Rule

The second rule gives an upper bound on the number of vertices of the 9-gon that can lie between two non-empty sectors.

Rule 2 *Let A_1, A_2 denote two consecutive vertices of* $\mathrm{conv}(I)$ *lying in distinct sectors (a_1Pb_1) and (a_2Pb_2), where a_1 and b_1 respectively a_2 and b_2 are consecutive vertices of* $\mathrm{conv}(J)$. *Suppose that a_1, b_1, a_2, b_2 are part of a chain of consecutive vertices of* $\mathrm{conv}(J)$. *Let $\mathcal{S} := (A_1b_1a_2A_2)$ if $A_1b_1a_2A_2$ is a convex quadrilateral and $\mathcal{S} := (A_1b_1A_2) \cup (A_1a_2A_2)$ otherwise. Then at most two vertices of the* 9*-gon can lie within $\mathcal{S}$.*

Remark 1 It is possible in Rule 2 that $b_1 = a_2$.

Proof A 3-sector that does not contain any points of J and covers the region $\mathcal{S}$ can be constructed by choosing A_1, A_2 and an appropriate a_r among the consecutive vertices of $\mathrm{conv}(J)$ between b_1 and a_2 (inclusively); see Fig. 27. □

10.3 Application of Rules 1 and 2

The first two rules are already sufficient to settle the cases $(7, 5, \geq 1)$ with distributions $\Pi(2, 2, 2, 1, 0)$, $(7, 6, \geq 1)$ with distributions $\Pi(2, 2, 2, 1, 0, 0)$, $(7, 7, \geq 1)$ with distributions $\Pi(2, 2, 2, 1, 0, 0, 0)$, $(7, 8, \geq 1)$ with distributions $\Pi(2, 2, 2, 1, 0, 0, 0, 0)$, $(8, 5, \geq 1)$ with distributions $\Pi(2, 2, 2, 2, 0)$, $(8, 6, \geq 1)$ with distributions $\Pi(2, 2, 2, 2, 0, 0)$, $(8, 7, \geq 1)$ with distributions $\Pi(2, 2, 2, 2, 0, 0, 0)$ and $(8, 8, \geq 1)$ with distributions $\Pi(2, 2, 2, 2, 0, 0, 0, 0)$. To see this, apply Rule 1 whenever two consecutive vertices of $\mathrm{conv}(I)$ lie in the same sector. Note that two such vertices correspond to a 2 in the underlying distribution. For consecutive vertices of $\mathrm{conv}(I)$ lying in distinct sectors, apply Rule 2. Note that in the cases at hand, Rule 2 needs to be applied exactly four times as there are always exactly four non-zero entries in the corresponding distribution sequences. It follows that at most $4 \cdot 2 = 8$ vertices of the 9-gon can be placed without the occurrence of an empty convex hexagon. An example is given in Fig. 28.

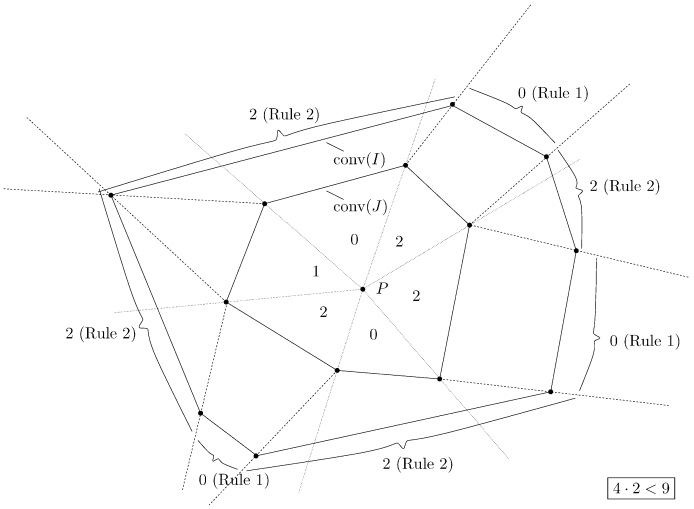


Fig. 28 Application of Rules 1 and 2: Example for the case $(7, 6, \geq 1)$ with distribution $(2, 0, 2, 1, 0, 2)$

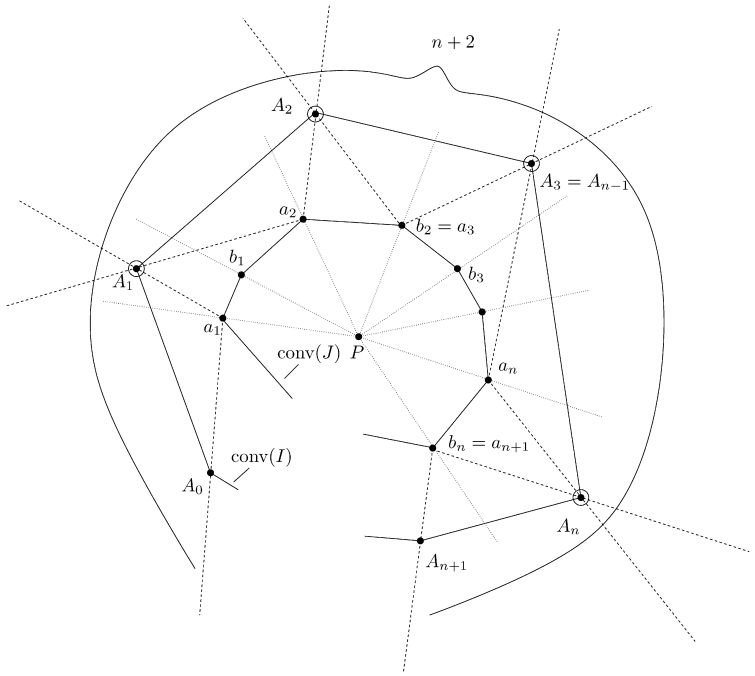
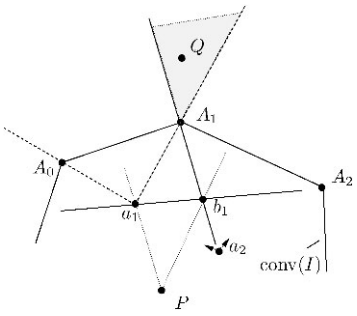


Fig. 29 Rule 3

10.4 The Third Rule

The third rule deals with a sequence of sectors, where each sector contains at most one vertex of conv(I). See also Fig. 29.

Fig. 30 Proof of Rule 3: $n = 1$



Rule 3 *Let* $1 \le n \le i-2$. *Consider a sequence* $A_0, A_1, \ldots, A_{n+1}$ *of consecutive vertices of* $\mathrm{conv}(I)$. *For* $1 \le l \le n+1$, *let* $A_l \in (a_l P b_l)$, *where* a_l *and* b_l *are consecutive vertices of* $\mathrm{conv}(J)$. *Suppose that for* $1 \le l \le n$, *each sector* $(a_l P b_l)$ *contains exactly one vertex of* $\mathrm{conv}(I)$ *and that* $a_1, b_1, a_2, b_2, \ldots, a_{n+1}, b_{n+1}$ *are part of a chain of consecutive vertices of* $\mathrm{conv}(J)$. *Then at most* $n+2$ *vertices of the* 9*-gon lie in the union of sectors* $\bigcup_{l=1}^{n+1}(A_{l-1}a_l A_l)$.

Remark 2 It is possible in Rule 3 that $b_l = a_{l+1}$ $(1 \le l \le n)$ or $b_{n+1} = a_1$. Furthermore, it is possible that A_0 and A_{n+1} both lie in $(a_{n+1} P b_{n+1})$.

Proof We prove by induction over n.

If $n = 1$, we can argue that it is not possible that A_1 lies above the line $\overline{a_1 b_1}$ while A_0 and A_2 lie below it, where *lying below* refers to lying in the halfplane defined by $\overline{a_1 b_1}$ that includes P. Otherwise, $|T_{a_1 b_1}| = 1$ and a 9-gon H' with smaller $|X \cap \mathrm{conv}(H')|$ could be constructed. (Set $t = 2$ in Observation 1 in Sect. 8.) Assume that A_0 also lies above $\overline{a_1 b_1}$. (The case that only A_1 and A_2 lie above the line is similar.) Construct the 4-sector $(A_0 a_1 b_1 A_1)$ together with the 3-sector $(A_1 a_2 A_2)$. If necessary, replace a_2 by an appropriate $a_2' \in \triangle A_1 a_2 A_2$ to obtain a new 3-sector $(A_1 a_2' A_2)$ with no points of J lying in $\triangle A_1 a_2' A_2$. Together, the 4- and the 3-sector cover (at least) the region of $(A_0 a_1 A_1) \cup (A_1 a_2 A_2)$. This is clear for points lying in $(A_1 a_2 A_2)$ since $(A_1 a_2' A_2)$ covers (at least) this region. Note that there cannot be a point $Q \in ((A_0 a_1 A_1) \setminus (A_1 a_2' A_2)) \setminus (A_0 a_1 b_1 A_1)$. Such a point would have to lie in the shaded region in Fig. 30. If a_2 lies to the right of $\overline{b_1 A_1}$ (or $a_2 = b_1$) then $Q \in (A_1 a_2' A_2)$. Otherwise $b_1 \in (A_1 a_2 A_2)$ and we could have chosen $a_2' := b_1$. The 4- and the 3- sector allow for at most $1 + 2 = 3$ vertices of the 9-gon without the occurrence of an empty convex hexagon.

For the induction step, assume that the claim is true for $1, 2, \ldots, n-1$. By the induction hypothesis, we know that at most $(n-1)+2$ vertices of the 9-gon can lie in the union of sectors $\bigcup_{l=1}^{n}(A_{l-1}a_l A_l)$. At most two additional vertices of the 9-gon can lie in the sector $(A_n a_n A_{n+1}) \setminus \bigcup_{l=1}^{n}(A_{l-1}a_l A_l)$ without the occurrence of an empty convex hexagon as it is part of the 3-sector $(A_n a_n A_{n+1})$. Therefore, the number of vertices of the 9-gon that can lie in the union of sectors $\bigcup_{l=1}^{n+1}(A_{l-1}a_l A_l)$ is at most $(n-1+2)+2 = n+3$ if no empty convex hexagon is to occur. It also follows from the induction hypothesis that at most $(n-1)+2$ vertices of the 9-gon can lie in the union of sectors $\bigcup_{l=2}^{n+1}(A_{l-1}a_l A_l)$ without the occurrence of an empty

convex hexagon. Accordingly, at most two additional vertices of the 9-gon can lie in the sector $(A_0a_1A_1) \setminus \bigcup_{l=2}^{n+1}(A_{l-1}a_lA_l)$ if no empty convex hexagon is to occur. Therefore, the above bound is sharp if and only if exactly two vertices of the 9-gon lie in the sectors $(A_0a_1A_1)$ and $(A_na_{n+1}A_{n+1})$ respectively.

It follows that A_0 must lie below the line $\overline{a_1b_1}$ and A_{n+1} must lie below the line $\overline{a_nb_n}$ as otherwise one could again replace one of the 3-sectors $(A_0a_1A_1)$ and $(A_na_{n+1}A_{n+1})$ by the 4-sector $(A_0a_1b_1A_1)$ respectively $(A_na_nb_nA_{n+1})$ as above. This sector could hold only one vertex of the 9-gon (without the occurrence of an empty convex hexagon) and the union of all sectors would still cover the same region.

This implies $|\bigcup_{l=1}^{n} T_{a_lb_l}| = n < n+1$, though, and a 9-gon H' with smaller $|X \cap \mathrm{conv}(H')|$ can be constructed by Observation 1. To see this, note that $a_1, b_1, a_2, b_2, \ldots, a_n, b_n$ are part of a chain of consecutive vertices of $\mathrm{conv}(J)$ of length $L \geq n+1$. Therefore, the claim follows. □

10.5 Application of Rules 1–3

Based on the three rules we can now settle all the remaining subcases of $(\geq 7, \geq 5, \geq 1)$ with the exception of $(7, 7, \geq 1)$ with distribution $(1, 1, 1, 1, 1, 1, 1)$, $(7, 8, \geq 1)$ with distribution $(1, 1, 1, 1, 1, 1, 1, 0)$ and $(8, 8, \geq 1)$ with distribution $(1, 1, 1, 1, 1, 1, 1, 1)$. (These cases do not allow for a direct application of Rule 3. They are treated individually in the following subsections.) In the other cases, at least one 2 appears in the distribution sequence. We can argue as follows:

Whenever two consecutive vertices of $\mathrm{conv}(I)$ lie within the same sector, apply Rule 1. Note that two such vertices correspond to a 2 in the underlying distribution. No vertices of the 9-gon can lie in the corresponding sectors.

Now take maximal series of consecutive sectors containing at most one vertex of $\mathrm{conv}(I)$ each and apply Rule 3 (respectively Rule 2 if none of them contains a vertex). The number of vertices of the 9-gon that can lie in the union of all corresponding sectors is equal to $q + s \cdot 2$, where q is the total number of 1's in the underlying distribution and s is the number of distinct series. Note that s is equal to the number of gaps between two occurrences of a 2 in the distribution sequence. As this number is equal to the number of 2's in the sequence, it follows that $q + s \cdot 2$ is equal to the sum of the elements of the distribution sequence. It can easily be verified that this sum is always smaller than 9. Therefore, in all these cases an empty convex hexagon occurs. (An example is given in Fig. 31.)

10.6 The Case $(1, 1, 1, 1, 1, 1, 1)$

This case can be dealt with by applying Rule 3 with $n = 5$ seven times with each vertex of $\mathrm{conv}(I)$ as a starting point. Each 3-sector $(A_{r-1}a_rA_r)$ is left out exactly once. Therefore, in the union of all sectors at most $(7 \cdot (5+2))/6 < 9$ vertices of the 9-gon can lie without the occurrence of an empty convex hexagon.

10.7 The Case $(1, 1, 1, 1, 1, 1, 1, 0)$

For the case $(1, 1, 1, 1, 1, 1, 1, 0)$, label the vertices of the polygon $\mathrm{conv}(J)$ in clockwise order a_l $(1 \leq l \leq 8)$ and assume that the sector (a_6Pa_7) is the one that does not

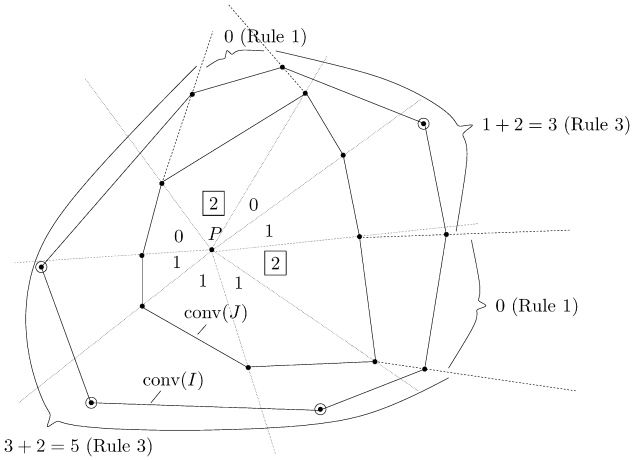
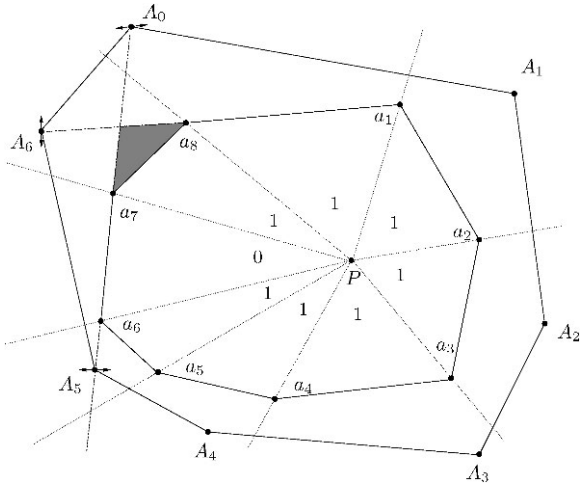


Fig. 31 The case $(8, 8, \geq 1)$ with $(2, 1, 1, 1, 0, 2, 0, 1)$

Fig. 32 The case $(7, 8, \geq 1)$ with $(1, 1, 1, 1, 1, 1, 1, 0)$



contain a vertex of conv(I). Applying Rule 3 with $n = 5$, we can conclude that at most $5 + 2 = 7$ vertices of the 9-gon can lie in the union of sectors $\bigcup_{l=1}^{6}(A_{l-1}a_lA_l)$; see Fig. 32. Consider A_6. Note that it is not possible that A_6 lies below the line $\overline{a_1a_8}$ *and* below the line $\overline{a_6a_7}$ (where *below* refers to the halfplane that includes P), as otherwise a 9-gon $H' := (a_8a_1a_2 \cdots a_7A_6)$ with smaller $|X \cap \text{conv}(H')|$ is present.

If A_6 lies above the line $\overline{a_1a_8}$, only one vertex of the 9-gon can lie in the then existing 4-sector $(A_6a_8a_1A_0)$ and therefore, without the occurrence of an empty convex hexagon, at most eight vertices of the 9-gon can lie in the union of sectors

$$\bigcup_{l=1}^{6}(A_{l-1}a_lA_l) \cup (A_6a_8a_1A_0),$$

which by construction covers the complete region outside of conv(I).

Similarly, if A_0 lies above the line $\overline{a_6a_7}$ (and therefore also A_6 by construction), at most eight vertices of the 9-gon can lie in the union of sectors

$$\bigcup_{l=1}^{6}(A_{l-1}a_lA_l) \cup (A_6a_6a_7A_0)$$

(which by construction covers the complete region outside of $\text{conv}(I)$) without the occurrence of an empty convex hexagon.

Finally, if A_0 lies below the line $\overline{a_6a_7}$ and A_6 lies above it, we know that A_5 must also lie above the line $\overline{a_6a_7}$ as otherwise $|T_{a_6a_7}| < 2$ and a 9-gon H' with smaller $|X \cap \text{conv}(H')|$ could be constructed by Observation 1. Therefore, the 4-sector $(A_5a_6a_7A_6)$ exists, which can only hold one vertex of the 9-gon without the occurrence of an empty convex hexagon. Now, applying Rule 3 with $n = 5$ yields that at most seven vertices of the 9-gon can lie in the union of sectors

$$(A_6a_8A_0) \cup \bigcup_{l=1}^{5}(A_{l-1}a_lA_l)$$

without the occurrence of an empty convex hexagon. Therefore, without the occurrence of an empty convex hexagon, at most eight vertices of the 9-gon can lie in the union of sectors

$$(A_6a_8A_0) \cup \bigcup_{l=1}^{5}(A_{l-1}a_lA_l) \cup (A_5a_6a_7A_6)$$

which by construction covers the complete region outside of $\text{conv}(I)$.

10.8 The Case (1, 1, 1, 1, 1, 1, 1, 1)

Note that in the case (1, 1, 1, 1, 1, 1, 1, 1), applying the induction argument with $n = 6$ eight times with each vertex of $\text{conv}(J)$ as a starting point (in analogy to our approach to the case $(7, 7, \geq 1)$ with distribution (1, 1, 1, 1, 1, 1, 1) in Sect. 10.6) only gives us an estimate of a total of $(8 \cdot (6 + 2))/7 > 9$ vertices of the 9-gon that can lie in the union of all sectors. Therefore, a different approach for this subcase is required.

Label the vertices of $\text{conv}(J)$ in clockwise order as a_r $(1 \leq r \leq 8)$. Consider four consecutive vertices of the convex 8-gon $\text{conv}(J)$, a_s, a_t, a_u and a_v. Note that no vertex of $\text{conv}(I)$ can lie below the line $\overline{a_sa_t}$ and below the line $\overline{a_ua_v}$ (where *below* refers to the halfplane that includes P) as otherwise we could use such a vertex to construct a 9-gon H' with smaller $|X \cap \text{conv}(H')|$. Denote by R_t the region above both lines $\overline{a_sa_t}$ and $\overline{a_ta_u}$; see Fig. 33. The union of all regions R_r $(1 \leq r \leq 8)$ defines the feasible region for vertices of $\text{conv}(I)$. Label the vertices of $\text{conv}(I)$ as A_m $(A_m \in (a_mPa_{m+1}),\ 1 \leq m \leq 8,\ a_9 := a_1)$. Note that A_m lies in R_m or R_{m+1} (or both) $(1 \leq m \leq 8,\ R_9 := R_1)$. Consider the following three possibilities:

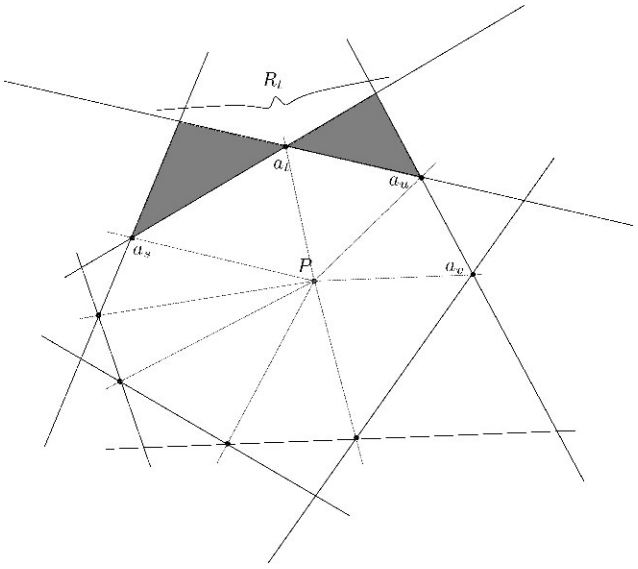


Fig. 33 The case $(8, 8, \geq 1)$ with $(1, 1, 1, 1, 1, 1, 1, 1)$: Definition R_t

10.8.1 There Exists a Region R_w with no A_m Lying in It

There can be at most one such region as otherwise one could construct a 9-gon H' with smaller $|X \cap \text{conv}(H')|$. To see this, eliminate successively the possibilities that the next R_z with this property is R_{w+1}, R_{w+2}, R_{w+3} or R_{w+4}. In the first case, $|T_{a_w a_{w+1}}| < 2$ and in the other three cases one can replace one to three vertices of the 8-gon $a_1a_2 \cdots a_8$ by two to four points A_m in such a way that a 9-gon H' with smaller $|X \cap \text{conv}(H')|$ appears.

Let a_1 be the vertex associated with the region that does not contain any A_m. Since the existence of such a region is independent of the choice of P, we may assume that P lies in the pentagon $a_6a_2a_3a_4a_5$. Such a P must exist for otherwise an empty convex hexagon appears; see also Fig. 34. (A different choice of P might result in a different distribution sequence. If this is the case, we arrive at a subcase that has already been settled.) As a consequence, at most three vertices of the 9-gon can lie in the sector $(A_6a_6Pa_2A_1)$ as otherwise a 9-gon H' with smaller $|X \cap \text{conv}(H')|$ could be constructed (as $5 + 4 = 9$).

We claim that $A_m \in R_m \cap R_{m+1}$ for $2 \leq m \leq 7$; that is, each A_m lies above both lines $\overline{a_{m-1}a_m}$ and $\overline{a_{m+1}a_{m+2}}$ $(2 \leq m \leq 7,\ a_9 := a_1)$. To see this, start from the line $\overline{a_1a_2}$ and work clockwise to prove that A_m is above the line $\overline{a_{m-1}a_m}$ $(2 \leq m \leq 7)$. Note that configurations, where $|\bigcup_{n=1}^{l-1} T_{a_n a_{n+1}}| < l$ $(2 \leq l \leq 8)$ yield a 9-gon H' with smaller $|X \cap \text{conv}(H')|$ by Observation 1 (Sect. 8). Now start from the line $\overline{a_1a_8}$ and work counter clockwise to prove that A_m is above the line $\overline{a_{m+1}a_{m+2}}$ $(7 \geq m \geq 2,$ $a_9 := a_1)$.

Finally, as we are assuming that no A_r lies in R_1, it follows that $A_1 \in R_2$ and $A_8 \in R_8$. Therefore, this case can be settled as indicated in Fig. 34.

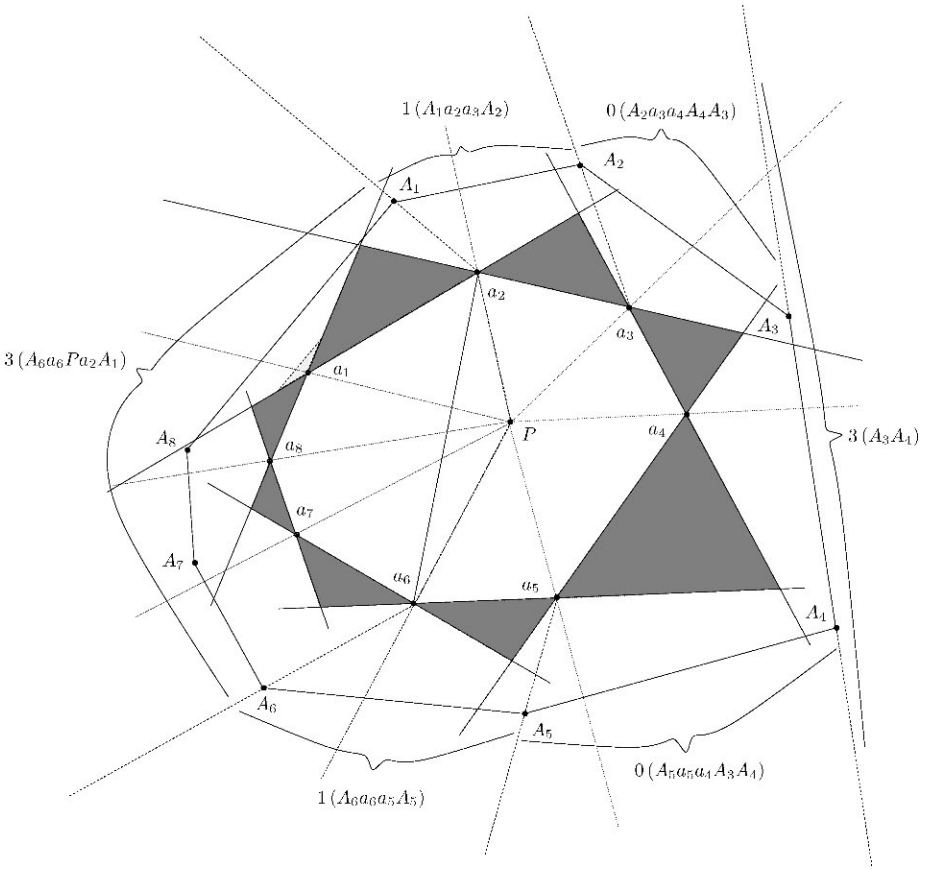


Fig. 34 The case $(8, 8, \geq 1)$ with $(1, 1, 1, 1, 1, 1, 1, 1)$: R_1 contains no A_m

10.8.2 (At Least) One A_m Lies in Each Region R_r ($1 \leq r \leq 8$) and, Say, $A_1 \in R_1 \setminus R_2$

We claim that this implies that $A_u \in R_u$ $(3 \leq u \leq 8)$ as otherwise a 9-gon H' with smaller $|X \cap \text{conv}(H')|$ appears. To see this, first consider the point A_8. If $A_8 \in R_1 \setminus R_8$, the 9-gon $H' := A_1a_2a_3 \cdots a_8A_8$ with smaller $|X \cap \text{conv}(H')|$ occurs. Therefore, $A_8 \in R_8$. Next, consider A_7, then A_6 and so on. Finally, as we are assuming that at least one A_m lies in each region R_r $(1 \leq r \leq 8)$ it follows that $A_2 \in R_2$; see Fig. 35. In this case, the region outside of $\text{conv}(I)$ can be partitioned into eight 4-sectors $(A_la_la_{l+1}A_{l+1})$ $(1 \leq l \leq 8,\ a_9 := a_1,\ A_9 := A_1)$ that together allow at most eight vertices of the 9-gon without the occurrence of an empty convex hexagon.

10.8.3 Each A_m Lies in Both R_m and R_{m+1}

Again, the region outside of $\text{conv}(I)$ can be partitioned into eight 4-sectors $(A_la_la_{l+1}A_{l+1})$ $(1 \leq l \leq 8,\ a_9 := a_1,\ A_9 := A_1)$ that together allow at most eight vertices of the 9-gon without the occurrence of an empty convex hexagon. □

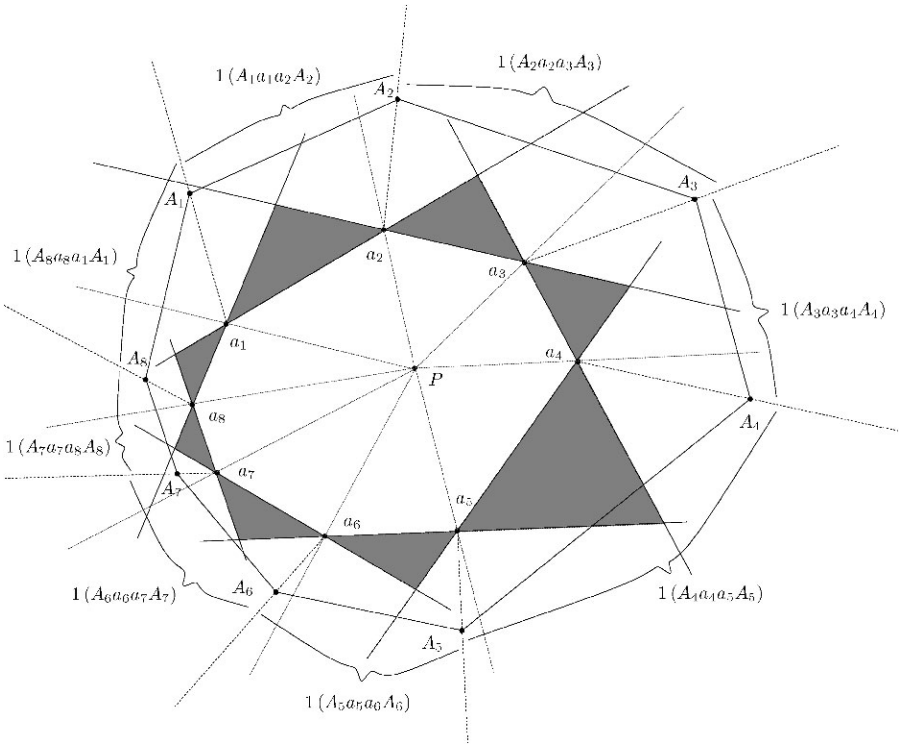


Fig. 35 The case $(8, 8, \geq 1)$ with $(1, 1, 1, 1, 1, 1, 1, 1)$: $A_1 \in R_1 \setminus R_2$

Acknowledgements The author would like to thank Pavel Valtr and David Wood as well as the anonymous referees for their detailed comments on preliminary versions of this paper.

References

1. Bárány, I., Károlyi, Gy.: Problems and results around the Erdős-Szekeres convex polygon theorem. In: Discrete and Computational Geometry, Tokyo, 2000. Lect. Notes Comput. Sci., vol. 2098. Springer, Berlin (2001)
2. Brass, P., Moser, W., Pach, J.: Research Problems in Discrete Geometry. Springer, New York (2005)
3. Erdős, P.: Some more problems on elementary geometry. Aust. Math. Soc. Gaz. **5**, 52–54 (1978)
4. Erdős, P.: Some applications of graph theory and combinatorial methods to number theory and geometry. In: Algebraic Methods in Graph Theory, vols. I, II, Szeged, 1978. Colloq. Math. Soc. János Bolyai, vol. 25, pp. 137–148. North-Holland, Amsterdam (1981)
5. Erdős, P., Szekeres, G.: A combinatorial problem in geometry. Compos. Math. **2**, 463–470 (1935)
6. Erdős, P., Szekeres, G.: On some extremum problems in elementary geometry. Ann. Univ. Sci. Bp. Eötvös Sect. Math. **3–4**, 53–62 (1960/1961)
7. Harborth, H.: Konvexe Fünfecke in ebenen Punktmengen. Elem. Math. **33**, 116–118 (1978)
8. Horton, J.D.: Sets with no empty convex 7-gons. Can. Math. Bull. **26**, 482–484 (1983)
9. Morris, W., Soltan, V.: The Erdős-Szekeres problem on points in convex position—a survey. Bull. Am. Math. Soc. **37**, 437–458 (2000)
10. Overmars, M.: Finding sets of points without empty convex 6-gons. Discrete Comput. Geom. **29**, 153–158 (2003)
11. Tóth, G., Valtr, P.: The Erdős-Szekeres theorem: upper bounds and related results. In: Goodman, J.E., Pach, J., Welzl, E. (eds.) Combinatorial and Computational Geometry. MSRI Publications, vol. 52, pp. 557–568. Cambridge University Press, Cambridge (2005)

Affinely Regular Polygons as Extremals of Area Functionals

Paolo Gronchi · Marco Longinetti

Abstract For any convex n-gon P we consider the polygons obtained by dropping a vertex or an edge of P. The area distance of P to such $(n-1)$-gons, divided by the area of P, is an affinely invariant functional on n-gons whose maximizers coincide with the affinely regular polygons. We provide a complete proof of this result.

We extend these area functionals to planar convex bodies and we present connections with the affine isoperimetric inequality and parallel X-ray tomography.

Keywords Affinely regular polygons · Geometric tomography · Affine length

1 Introduction

Given a convex polygon P with n vertices z_j ordered counterclockwise, we define $W_j(P)$ to be the triangle $z_{j-1}z_jz_{j+1}$ and $T_j(P)$ to be the (possibly unbounded) triangle outside P bounded by the side z_jz_{j+1} and the continuations of the two adjacent sides (see Fig. 1). Henceforth, the index j is taken modulo n, and $|C|$ denotes the area of C.

In this paper we consider the following affinely invariant functionals defined on the class $\mathcal{P}_n$ of planar convex n-gons, i.e., polygons with *exactly* n vertices:

P. Gronchi (✉)
Dipartimento di Matematica e Applicazioni per l'Architettura, Università degli Studi di Firenze, Piazza Ghiberti 27, 50122 Firenze, Italy
e-mail: paolo@fi.iac.cnr.it

M. Longinetti
Dipartimento di Ingegneria Agraria e Forestale, Università degli Studi di Firenze, P. le delle Cascine 15, 50139 Firenze, Italy
e-mail: marco.longinetti@unifi.it

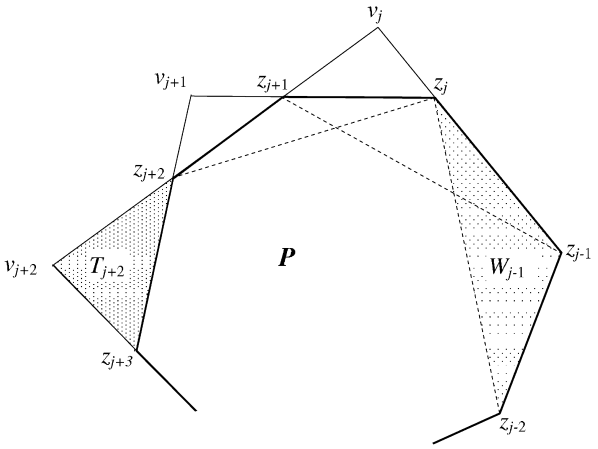


Fig. 1 Triangles W_j and T_j of P

$$F(P) = \min_{j=1,\dots,n} \frac{|W_j(P)|}{|P|}, \tag{1}$$

$$G(P) = \min_{j=1,\dots,n} \frac{|T_j(P)|}{|P|}, \tag{2}$$

and we are interested in the maximizers of these functionals.

In Theorem 1.8 it is shown that *the maximizers of the above functionals are affinely regular n-gons, i.e., affine images of regular n-gons*. This class, denoted by $\mathcal{R}_n$, often appears in geometric problems with affine invariance [1, 3, 5, 17, 20].

A characterization of $\mathcal{R}_n$ as extremals of area functionals was obtained by Renyi and Sulanke [20]. They proved in Satz 2 that $\prod_{i=1}^{n} |W_i(P)|/|P|^n$ attains its maximum on $\mathcal{R}_n$.

The functional F was first introduced by Lopez and Reisner [17] in connection with algorithms for the approximation of a convex set by polygons. They showed that Theorem 1.8 for the functional F is a consequence of the result by Renyi and Sulanke.

The functional G was first introduced by Longinetti [15], where Theorem 1.8 is proved for $n = 5, 6$, via elementary geometric arguments. The functional G and a similar functional (not affinely invariant) considered in [14] are related to Hammer's X-ray problem for planar convex bodies proposed in [10]: How many X-ray pictures of a convex body must be taken in order to permit its reconstruction? The solution of this problem is given by Gardner and McMullen [6]. We refer to [7, Chap. 1] for an overview of this topic. In Sect. 6 we present in detail the connection between the functional G and the stability of the reconstruction in the Hammer's problem. In Sect. 5, we discuss some extensions of functionals F and G to the class of planar convex bodies related to the affine length of a convex body and to the affine isoperimetric inequality. These functionals are also related to the approximation of planar convex bodies by polygons [9, 17]. In particular, F is related to the approximation of an n-gon P by $(n-1)$-gons contained in P. Similarly, G is related with the approximation of P by $(n-1)$-gons containing P. Because of this we use the word *inner* or *outer* in connection with F or G, respectively. In higher dimension, similar function-

als involving polytopes obtained by dropping a vertex or a facet were investigated by Reisner et al. [19].

As a first remark we deal with the trivial cases $n = 3, 4$. For $n = 3$, we have $F(P) = 1$ and $G(P) = \infty$, for every P. For $n = 4$, by elementary arguments, one can prove that the maximizers of F have diagonals which divide them into triangles of equal area, and $G(P) = \infty$ only for parallelograms. Hence, all maximizers of F and G are parallelograms, and vice versa. As mentioned, some instances of Theorem 1.8 were already proved. In this paper we complete this result in a common framework to all n, for both the inner and outer functional.

We now provide a guide to the proof of Theorem 1.8. The functionals F and G are continuous with respect to the Hausdorff metric in $\mathcal{P}_n$, which is not compact, since n-gons can converge to polygons with fewer vertices. Hence, we have first to prove the existence of the maximizers in Lemma 2.1.

A first step towards the characterization of these maximizers is to show that they satisfy an inner (or outer) *equal-area property* which has some interest by itself. Consider the following classes:

$$\Phi_n = \left\{ P \in \mathcal{P}_n : |W_1(P)| = |W_2(P)| = \cdots = |W_n(P)| \right\}, \tag{3}$$

$$\Gamma_n = \left\{ P \in \mathcal{P}_n : |T_1(P)| = |T_2(P)| = \cdots = |T_n(P)| \right\}. \tag{4}$$

Henceforth, we say that a polygon P has the inner (outer) equal-area property when P belongs to Φ_n (Γ_n, respectively). Section 2.1 contains the proof of the following proposition.

Proposition 1.1 *F and G attain their maximum on Φ_n and Γ_n, respectively.*

It can be proved that the classes Φ_5 and Γ_5 coincide with the class of affinely regular pentagons. For $n > 5$, it is easy to see that the above class is larger than the class $\mathcal{R}_n$ of affinely regular polygons. For example, hexagons in Γ_6 are, up to an affine transformation, the intersection of two concentric equilateral triangles (see the shaded polygon in Fig. 2). A proof can be found in [15]. Similarly, hexagons in Φ_6 are, up to an affine transformation, equiangular (see the larger polygon in Fig. 2). In [11] the polygons of Φ_n are considered and a larger class of Φ_n containing not necessarily

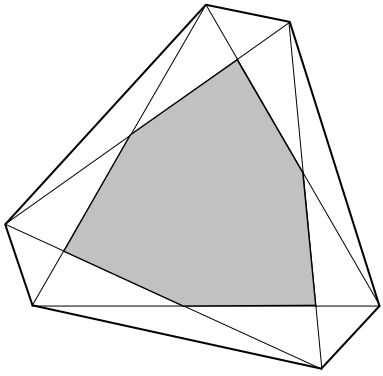


Fig. 2 Polygons from Φ_6 and Γ_6

convex n-gons is explicitly parametrized by $n-5$ real parameters modulo the action of the affine group.

Roughly speaking, (3) or (4) involve $n-1$ independent constraints. Since n-gons depend on the $2n$ coordinates of its vertices, it follows that Φ_n and Γ_n depend on $n+1$ parameters, and therefore, modulo the action of the affine group, depend on $n-5$ parameters. So, in order to show that all maximizers of F and G are in $\mathcal{R}_n$ other significant properties must be proved. A subclass of Φ_{2m} was characterized by Bianchi and Longinetti [2, Lemma 1].

Let l_j be the length of the side $z_j z_{j+1}$ of P, i.e. $l_j = \|z_j - z_{j+1}\|$ and d_j the length of the diagonal $z_{j-1}z_{j+2}$.

Definition 1.2 We define $\mathcal{F}_n$ as the class of convex n-gons such that

$$\left(\frac{d_{j+1}-l_{j+1}}{d_{j+1}}\right)\left(\frac{d_j-l_j}{l_j}\right) = \left(\frac{d_{j-2}-l_{j-2}}{d_{j-2}}\right)\left(\frac{d_{j-1}-l_{j-1}}{l_{j-1}}\right) \quad \text{for } j = 1,\dots,n. \tag{5}$$

We say that P has the *inner-ratio property* when $P \in \mathcal{F}_n$. Assuming that all triangles T_j of P are bounded, we define v_j as the vertex of T_j not in P. Also, set $s_j = \|v_j - z_j\|$ and $p_j = \|z_{j+1} - v_j\|$. Let e_j be the length of the segment joining the outer points v_{j+1}, v_{j-1}. Notice that $e_j = p_{j-1} + l_j + s_{j+1}$.

Definition 1.3 We define $\mathcal{G}_n$ as the class of convex n-gons such that

$$\frac{s_j e_{j-1}}{l_{j-1}(p_{j-2}+l_{j-1})} = \frac{p_j e_{j+1}}{l_{j+1}(s_{j+2}+l_{j+1})} \quad \text{for } j = 1,\dots,n. \tag{6}$$

We call this the *outer-ratio property*. Notice that the outer-ratio property is an equality between ratios of lengths of segments on the two lines through v_j in Fig. 1.

The following theorems, proved in Sects. 2.1, 2.2, are important steps toward the goal.

Theorem 1.4 *If $P^\star$ is a maximizer of F in $\mathcal{P}_n$, then $P^\star \in \mathcal{F}_n$.*

Theorem 1.5 *If $P^\star$ is a maximizer of G in $\mathcal{P}_n$, then $P^\star \in \mathcal{G}_n$.*

In Sect. 3, through an algebraic manipulation, we will prove that in (5) the ratios $\lambda_j = d_j/l_j$ are independent of j. Analogously, for the outer problem, we will prove that in (6) the ratios $\zeta_j = s_j/l_{j-1} = p_{j-1}/l_j$ are independent of j. In Sect. 4 we go back to planar geometry and prove the following theorems.

Theorem 1.6

$$\Phi_n \cap \mathcal{F}_n = \mathcal{R}_n.$$

Theorem 1.7

$$\Gamma_n \cap \mathcal{G}_n = \mathcal{R}_n.$$

These results permit us to obtain the goal of the paper.

Theorem 1.8 *All maximizers of F or G on $\mathcal{P}_n$ are affinely regular polygons.*

2 First Variations at Extremal Polygons

Observe that for each $P \in \mathcal{P}_n$ the triangles $W_j(P)$ have positive area, hence $F(P) > 0$, but if P is close to a polygon with $n-1$ vertices, then $F(P)$ is close to zero. Therefore $\inf_{P\in\mathcal{P}_n} F(P) = 0$ and F has no minimum in $\mathcal{P}_n$. Similarly, $\inf_{P\in\mathcal{P}_n} G(P) = 0$ and G has no minimum in $\mathcal{P}_n$. Then *all the extremals of the functionals F and G in $\mathcal{P}_n$ are maximizers.* In this section we obtain the more significant properties of these polygons. We use only elementary arguments of Euclidean geometry in the proof of Proposition 1.1. The idea of the proof is to consider a maximizer $P^\star$ of each functional and a suitable local variation $P_\varepsilon^\star$ of one or two vertices. An analysis of the sign of the area difference Δ_ε yields the results.

To prove directly Theorems 1.4 and 1.5, which are the principal goal of this section, a more complicated perturbation $P_\varepsilon^\star$ of $P^\star$ involving five consecutive vertices of $P^\star$ can be carried out. In this case one has to take into account only the first order terms of Δ_ε. This computation, using Proposition 1.1, can be explicitly obtained in terms of the sides and angles of $P^\star$.

Here, we prefer to give a different proof, less geometric, via partial differentiation of area functionals (Lemmas 2.2 and 2.3) with respect to the vertices z_j. This yields the Lagrange multiplier systems (21) and (28) for the area $|P|$ under the corresponding equal-area property constraints. At the end of each subsection, an algebraic manipulation of such systems will give the proof of Theorem 1.4 for the inner problem and the proof of Theorem 1.5 for the outer one.

We begin with the following result.

Lemma 2.1 *The functionals F and G have maxima in $\mathcal{P}_n$.*

Proof From the well-known John theorem about the maximal ellipse contained in P, see [12], we can restrict to n-gons with fixed area 1, whose boundaries are contained in a circular annulus A of radii r and $2r$. Clearly $r \geq 1/\sqrt{4\pi}$ and the diameter of P is less or equal to $D = 4/\sqrt{\pi}$. Represent each polygon in $\mathcal{P}_n$ as a point in $\mathbb{R}^{2n}$ with coordinates the coordinates of its vertices. With respect to the standard metric on $\mathbb{R}^{2n}$, F is continuous and the class of k-gons, $k \leq n$, with vertices contained in A is compact. Since $F > 0$, it is trivial to show that a convergent maximizing sequence on $\mathcal{P}_n$ has vertices which converge to n distinct points, no three collinear.

Turning to G, some difficulties arise since $|T_j|$ may be infinite. Let α_j be the exterior angle of P at the vertex z_j. If $T_j(P)$ is bounded then

$$|T_j(P)| = \frac{1}{2} l_j^2 (\cot\alpha_{j+1} + \cot\alpha_j)^{-1}. \tag{7}$$

The function $\cot x$ is convex in $(0, \pi/2)$ and symmetric with respect to the point $\pi/2$. Thus

$$\cot\left(\frac{x_1 + x_2}{2}\right) \leq \frac{\cot x_1 + \cot x_2}{2} \quad \text{if } x_1, x_2 > 0 \text{ and } x_1 + x_2 < \pi. \tag{8}$$

Since $\alpha_j > 0$ and $\sum_{j=1}^{n} \alpha_j = 2\pi$, there exists j such that

$$\frac{\alpha_{j+i} + \alpha_j}{2} \leq \frac{2\pi}{n}.$$

Therefore, by (7) and (8) it follows that

$$\min_{j=1,\ldots,n} |T_j(P)| < \frac{1}{4} D^2 \tan\frac{2\pi}{n}.$$

Hence, G is bounded from above on $\mathcal{P}_n$, for $n \geq 5$.

Now consider a maximizing sequence $\{P^m\}$ of n-gons with vertices in the circular annulus A, which converges to $P^\star$. It remains to show that $P^\star$ has exactly n vertices. We can suppose

$$\min_{j=1,\ldots,n} |T_j(P^m)| > \frac{1}{2} \sup_{P \in \mathcal{P}_n} G(P) = \mu > 0, \quad \text{for all } m. \tag{9}$$

For each P with vertices in A, consider a triangle similar to $T_j(P)$ bounded by the continuations of its sides not in P and a line parallel to the side $z_{j+1}z_j$ through the center of A. If h_j denotes the altitude of $T_j(P)$ to the side $z_{j+1}z_j$, then this larger triangle has an altitude smaller than $h_j + 2r$ and a base larger than $2r$. Hence

$$\frac{l_j}{h_j} \geq \frac{2r}{h_j + 2r},$$

i.e., $l_j \geq (2r - l_j)h_j/2r$. Since $l_j h_j = 2|T_j(P)| \geq 2\mu$, we deduce $l_j^2 \geq (2r - l_j)\mu/r$ and consequently that the sides of the polygons P^m are uniformly larger than a positive constant. This implies that n distinct points $z_j^\star$ of $P^\star$ are limits of the sequences of the vertices of P^m. Moreover, from (9), the limit of the area $|T_j(P^m)|$ is positive. Hence, no three consecutive points $z_j^\star$ are collinear and they are all distinct vertices of $P^\star$. □

Proof of Proposition 1.1 The claim that G attains its maximum on Γ_n was already proved in [15, Theorem 1].

For the inner case, assume that $P^\star$ is a maximizer of F. Let W_r be a triangle of maximal area among the n triangles W_j. Moving the vertex z_r towards the interior of $P^\star$ reduces the area of $P^\star$. Since F cannot increase, the value $\min|W_j|$ has to decrease. Hence, either W_{r-1} or W_{r+1} is of minimal area. The freedom we have in choosing the direction along which z_r moves easily implies that they are both of minimal area. Therefore,

$$|z_{r-1}z_{r-2}z_{r-3}| = |W_{r-2}| \geq |W_{r-1}| = |z_r z_{r-1} z_{r-2}|,$$
$$|z_{r+1}z_{r+2}z_{r+3}| = |W_{r+2}| \geq |W_{r+1}| = |z_r z_{r+1} z_{r+2}|.$$

Then, the distance of z_{r-3} from the line containing the edge $z_{r-2}z_{r-1}$ is larger or equal to that of z_r. Analogously, the distance of z_{r+3} from the line containing $z_{r+2}z_{r+1}$ is larger or equal to that of z_r. From the convexity of $P^\star$, the triangle $U = z_{r-3}z_r z_{r+3}$ is ordered counterclockwise and contained in $P^\star$. Therefore, any movement of z_r inside U increases the area of both W_{r-1} and W_{r+1} and decreases the area of P. The maximality of $P^\star$ implies that $|W_r|$ cannot be larger than $\min |W_j|$. □

We represent with complex variable $x_j + iy_j$ the vertices z_j, i.e. z_j represents both a point in the plane and a complex number; so the area functionals $|T_j|$, $|W_j|$ and $|P|$ are real functions of complex variables z_j. We use partial derivatives of functions with respect to complex variables with the notation

$$f_{\overline{z}} = \frac{\partial f}{\partial \overline{z}} = \frac{1}{2}\left(\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y}\right).$$

Lemma 2.2 *If T is the triangle with vertices a, b, c ordered counterclockwise, then*

$$\frac{4}{i}\frac{\partial |T|}{\partial \overline{a}} = c - b. \tag{10}$$

Proof If $a = x_a + iy_a$, $b = x_b + iy_b$, and $c = x_c + iy_c$ the result is obtained by an elementary computation starting from the formula

$$\begin{aligned}\frac{4}{i}|T| &= \frac{2}{i}\big((x_c - x_b)(y_a - y_b) - (x_a - x_b)(y_c - y_b)\big)\\ &= (c-b)(\overline{a} - \overline{b}) - (\overline{c} - \overline{b})(a - b).\end{aligned} \tag{11}$$

□

Lemma 2.3 *Let b, c, d and e be the vertices of a convex quadrilateral ordered clockwise. Let abc be the triangle T outside the quadrilateral bounded by bc and the continuations of the two sides eb and dc. The area of $|T|$ depending on b, c, d, e, satisfies*:

$$-4i|T|_{\overline{b}} = (a - c) + (a - b)|ae|/|be|, \tag{12}$$

$$-4i|T|_{\overline{c}} = (b - a) + (c - a)|ad|/|cd|, \tag{13}$$

$$-4i|T|_{\overline{d}} = (a - c)|ac|/|cd|, \tag{14}$$

$$-4i|T|_{\overline{e}} = (b - a)|ab|/|be|. \tag{15}$$

Proof Observe that the vertex a is a function of b, c, d, e and each proof starts by differentiating (11) with respect to these variables. For example, in order to get (13) we have

$$4i|T|_{\overline{c}} = -(c - b)\overline{a}_{\overline{c}} + (\overline{c} - \overline{b})a_{\overline{c}} + (a - b). \tag{16}$$

To compute the partial derivatives of a and $\overline{a}$ with respect to $\overline{c}$ we consider the collinear conditions of a with b and e and with c and d, i.e.

$$(a-b)(\overline{e}-\overline{b})-(e-b)(\overline{a}-\overline{b})=0, \tag{17}$$

$$(a-c)(\overline{c}-\overline{d})-(c-d)(\overline{a}-\overline{c})=0, \tag{18}$$

and, by differentiating, we obtain

$$(\overline{e}-\overline{b})a_{\overline{c}}-(e-b)\overline{a}_{\overline{c}}=0,$$
$$(\overline{c}-\overline{d})a_{\overline{c}}-(c-d)\overline{a}_{\overline{c}}=d-a.$$

Solving for $a_{\overline{c}}, \overline{a}_{\overline{c}}$ in the previous equations we obtain

$$a_{\overline{c}}=(e-b)(d-a)\big/\big((e-b)(\overline{c}-\overline{d})-(\overline{e}-\overline{b})(c-d)\big),$$
$$\overline{a}_{\overline{c}}=(\overline{e}-\overline{b})(d-a)\big/\big((e-b)(\overline{c}-\overline{d})-(\overline{e}-\overline{b})(c-d)\big).$$

By substituting in (16) we get

$$\frac{4}{i}\frac{\partial|T|}{\partial\overline{c}}=(b-a)+(d-a)\frac{(c-b)(\overline{e}-\overline{b})-(\overline{c}-\overline{b})(e-b)}{(e-b)(\overline{c}-\overline{d})-(\overline{e}-\overline{e})(c-d)}.$$

The proof of (13) is obtained via the formula

$$(c-b)(\overline{e}-\overline{b})-(\overline{c}-\overline{b})(e-b)=\frac{|ac|}{|cd|}\big((e-b)(\overline{c}-\overline{d})-(\overline{e}-\overline{b})(c-d)\big). \tag{19}$$

Indeed, by subtracting the left-hand side of (17) from the left hand side of (19) we deduce

$$(c-b)(\overline{e}-\overline{b})-(\overline{c}-\overline{b})(e-b)=(c-a)(\overline{e}-\overline{b})-(\overline{c}-\overline{a})(e-b).$$

Since a, c, d are collinear the vector $(c-a)$ is proportional to the vector $(d-c)$ by the factor $|ac|/|cd|$. This proves (19).

The proof of (12) can be obtained in a similar way or simply interchanging b and c, e and d, by a reflection. We remark that such a reflection changes the sign as in formula (11).

The proof of (14) follows from similar computations. More explicitly, by differentiating (11) with respect to $\overline{d}$ we obtain

$$4i|T|_{\overline{d}}=-(c-b)\overline{a}_{\overline{d}}+(\overline{c}-\overline{b})a_{\overline{d}}. \tag{20}$$

By differentiating with respect to $\overline{d}$ the constraints (17) and (18), we get

$$(\overline{e}-\overline{b})a_{\overline{d}}-(e-b)\overline{a}_{\overline{d}}=0,$$
$$(\overline{c}-\overline{d})a_{\overline{d}}-(c-d)\overline{a}_{\overline{d}}=a-c.$$

This implies

$$a_{\overline{d}}=(e-b)(a-c)\big/\big((e-b)(\overline{c}-\overline{d})-(\overline{e}-\overline{b})(c-d)\big),$$
$$\overline{a}_{\overline{d}}=(\overline{e}-\overline{b})(a-c)\big/\big((e-b)(\overline{c}-\overline{d})-(\overline{e}-\overline{b})(c-d)\big).$$

By substituting in (20) we get

$$\frac{4}{i}\frac{\partial|T|}{\partial\overline{d}} = (a-c)\frac{(c-b)(\overline{e}-\overline{b}) - (\overline{c}-\overline{b})(e-b)}{(e-b)(\overline{c}-\overline{d}) - (\overline{e}-\overline{b})(c-d)},$$

and from (19) we prove (14). A similar argument proves (15). □

2.1 Inner Case

Proposition 2.4 *Let $P^\star$ be a maximizer of F on $\mathcal{P}_n$. Then there exist real numbers $\mu_1, \ldots, \mu_n$ such that the following equations hold for the vertices of $P^\star$:*

$$(z_{j-1} - z_{j+1}) = \mu_{j-1}(z_{j-1} - z_{j-2}) + \mu_j(z_{j-1} - z_{j+1}) + \mu_{j+1}(z_{j+2} - z_{j+1}), \quad (21)$$

for $j = 1, \ldots, n$.

Proof By Proposition 1.1, each maximizer $P^\star$ satisfies the inner equal-area property. Choosing a suitable affine transformation we can assume that $|W_k(P)| = 1$ for $k = 1, \ldots, n$. By the Lagrange multipliers argument we have that at $P^\star$ the gradient of the area functional $|P|$ is a linear combination of the gradients of the constraints $|W_k(P)| = 1$, i.e. there exist real multipliers μ_k such that at $P^\star$,

$$\frac{\partial|P|}{\partial\overline{z_j}} = \sum_k \mu_k \frac{\partial|W_k|}{\partial\overline{z_j}} \quad \text{for } j = 1, \ldots, n.$$

Since W_k depends only on z_{k-1}, z_k, z_{k+1} we have that

$$\frac{\partial|W_k|}{\partial\overline{z_j}} = 0 \quad \text{for } j \notin \{k-1, k, k+1\}.$$

Now we apply Lemma 2.2 to the triangles W_{j-1}, W_j, W_{j+1} with respect to the vertex z_j and find that

$$\frac{4}{i}\frac{\partial|W_{j-1}|}{\partial\overline{z_j}} = z_{j-1} - z_{j-2}, \qquad \frac{4}{i}\frac{\partial|W_j|}{\partial\overline{z_j}} = z_{j-1} - z_{j+1},$$
$$\frac{4}{i}\frac{\partial|W_{j+1}|}{\partial\overline{z_j}} = z_{j+2} - z_{j+1}.$$

Since P can be decomposed in the disjoint subsets $W_j, P \setminus W_j$ and the latter does not depend on the vertex z_j, we have also

$$\frac{4}{i}\frac{\partial|P|}{\partial\overline{z_j}} = \frac{4}{i}\frac{\partial|W_j|}{\partial\overline{z_j}} = z_{j-1} - z_{j+1}. \quad (22)$$

The six previous equations prove (21). □

The system (21) is a system of $2n$ real equations in the $3n$ real unknowns z_j, μ_j, and so it cannot determine the maximizers. The inner equal-area property adds more information. In particular, we have that for each j the vector $(z_{j+2} - z_{j-1})$ is parallel to $(z_{j+1} - z_j)$. So any maximizer $P^\star$ also satisfies the following system for suitable positive λ_j:

$$(z_{j+2} - z_{j-1}) = \lambda_j (z_{j+1} - z_j) \quad \text{for } j = 1, \ldots, n. \tag{23}$$

We notice that the λ_j's are well defined and positive because all vertices of $P^\star$ are distinct and $P^\star$ is convex. Moreover, we can prove that

$$\lambda_j > 1 \quad \text{for } j = 1, \ldots, n. \tag{24}$$

Indeed, suppose that there exists a $\lambda_j \leq 1$. Up to an affine transformation we can assume that z_{j-1}, z_j, z_{j+1} are three consecutive vertices of a square Q. Since $\lambda_j \leq 1$, by (23), z_{j+2} belongs to Q. Using (23) again, we obtain that $z_{j+3} - z_j$ has the direction of $z_{j+2} - z_{j+1}$. This means that z_{j+3} belongs to a line supporting Q at z_j. Any choice of z_{j+3} on such a line gives a contradiction, since either z_{j-1} belongs to the convex hull of the other vertices, or the side $z_{j+2}z_{j+3}$ intersects the boundary of Q.

By the definition (23) of λ_j, we have $\lambda_j = d_j / l_j$ and then the inner-ratio property (5) can be rewritten as

$$\frac{(\lambda_{j-2} - 1)(\lambda_{j-1} - 1)}{\lambda_{j-2}} = \frac{(\lambda_{j+1} - 1)(\lambda_j - 1)}{\lambda_{j+1}} \quad \text{for } j = 1, \ldots, n. \tag{25}$$

This is the form we get in the following proof.

Proof of Theorem 1.4 From (23) we get

$$z_{j+2} = z_{j-1} + \lambda_j (z_{j+1} - z_j) \quad \text{and}$$
$$z_{j-2} = z_{j+1} - \lambda_{j-1} (z_j - z_{j-1}).$$

By substituting in (21) and rearranging the terms we obtain

$$(z_{j+1} - z_j)(\mu_{j-1} + \mu_j + \mu_{j+1} - 1 - \lambda_j \mu_{j+1}) + (z_j - z_{j-1})(\mu_{j-1} + \mu_j + \mu_{j+1} - 1 - \lambda_{j-1} \mu_{j-1}) = 0.$$

Since $P^\star \in \mathcal{P}_n$ the vertices z_j are distinct and no three collinear. This implies that the vectors $(z_{j+1} - z_j)$, $(z_j - z_{j-1})$ are linearly independent and their coefficients in the previous equation must be zero. Hence, we get for $j = 1, \ldots, n$,

$$\begin{cases} \mu_{j-1} + \mu_j + \mu_{j+1} - 1 = \lambda_j \mu_{j+1}, \\ \mu_{j-1} + \mu_j + \mu_{j+1} - 1 = \lambda_{j-1} \mu_{j-1}. \end{cases} \tag{26}$$

We infer that the λ_j's are such that the linear system (26) has a solution $\mu_1, \mu_2, \ldots, \mu_n$. Taking into account the six equations involving only the five unknowns

$\mu_{j-2}, \mu_{j-1}, \mu_j, \mu_{j+1}$, and μ_{j+2}, we deduce that the determinant of the matrix

$$\begin{pmatrix} 1 & 1 & 1-\lambda_{j-1} & 0 & 0 & 1 \\ 0 & 1 & 1 & 1-\lambda_j & 0 & 1 \\ 0 & 0 & 1 & 1 & 1-\lambda_{j+1} & 1 \\ 1-\lambda_{j-2} & 1 & 1 & 0 & 0 & 1 \\ 0 & 1-\lambda_{j-1} & 1 & 1 & 0 & 1 \\ 0 & 0 & 1-\lambda_j & 1 & 1 & 1 \end{pmatrix}$$

has to be zero. After some manipulations, we get the following equation

$$\lambda_j\lambda_{j-2}(1-\lambda_{j+1})\lambda_{j-1}(1-\lambda_j) - \lambda_{j-1}\lambda_{j+1}(1-\lambda_{j-2})\lambda_j(1-\lambda_{j-1}) = 0. \quad (27)$$

Since all $\lambda_j > 0$, the previous equation can be simplified as in (25). □

2.2 Outer Case

In order to get the Lagrange multipliers system for the outer case in a simple way we recall that v_j is the intersection of the two lines through the consecutive vertices z_{j-1}, z_j and the vertices z_{j+1}, z_{j+2}, (see Fig. 1), $l_j = \|z_{j+1} - z_j\|$, i.e., the length of the jth side, and $s_j = \|v_j - z_j\|$, $p_j = \|z_{j+1} - v_j\|$.

Proposition 2.5 *Let $P^\star$ be a maximizer of the functional G on $\mathcal{P}_n$, $n \geq 5$. Then there exist real numbers $\eta_1, \ldots, \eta_n$ such that the following equations hold for the vertices of $P^\star$, for $j = 1, \ldots, n$:*

$$\begin{aligned}(z_{j-1} - z_{j+1}) = {} & \eta_{j-2}(z_{j-1} - v_{j-2})\frac{p_{j-2}}{l_{j-1}} \\ & + \eta_{j-1}\left((v_{j-1} - z_{j-1}) + (v_{j-1} - z_{j+1})\frac{p_{j-1}}{l_j}\right) \\ & + \eta_j\left((z_{j+1} - v_j) + (z_{j-1} - v_j)\frac{s_j}{l_{j-1}}\right) \\ & + \eta_{j+1}(v_{j+1} - z_{j+1})\frac{s_{j+1}}{l_j}. \end{aligned} \quad (28)$$

Proof By Proposition 1.1 each maximizer $P^\star$ satisfies the outer equivalent triangle property. Choosing a suitable affine transformation we can assume that $|T_j(P)| = 1$, for $j = 1, \ldots, n$. By the Lagrange multipliers argument, at $P^\star$ the gradient of the area functional $|P|$ is a linear combination of the gradients of the constraints $|T_k(P)| = 1$, i.e. there exist real multipliers η_j such that, at $P^\star$,

$$\frac{\partial |P|}{\partial \overline{z_j}} = \sum_k \eta_k \frac{\partial |T_k|}{\partial \overline{z_j}} \quad \text{for } j = 1, \ldots, n.$$

Since $|T_k|$ depends only on $z_{k-1}, z_k, z_{k+1}, z_{k+2}$ we get

$$\frac{4}{i}\frac{\partial |T_k|}{\partial \overline{z_j}} = 0 \quad \text{for } k \notin \{j-2, j-1, j, j+1\}.$$

By Lemma 2.3, we obtain

$$\frac{4}{i}|T_{j-2}|_{\overline{z_j}} = (z_{j-1} - v_{j-2})\frac{p_{j-2}}{l_{j-1}},$$
$$\frac{4}{i}|T_{j-1}|_{\overline{z_j}} = (v_{j-1} - z_{j-1}) + (v_{j-1} - z_{j+1})\frac{p_{j-1}}{l_j},$$
$$\frac{4}{i}|T_j|_{\overline{z_j}} = (z_{j+1} - v_j) + (z_{j-1} - v_j)\frac{s_j}{l_{j-1}},$$
$$\frac{4}{i}|T_{j+1}|_{\overline{z_j}} = (v_{j+1} - z_{j+1})\frac{s_{j+1}}{l_j}.$$

The six previous equations and (22) yield (28). □

Since the lengths p_j, l_j, s_j are functions of the vertices z_j, (28) is a system of $2n$ real equations in the unknowns z_j, η_j, i.e. $3n$ real unknowns, and so it cannot determine the maximizers. From the outer equivalent triangle property of any maximizer $P^\star$ we have that, for each j, the vector $(v_j - v_{j-1})$ is parallel to $(z_{j+1} - z_{j-1})$. Hence $P^\star$ satisfies also the following system for suitable positive ζ_j:

$$(v_j - v_{j-1}) = \zeta_j(z_{j+1} - z_{j-1}) \quad \text{for } j = 1, \ldots, n. \tag{29}$$

We notice that the triangles v_j, z_j, v_{j-1} and z_{j+1}, z_j, z_{j-1} are similar and then

$$\zeta_j = \frac{|v_j - v_{j-1}|}{|z_{j+1} - z_{j-1}|} = \frac{s_j}{l_{j-1}} = \frac{p_{j-1}}{l_j}. \tag{30}$$

Thus, the outer-ratio property (6) can be rewritten as

$$\frac{(1 + \zeta_{j+2})}{\zeta_{j+1}(1 + \zeta_{j+1} + \zeta_{j+2})} = \frac{(1 + \zeta_{j-1})}{\zeta_j(1 + \zeta_j + \zeta_{j-1})} \quad \text{for } j = 1, \ldots, n. \tag{31}$$

Proof of Theorem 1.5 Since v_j, z_j, and z_{j-1} are collinear, (30) implies that we can express v_j in terms of z_j, z_{j-1} and positive ζ_j

$$v_j = z_j + \zeta_j(z_j - z_{j-1}).$$

Similarly, (30) also implies that

$$v_{j+1} = z_{j+1} + \zeta_{j+1}(z_{j+1} - z_j),$$
$$v_{j-1} = z_j + \zeta_{j-1}(z_j - z_{j+1}),$$
$$v_{j-2} = z_{j-1} + \zeta_{j-1}(z_{j-1} - z_j).$$

This permits us to eliminate in (28) all the vectors v_j, i.e.,

$$(z_{j-1} - z_{j+1}) = \eta_{j-2}\zeta_{j-1}^2(z_j - z_{j-1}) + \eta_{j-1}[(z_j - z_{j-1}) + \zeta_j(z_j - z_{j+1}) + (z_j - z_{j+1})(1 + \zeta_j)\zeta_j]$$

$$+\eta_j[(z_{j+1}-z_j)-\zeta_j(z_j-z_{j-1})+(z_{j-1}-z_j)(1+\zeta_j)\zeta_j]$$
$$+\eta_{j+1}(z_{j+1}-z_j)\zeta_{j+1}^2. \tag{32}$$

Rearranging the terms we obtain

$$(z_{j-1}-z_j)\big(1+\eta_{j-2}\zeta_{j-1}^2+\eta_{j-1}-\eta_j(2\zeta_j+\zeta_j^2)\big)$$
$$+(z_j-z_{j+1})\big(1-\eta_{j-1}(2\zeta_j+\zeta_j^2)+\eta_j+\eta_{j+1}\zeta_{j+1}^2\big)=0.$$

Since $P^\star\in\mathcal{P}_n$, the vertices z_j are distinct and not collinear. This implies that the vectors $(z_{j+1}-z_j)$, (z_j-z_{j-1}) are linearly independent and their coefficients in the previous equation must be zero, i.e.,

$$\begin{cases}\eta_{j-2}\zeta_{j-1}^2+\eta_{j-1}-\eta_j(2\zeta_j+\zeta_j^2)=-1,\\ -\eta_{j-1}(2\zeta_j+\zeta_j^2)+\eta_j+\eta_{j+1}\zeta_{j+1}^2=-1\end{cases} \tag{33}$$

for $j=1,\dots,n$. We shift by one the index in the first equation and adding to and subtracting from the second one, we obtain

$$\begin{cases}\eta_{j-1}\zeta_j-\eta_j+\eta_{j+1}\zeta_{j+1}=1,\\ \eta_{j-1}\zeta_j(1+\zeta_j)-\eta_{j+1}\zeta_{j+1}(1+\zeta_{j+1})=0\end{cases} \tag{34}$$

for $j=1,\dots,n$.

We infer that the ζ_j's are such that the linear system (34) has a solution $\eta_1,\eta_2,\dots,\eta_n$. Taking into account the six equations involving only the five unknowns $\eta_{j-2},\dots,\eta_{j+2}$, we deduce that the determinant of the matrix

$$\begin{pmatrix}\zeta_{j-1} & -1 & \zeta_j & 0 & 0 & 1\\ 0 & \zeta_j & -1 & \zeta_{j+1} & 0 & 1\\ 0 & 0 & \zeta_{j+1} & -1 & \zeta_{j+2} & 1\\ \zeta_{j-1}+\zeta_{j-1}^2 & 0 & -\zeta_j-\zeta_j^2 & 0 & 0 & 0\\ 0 & \zeta_j+\zeta_j^2 & 0 & -\zeta_{j+1}-\zeta_{j+1}^2 & 0 & 0\\ 0 & 0 & \zeta_{j+1}+\zeta_{j+1}^2 & 0 & -\zeta_{j+2}-\zeta_{j+2}^2 & 0\end{pmatrix}$$

has to be zero. After some manipulations we obtain

$$\zeta_{j-1}(1+\zeta_j)(1+\zeta_{j+1})\zeta_{j+2}[\zeta_j(1+\zeta_{j+2})(1+\zeta_j+\zeta_{j-1})$$
$$-\zeta_{j+1}(1+\zeta_{j-1})(1+\zeta_{j+1}+\zeta_{j+2})]=0.$$

Since $\zeta_j>0$, the previous equation can be simplified as in the statement (31) and we conclude the proof. □

3 Circulant Systems

In this section we have collected the steps of the proofs which are not strictly related to the geometry of the problem. We focus on the systems (25), (31) and we prove that the only suitable solutions are those independent of j.

First we deal with the inner problem.

Proposition 3.1 *Any solution of the system of equations*

$$\frac{(\lambda_{j-2}-1)(\lambda_{j-1}-1)}{\lambda_{j-2}} = \frac{(\lambda_{j+1}-1)(\lambda_j-1)}{\lambda_{j+1}},$$

for $j = 1, 2, \ldots, n$, *with* $\lambda_j > 1$ *for all* j, *satisfies* $\lambda_i = \lambda_j$ *for all* i *and* j.

Proof Set $\gamma_j = 1/(\lambda_j - 1)$, for all j. Notice that $\lambda_j > 1$ implies $\gamma_j > 0$. Then (25) reduce to

$$\gamma_{j-1}(1+\gamma_{j-2}) = \gamma_j(1+\gamma_{j+1}). \tag{35}$$

Consider the maximum number among the expressions $\gamma_j(1+\gamma_{j-1})$, for all j. Assume this maximum is attained by $\gamma_M(1+\gamma_{M-1})$. From (35) we obtain

$$\gamma_M(1+\gamma_{M-1}) \geq \gamma_j(1+\gamma_{j+1}),$$

for every $j = 1, 2, \ldots, n$. Now, from $\gamma_M(1+\gamma_{M-1}) \geq \gamma_{M-1}(1+\gamma_M)$ we have

$$\gamma_M \geq \gamma_{M-1}. \tag{36}$$

From $\gamma_M(1+\gamma_{M-1}) \geq \gamma_M(1+\gamma_{M+1})$ we find

$$\gamma_{M-1} \geq \gamma_{M+1}. \tag{37}$$

From (35) we deduce that $\gamma_M(1+\gamma_{M-1}) = \gamma_{M+1}(1+\gamma_{M+2})$ and so we can also write $\gamma_{M+1}(1+\gamma_{M+2}) \geq \gamma_{M+1}(1+\gamma_M)$, which implies that

$$\gamma_{M+2} \geq \gamma_M, \tag{38}$$

and $\gamma_{M+1}(1+\gamma_{M+2}) \geq \gamma_{M+2}(1+\gamma_{M+1})$, which yields

$$\gamma_{M+1} \geq \gamma_{M+2}. \tag{39}$$

By combining inequalities (36–39) we discover $\gamma_{M+1} \geq \gamma_{M+2} \geq \gamma_M \geq \gamma_{M-1} \geq \gamma_{M+1}$ and so $\gamma_{M-1} = \gamma_M = \gamma_{M+1} = \gamma_{M+2}$. Taking into account (35) again yields $\gamma_i = \gamma_j$, for all i and j. □

We now turn to the outer problem.

Proposition 3.2 *Any solution of the system of equations*

$$\frac{1+\zeta_{j+2}}{\zeta_{j+1}(1+\zeta_{j+1}+\zeta_{j+2})} = \frac{1+\zeta_{j-1}}{\zeta_j(1+\zeta_j+\zeta_{j-1})}$$

for $j = 1, 2, \ldots, n$, *with* $\zeta_j > 0$ *for all* j, *satisfies* $\zeta_i = \zeta_j$ *for all* i *and* j.

Proof Set

$$a_j = \frac{\zeta_j}{1+\zeta_j+\zeta_{j+1}}, \qquad b_j = \frac{\zeta_{j+1}}{1+\zeta_j+\zeta_{j+1}}. \tag{40}$$

Notice that $\zeta_j > 0$, for all j, implies that all a_j and b_j are positive and that

$$a_j + b_j < 1. \tag{41}$$

The system (31) can be rewritten in terms of a_j, b_j as

$$\frac{1 - a_{j+1}}{b_j} = \frac{1 - b_{j-1}}{a_j}. \tag{42}$$

Furthermore, it is easy to check that the relation

$$\frac{1 - a_j}{b_j} = \frac{1 - b_{j+1}}{a_{j+1}} \tag{43}$$

also holds for every $j = 1, 2, \ldots, n$.

Now consider the maximum number among the expressions $(1 - a_{j+1})/b_j$ or $(1 - a_j)/b_j$ involved in (42) and (43) and call it C. Assume first that C appears in (42), i.e., $C = \frac{1-a_{M+1}}{b_M}$, for some M. From $\frac{1-a_{M+1}}{b_M} \geq \frac{1-a_M}{b_M}$ we infer

$$a_M \geq a_{M+1}. \tag{44}$$

From

$$\frac{1 - b_{M-1}}{a_M} = \frac{1 - a_{M+1}}{b_M} \geq \frac{1 - b_M}{a_M}$$

we deduce

$$b_M \geq b_{M-1}. \tag{45}$$

From

$$\frac{1 - a_{M+1}}{b_M} = \frac{1 - b_{M-1}}{a_M} \geq \frac{1 - a_M}{b_{M-1}}$$

we deduce $b_{M-1} - b_{M-1}^2 \geq a_M - a_M^2$ and then

$$(b_{M-1} - a_M)(1 - b_{M-1} - a_M) \geq 0.$$

Since $b_{M-1} \leq b_M$ by (45), inequality (41) implies that $(1 - b_{M-1} - a_M) > 0$. Therefore

$$b_{M-1} \geq a_M. \tag{46}$$

Analogously, from

$$\frac{1 - a_{M+1}}{b_M} \geq \frac{1 - a_{M+2}}{b_{M+1}} = \frac{1 - b_M}{a_{M+1}}$$

we infer

$$(a_{M+1} - b_M)(1 - a_{M+1} - b_M) \geq 0.$$

Inequalities (41), (44) ensure that $(1 - a_{M+1} - b_M) > 0$ and so

$$a_{M+1} \geq b_M. \tag{47}$$

Combining inequalities (44–47) we obtain

$$b_{M-1} \geq a_M \geq a_{M+1} \geq b_M \geq b_{M-1}$$

and so

$$b_{M-1} = a_M = a_{M+1} = b_M.$$

Now it is easy to use (42) and (43) to deduce that all a_j and b_j have to be equal, and from this fact and (40) that all the ζ_j's are equal.

Assume now that C appears in (43), i.e., $C = (1 - a_M)/b_M$, for some M. From $(1 - a_M)/b_M \geq (1 - b_M)/a_M$ and (41) we deduce

$$a_M \geq b_M. \tag{48}$$

From $\frac{1-b_{M+1}}{a_{M+1}} = \frac{1-a_M}{b_M} \geq \frac{1-a_{M+1}}{b_{M+1}}$ and (41) we deduce

$$b_{M+1} \geq a_{M+1}. \tag{49}$$

From $\frac{1-b_{M+1}}{a_{M+1}} = \frac{1-a_M}{b_M} \geq \frac{1-b_M}{a_{M+1}}$ we infer

$$b_M \geq b_{M+1}. \tag{50}$$

From $\frac{1-a_M}{b_M} \geq \frac{1-a_{M+1}}{b_M}$ we obtain

$$a_{M+1} \geq a_M. \tag{51}$$

Once again we have obtained a sequence of inequalities that can be satisfied only if $a_M = a_{M+1} = b_M = b_{M+1}$. As in the previous case, it is now easy to conclude that all ζ_j have to be equal. □

G. Ottaviani has analyzed the case $n = 7$ for the inner and outer problem. By using the computer algebra system Macaulay, he found that in the complex variables $\gamma_1, \ldots, \gamma_7$, the system (35) has a solution set which consists of the trivial line $\gamma_1 = \cdots = \gamma_7$ and of an algebraic surface of degree 14. It is interesting that the line and the surface are disjoint components. Similarly, he found that for the outer problem the solution set of the system (31) contains the trivial line and a surface of degree 71.

4 Proofs of Theorems 1.6 and 1.7

In this section we complete the characterization of the maximizers of F and G, thus proving Theorems 1.6, 1.7. To do this we turn back to geometry.

Proof of Theorem 1.6 As noted in Sect. 2.1, each polygon P contained in $\Phi_n \cap \mathcal{F}_n$ satisfies (23–25). By Proposition 3.1 we deduce that there exists a $\lambda > 1$ such that

$$z_{j+2} - z_{j-1} = \lambda(z_{j+1} - z_j), \quad \text{for all } j. \tag{52}$$

Such a condition is satisfied only by affinely regular n-gons. This is proved in [5, Statement 3, Theorem 1], where the result is attributed to Coxeter [3]. We provide a proof for completeness and for the convenience of the reader.

Up to an affine transformation, we can assume that three consecutive edges of P, z_1z_2, z_2z_3 and z_3z_4, say, all have length one. By (52), the quadrilateral $z_1z_2z_3z_4$ is an isosceles trapezium, and so the angles of P at z_2 and z_3 are also equal. Denote by r_2 the bisector of the angle at z_2. By assumption, r_2 is an axis of symmetry of the segment z_1z_3. Since the vertex z_n, consecutive to z_1, is determined by the relation $z_n - z_3 = \lambda(z_1 - z_2)$ and z_4 by $z_4 - z_1 = \lambda(z_3 - z_2)$, it is easily seen that r_2 is an axis of symmetry of the segment z_4z_n. Similarly, z_{n-1} and z_5 are uniquely determined by the previous vertices (and λ) and then r_2 is also an axis of symmetry of the segment z_5z_{n-1}. This argument shows that r_2 is a symmetry axis of P. Analogously, considering the vertex z_3 of the trapezium $z_1z_2z_3z_4$, the bisector r_3 of the angle at z_3 is a symmetry axis of P. This clearly implies that P is a regular polygon. □

Proof of Theorem 1.7 As noted in Sect. 2.2, each polygon P contained in $\Gamma_n \cap \mathcal{G}_n$ satisfies systems (29), (31). By Proposition 3.2 we deduce that there exists a $\zeta > 0$ such that

$$v_j - v_{j-1} = \zeta(z_{j+1} - z_{j-1}), \quad \text{for all } j. \tag{53}$$

This means that the pairs of triangles $v_{j-1}v_jz_j$, $z_{j+1}z_{j-1}z_j$ are all similar with the same ratio of similarity. Hence,

$$v_j - z_j = \zeta(z_j - z_{j-1}), \qquad v_j - z_{j+1} = \zeta(z_{j+1} - z_{j+2}).$$

Therefore the triangles $v_jz_{j+1}z_j$, $v_jz_{j+2}z_{j-1}$ are also similar and

$$z_{j+2} - z_{j-1} = (1 + \zeta)(z_{j+1} - z_j), \quad \text{for all } j. \tag{54}$$

Such a condition is just like (52) and the argument of the previous proof implies that P is an affinely regular polygon. □

Theorems 1.6 and 1.7 were the last ingredients in the proof of Theorem 1.8, whose quantitative form is the following.

Theorem 4.1 *Let $P \in \mathcal{P}_n$, $n \geq 5$, then*

$$\min_{j=1,\dots,n} |T_j(P)| \leq |P| \frac{2\sin^2(\pi/n)}{n\cos(2\pi/n)}, \tag{55}$$

$$\min_{j=1,\dots,n} |W_j(P)| \leq |P| \frac{4\sin^2(\pi/n)}{n}, \tag{56}$$

and equality holds if and only if P is an affinely regular n-gon.

Notice that (55) implies the inequality in [17, Lemma 3].

5 Extensions and Affine Length

Let C be a planar convex body, and let γ be a connected closed proper subset of ∂C, i.e., a closed arc with endpoints a and b. Define $C(\gamma)$ as the intersection of all half-planes supporting C at boundary points in $\partial C \setminus \gamma$. The boundary of $C(\gamma)$ is obtained by extending $\partial C \setminus \gamma$ with the continuations of the half-lines tangent to $\partial C \setminus \gamma$ at a and b. Consider the region $I(C, \gamma) = C(\gamma) \setminus C$ and the area ratio $|I(C, \gamma)|/|C|$. Now let Γ_n be a finite family of connected closed subsets γ_i of ∂C with disjoint interiors:

$$\Gamma_n = \big\{\gamma_1, \ldots, \gamma_n \colon\ \gamma_i \subset \partial C, \operatorname{int}(\gamma_i) \cap \operatorname{int}(\gamma_j) = \emptyset \text{ for } i \neq j,\ i, j = 1, \ldots, n\big\}.$$

Define

$$G(C, \Gamma_n) = \min_{j=1,\ldots,n} \frac{|I(C, \gamma_j)|}{|C|}, \tag{57}$$

and

$$G_n(C) = \max_{\Gamma_n} G(C, \Gamma_n). \tag{58}$$

Theorem 5.1 *For $n > 4$ and for any planar convex body C we have*

$$G_n(C) \leq \max_{P \in \mathcal{P}_n} G(P). \tag{59}$$

Equality holds if and only if C is an affinely regular n-gon.

Proof Let C be a planar convex body and let Γ be a finite family of arcs γ_i on ∂C with endpoints a_i, b_i. We can consider families of arcs which partition ∂C, i.e., such that $a_{i+1} = b_i$ for all i. This can be proved by taking a partition $\Gamma^\star$ with arcs $\gamma_i^\star$ so that $\gamma_i \subset \gamma_i^\star$. Since $I(C, \gamma_i) \subset I(C, \gamma_i^\star)$ we have $G(C, \Gamma) \leq G(C, \Gamma^\star)$.

Let P be the convex polygon whose vertices are the endpoints a_i, b_i of the arcs γ_i of Γ. Clearly, $P \subset C$ and P has exactly n edges. Moreover, if $T_i(P)$ is the outer triangle defined in the introduction corresponding to the edge with vertices a_i, b_i, then $I(C, \gamma_i) \subset T_i(P)$, and

$$G(C, \Gamma) \leq G(P).$$

This immediately yields (59) and the equality conditions. □

Corollary 5.2 *Let K and K' be two planar convex bodies such that their symmetric difference has $n > 4$ connected components $C_1, \ldots, C_n$. Then*

$$\min_{i=1,\ldots,n} |C_i| \leq |K \cap K'| \max_{P \in \mathcal{P}_n} G(P). \tag{60}$$

Equality holds if and only if $K \cap K'$ is an affinely regular polygon with $n = 2m$ edges and, up to an affine transformation, K and K' are two congruent regular m-gons, and K' is K rotated by π/m about its center.

Proof We consider the family Γ of the closed arcs

$$\gamma_i = \partial C_i \cap \partial(K \cap K').$$

Since

$$|C_i| \leq |I(K \cap K', \gamma_i)|,$$

we can apply the previous theorem to $C = K \cap K'$. Equality holds if and only if C is an affinely regular $2m$-gon and so K and K' have to be affinely regular m-gons. □

We present another possible extension of the previous area functionals. Let C be a convex body. Let D be a convex body containing C and such that $D \setminus C$ consists of at least m connected components $D_1, \ldots, D_m$. Let

$$H_m(C, D) = \min_{i=1,\ldots,m} |D_i|.$$

Arguing as above, when we look for the maximizers of $H_m(C, D)|C|^{-1}$ it is easy to see that one can assume that C and D are m-gons. The number $\#(D \setminus C)$ of the connected components of $D \setminus C$ is m, and hence the edges of D support C at its vertices. We get the following result.

Theorem 5.3 *For $m \geq 3$ and for any two planar convex bodies C, D with $C \subset D$, $\#(D \setminus C) = m$, we have*

$$\frac{H_m(C, D)}{|C|} \leq \frac{1}{m} \tan^2 \frac{\pi}{m}. \tag{61}$$

Equality holds if and only if C, D are affinely regular m-gons and the vertices of C are the midpoints of the edges of D.

Proof Standard compactness arguments show that $H_m(C, D)|C|^{-1}$ has a maximum. As already stated in advance, we can assume that such a maximum is attained when C and D are two m-gons P and S, $P \subset S$. Let z_j be the vertices of P and v_j the vertices of S labeled so that z_j belongs to the segment $v_j v_{j-1}$.

First we prove that all triangles S_j with vertices z_j, v_j, z_{j+1} have equal area. Indeed, if $|S_k| > \min_{j=1,\ldots,m} |S_j|$, then a small counterclockwise rotation of the side of S through z_k around z_k decreases the area of S_k and increases the area of S_{k-1}. Possible iterations of this procedure to different edges of S would permit to increase the value of our functional. The assumption about the maximality of the pair (P, S) implies

$$|S_k| = |S_{k-1}| \quad \text{for } k = 1, \ldots, n. \tag{62}$$

Arguing as in the proof of Proposition 2.4, we consider the Lagrange multipliers system relative to $|P|$ as a function of z_j, v_j, under the constraints $|S_j| \geq const$, $z_j \in v_{j-1} v_j$. Denote by μ_j the parameters corresponding to the constraints $|S_j| \geq const$, and by ν_j that corresponding to $0 = \det(z_j - v_{j-1}, z_j - v_j) \equiv A_j$. Hence, we

obtain the system

$$\frac{\partial |P|}{\partial \overline{z_j}} = \sum_{k=1}^{m} \mu_k \frac{\partial |S_k|}{\partial \overline{z_j}} + \sum_{k=1}^{m} \nu_k \frac{\partial A_k}{\partial \overline{z_j}} \quad \text{for } j = 1, \ldots, n,$$

$$0 = \frac{\partial |P|}{\partial \overline{v_j}} = \sum_{k=1}^{m} \mu_k \frac{\partial |S_k|}{\partial \overline{v_j}} + \sum_{k=1}^{m} \nu_k \frac{\partial A_k}{\partial \overline{v_j}} \quad \text{for } j = 1, \ldots, n.$$

Since we are looking for minimizers of $|P|$ with constraints $|S_j| \geq const$, we have

$$\mu_j \geq 0 \quad \text{for } j = 1, \ldots, n. \tag{63}$$

Taking into account the results in Sect. 2 we get:

$$(z_{j-1} - z_{j+1}) = \mu_{j-1}(v_{j-1} - z_{j-1}) + \mu_j(z_{j+1} - v_j) + \nu_j(v_{j-1} - v_j), \tag{64}$$

$$0 = \mu_j(z_j - z_{j+1}) + \nu_j(z_j - v_{j-1}) + \nu_{j+1}(v_{j+1} - z_{j+1}). \tag{65}$$

The constraint $A_j = 0$ can be written as $z_j = v_{j-1} + \rho_j(v_j - v_{j-1})$, where the variables ρ_j satisfy $0 < \rho_j < 1$. Consequently,

$$|S_j| = \left|(z_{j+1} - v_j) \times (z_j - v_j)\right| = \rho_{j+1}(1 - \rho_j)|U_j|,$$

where U_j denotes the triangle $v_{j-1}v_jv_{j+1}$. Identity (62) yields

$$\rho_{j+1}(1 - \rho_j)|U_j| = \rho_j(1 - \rho_{j-1})|U_{j-1}|. \tag{66}$$

Equations (64) and (65) become

$$\begin{aligned}(1 + \mu_{j-1})(1 - \rho_{j-1})(v_{j-2} - v_{j-1}) + (1 + \mu_j)\rho_{j+1}(v_j - v_{j+1}) \\ + (v_{j-1} - v_j)(1 - \nu_j) = 0,\end{aligned} \tag{67}$$

$$\begin{aligned}0 = (v_j - v_{j-1})(\rho_j(\mu_j + \nu_j) - \mu_j) \\ + (v_{j+1} - v_j)(-\rho_{j+1}(\mu_j + \nu_{j+1}) + \nu_{j+1}).\end{aligned} \tag{68}$$

Since $(v_j - v_{j-1})$ and $(v_{j+1} - v_j)$ are linearly independent, we obtain

$$\mu_j(1 - \rho_j) = \rho_j\nu_j, \tag{69}$$

$$\mu_j\rho_{j+1} = (1 - \rho_{j+1})\nu_{j+1}. \tag{70}$$

Taking the cross product of (67) with the vector $(v_j - v_{j-1})$ yields

$$(1 + \mu_{j-1})(1 - \rho_{j-1})U_{j-1} - (1 + \mu_j)\rho_{j+1}U_j = 0.$$

By means of (66) the latter reduces to

$$(1 + \mu_j)\rho_j = (1 + \mu_{j-1})(1 - \rho_j). \tag{71}$$

The systems (69), (70), (71) are $3n$ real equations in $3n$ real unknowns μ_j, ρ_j, ν_j. From the first two equations we get

$$\mu_j(1-\rho_j)^2 = \mu_{j-1}\rho_j^2. \tag{72}$$

A comparison with (71) implies

$$(1+\mu_j)^2\mu_j = (1+\mu_{j-1})^2\mu_{j-1}. \tag{73}$$

Since the function $y = (1+x)^2x$ is injective for $y > 0$, inequality (63) implies that $\mu_j = \mu$, for all j.

From (71) we deduce $\rho_j = 1/2$ and then, from (69), that $\nu_j = \mu$, for all j. Hence, (67) gives

$$(1+\mu)v_{j-2} - 2\mu v_{j-1} + 2\mu v_j - (1+\mu)v_{j+1} = 0,$$

which implies that $v_{j-2}v_{j+1}$ is parallel to $v_{j-1}v_j$ and their ratio is independent of j.

As in the proof of Theorem 1.6 (see (52)), we deduce that S is affinely regular. □

Given a planar convex body K and a natural number $n \geq 3$, we denote by $\mathcal{P}_n^{\mathrm{i}}(K)$ and $\mathcal{P}_n^{\mathrm{c}}(K)$ the class of n-gons inscribed in K or with K inscribed in them, respectively. For any n distinct points $z_i \in \partial K$ we consider the polygon $P = \mathrm{conv}\{z_1, \ldots, z_n\} \in \mathcal{P}_n^{\mathrm{i}}(K)$ and lines s_i supporting K at the vertices z_i of P. Let $S_i(K)$ be the triangle bounded by the lines s_{i-1}, s_i and the side at $z_{i-1}z_i$, and S be the polygon bounded by the lines s_i. Hence, S is an n-gon in $\mathcal{P}_n^{\mathrm{c}}(K)$. This is equivalent to choosing two n-gons P, S, with $P \in \mathcal{P}_n^{\mathrm{i}}(K)$, $S \in \mathcal{P}_n^{\mathrm{c}}(K) \cap \mathcal{P}_n^{\mathrm{c}}(P)$. In this case, for brevity, we say that the pair (P, S) belongs to $\mathcal{P}_n^{\mathrm{i,c}}(K)$. Define

$$AL_n(K) = 2n \max\Big\{ \min_{j=1,\ldots,n} |S_j(K)|^{\frac{1}{3}} \colon (P,S) \in \mathcal{P}_n^{\mathrm{i,c}}(K) \Big\}.$$

Since $n \geq 3$, $AL_n(K)$ is finite for every K. The symbol we chose to denote these functionals is justified by some properties listed below, which present AL_n as a discretization of the affine length of the boundary of K. Recalling the functional $H_n(P, S)$ introduced above, we have

$$AL_n(K) = 2n \max\big\{ H_n(P,S)^{\frac{1}{3}} \colon (P,S) \in \mathcal{P}_n^{\mathrm{i,c}}(K) \big\}.$$

As in the proof of Theorem 5.3 it turns out that the previous maximum is attained when all the $S_i(K)$ have the same area. In the sequel we denote by $\mathcal{E}_n$ the class of pairs of n-gons with the property that all the $S_i(K)$ have equal area.

Proposition 5.4 *For every planar convex body K and $n \geq 3$,*

$$AL_n(K) = 2n \max\big\{ H_n(P,S)^{\frac{1}{3}} \colon (P,S) \in \mathcal{P}_n^{\mathrm{i,c}}(K) \cap \mathcal{E}_n \big\}.$$

Proof Clearly, $|S_j(K)|$ is a continuous function of z_i and s_i for all i and j. If z_i moves towards z_{i+1}, then $|S_{i+1}(P)|$ decreases and $|S_i(P)|$ increases, unless one or both of them remain equal to zero. Furthermore, if s_i rotates around z_i counterclockwise, then

$|S_{i+1}(P)|$ decreases and $|S_i(P)|$ increases. These facts easily imply that, if (P, S) does not belong to $\mathcal{E}_n$, then we can move the supporting lines s_i together with the points z_i increasing all $|S_j(K)|$. □

By Proposition 5.4 the functional AL_n is then related to the arithmetic average of the cube root of $|S_j(K)|$, i.e.,

$$AL_n(K) = 2 \max_{(P,S)\in\mathcal{P}_n^{\mathrm{i,c}}(K)\cap\mathcal{E}_n} \sum_{j=1}^{n} |S_j(K)|^{\frac{1}{3}}.$$

This further clarifies the connection with the affine length of K, $\Omega_1(K)$, as defined, for example, by Ludwig [18, Sect. 3]. There Ludwig proved that all upper (or lower) semicontinuous and equi-affine invariant valuations on the space of planar compact convex sets (endowed with the Hausdorff metric) are linear combinations of three basic valuations: the Euler characteristic, the area, and the affine length.

Such a characterization is the main ingredient in the following proof.

Proposition 5.5 *If K is a planar convex body, then*

$$\inf_{n\in\mathbb{N}} AL_n(K) = \Omega_1(K).$$

Proof We first notice that $AL_n(K) = 0$ when K is a polygon with less than n vertices. Moreover, these are the only convex sets where AL_n vanishes.

Set $AL(K) = \inf_n AL_n(K)$. Since the functionals AL_n are continuous and equi-affine invariant, it is easy to verify that AL is equi-affine invariant and upper semicontinuous. Following the arguments used by Ludwig [18, Theorem 2], it can be proved that AL is a valuation, i.e.,

$$AL(H \cup K) + AL(H \cap K) = AL(H) + AL(K),$$

for every pair of convex bodies such that $H \cup K$ is convex. By [18, Theorem 1], it follows that AL is a linear combination of the affine length, the area, and the Euler characteristic. Since $AL_n(P)$ vanishes when P is a polygon with less than n vertices, AL has to be a multiple of the affine length. A simple calculation of AL at the unit disc yields the result. □

We now present some straightforward consequences of Theorem 5.3.

Theorem 5.6 *If K is a planar convex body and $n \geq 3$, then*

$$AL_n(K) \leq 2\left(n \tan\frac{\pi}{n}\right)^{\frac{2}{3}} \max_{P\in\mathcal{P}_n^{\mathrm{i}}(K)} |P|^{\frac{1}{3}},$$

and equality holds if and only if exists an affinely regular n-gon of maximal area in $\mathcal{P}_n^{\mathrm{i}}(K)$.

Corollary 5.7 *If K is a planar convex body and $n \geq 3$, then*

$$AL_n(K) \leq 2\left(n \tan \frac{\pi}{n}\right)^{\frac{2}{3}} |K|^{\frac{1}{3}},$$

and equality holds if and only if K is an affinely regular n-gon.

Corollary 5.7 and Proposition 5.5 yield the affine isoperimetric inequality

$$\Omega_1(K) \leq 2\pi^{\frac{2}{3}} |K|^{\frac{1}{3}},$$

without the equality conditions, which are known to characterize ellipses. Notice that equality in the formula of Theorem 5.6 implies that K has at least n points of intersection with a suitable ellipse, and so its Hausdorff distance from that ellipse decreases as n increases.

It is well known that the affine length of a planar convex body K is closely related to the approximation of K by polygons. We refer the interested reader to [9] for an extensive review. Theorem 5.6 implies that, for any planar convex body K and $n \geq 3$, there exists a polygon $P \in \mathcal{P}_n^{\mathrm{i}}(K)$ such that

$$|P| \geq \frac{AL_n(K)^3}{8n^2 \tan^2(\frac{\pi}{n})}.$$

Proposition 5.5 allows $AL_n(K)$ to be replaced with $\Omega_1(K)$ in the previous inequality. This yields a sharp lower bound for the approximation of K with polygons from $\mathcal{P}_n^{\mathrm{i}}(K)$, which also follows from the affine isoperimetric inequality and a result of Blaschke (see [9, Sect. 4]).

6 Applications to Tomography

The word tomography reminds people of the medical CAT scanner, where images are reconstructed from X-rays. Despite the common use of CAT scanners, the mathematical subject related to this reconstruction still deserves interesting and unsolved problems. One of them is stated in the question asked by P.C. Hammer in 1963: *How many parallel X-ray pictures of a convex body must be taken in order to permit its exact reconstruction?* The parallel X-ray of a planar convex body K in a direction θ provides the length of each chord of K parallel to θ.

The existence of finite sets of directions, with arbitrary large cardinality, such that the corresponding X-rays cannot determine a convex body among the others is shown with the following well-known example (see [8]).

Consider a regular n-gon Q centered at a fixed point o, and its rotation Q' by π/n about o. Let θ be a direction parallel to one of the edges of the convex hull of Q and Q'. It is easy to see that Q and Q' have the same parallel X-rays in the direction θ. This example arises in many papers, mainly related to Geometric and Discrete Tomography. We refer the interested reader to [7, 13].

Gardner and McMullen proved in [6] that *convex bodies are determined by X-rays taken in any set of directions that is not a subset of the directions of the edges of an*

affinely regular polygon. Since the cross ratio of any four directions of the edges of a regular polygon is an algebraic number, any set of four directions with a transcendental cross ratio uniquely determines a convex body by means of the corresponding X-rays.

Volčič [21] and Longinetti [16] show that the reconstruction of H is well posed when the set of directions guarantees uniqueness. Roughly speaking, if we know all the X-rays of H in such directions, and these X-rays contain errors ε as small as we want, then the corresponding reconstructions H_ε converge to H when ε goes to zero.

Now, consider the case when a finite number n of X-rays of H are exactly known, but the directions are determined up to an error δ. The error δ has to be small enough to distinguish the given n directions among them. For any positive δ, we cannot distinguish the set S of the given directions from the sets of non-uniqueness in the Gardner–McMullen theorem. Therefore, the results of well-posedness proved by Volčič cannot be used here. In [14] the following result is proved:

$$|K \triangle K'| \leq l^2 (8n)^{-1} \tan \frac{\pi}{n}, \tag{74}$$

where l is the length of the boundary of $K \cap K'$ and $K \triangle K'$ is the symmetric difference. Inequality (74) can be seen as a stability result and is optimal not only in the order but also in the constant, since equality holds if and only if the n directions, K and K' are chosen as in the above example. Inequality (74) is not affine invariant, while Hammer's problem is. An affine-invariant inequality in which l^2 is replaced by $|K \cap K'|$ is proved in [15] for sets of three directions. We show here that (55) can be used to generalize this result.

Let R be a connected component of $K \triangle K'$. For every direction $\theta \in S$, let θR be the connected component of $K \triangle K'$ different from R with the same X-ray of R in the direction θ. Let

$$W(R) = \bigcup_{h \in \mathbb{N},\, \theta_{i_j} \in S} \theta_{i_h} \cdots \theta_{i_1} R.$$

In [14, Proposition 2] (see also [7, Lemma 1.2.6]) it is proved that $W(R)$ consists of a finite number h of components and they are at least $2n$.

Suppose that $K \triangle K' = W(R)$. Since all components in $W(R)$ have the same area, Corollary 5.2 yields

$$\frac{|W(R)|}{|K \cap K'|} \leq h \cdot \max_{P \in \mathcal{P}_h} G(P).$$

From (55) we obtain the explicit bound $2 \sin^2(\pi/h)/\cos(2\pi/h)$, a decreasing function of h. Since $h \geq 2n$, we get the following result.

Theorem 6.1 *If K and K' are two planar convex bodies with the same X-rays in n different directions ($n \geq 3$), and such that $K \triangle K'$ consists of a finite number of connected components of equal area, then*

$$|K \triangle K'| \leq |K \cap K'| \frac{1 - \cos(\pi/n)}{\cos(\pi/n)}. \tag{75}$$

Equality holds if and only if, up to an affine transformation, the directions are equally spaced, K and K' are congruent regular n-gons, and K' is K rotated by π/n about its center.

We remark that (75) is stronger than (74), via the classical isoperimetric inequality for n-gons.

In a forthcoming paper by Dulio, Longinetti, Peri and Venturi [4], Theorem 6.1 will be presented without the assumption on the connected components of $K \triangle K'$. There, also an improvement of (75) is given by using more information about the set of directions, as, for example, the cross ratio of four directions.

References

1. Barlotti, A.: Una proprietà degli n-agoni che si ottengono trasformando in una affinità un n-agono regolare. Boll. Unione Mat. Ital. **510**(3), 96–98 (1955)
2. Bianchi, G., Longinetti, M.: Reconstructing plane sets from projections. Discrete Comput. Geom. **5**, 223–242 (1990)
3. Coxeter, H.S.M.: Affinely regular polygons. Abh. Math. Sem. Univ. Hamburg **62**, 249–253 (1992)
4. Dulio, P., Longinetti, M., Peri, C., Venturi, A.: Sharp affine stability estimates for Hammer's problems. Adv. Appl. Math. (2007). doi:10.1016/j.amm.2007.06.001
5. Fisher, J.C., Jamison, R.E.: Properties of affinely regular polygons. Geom. Dedicata **69**, 241–259 (1998)
6. Gardner, R.J., McMullen, P.: On Hammer's X-ray problem. J. Lond. Math. Soc. **21**, 171–175 (1980)
7. Gardner, R.J.: Geometric Tomography. Cambridge University Press, New York (1995)
8. Giering, O.: Bestimmung von Eibeireichen und Eikörpen durch Steiner-Symmetrisierungen. Sber. Bayer. Akad. Wiss. München, Math.-Nat. Kl., 225–253 (1962)
9. Gruber, P.M.: Aspects of approximation of convex bodies. In: Gruber, P.M., Wills, J.M. (eds.) Handbook of Convex Geometry, pp. 319–345. North-Holland, Amsterdam (1993)
10. Hammer, P.C.: Problem 2. In: Proc. of Symposium in Pure Mathematics. Convexity, vol. VII. American Mathematical Society, Providence (1963)
11. Harel, G., Rabin, J.M.: Polygons whose vertex triangles have equal area. Am. Math. Mon. **110**, 606–619 (2003)
12. John, F.: Extremum problems with inequalities as subsidary conditions. In: Courant Anniversary Volume, pp. 187–204. Interscience, New York (1948)
13. Kuba, A., Herman, G.T.: Discrete tomography: a historical overview. In: Herman, G.T., Kuba, A. (eds.) Discrete Tomography, pp. 3–34. Birkhäuser, Boston (1999)
14. Longinetti, M.: An isoperimeric inequality for convex polygons and convex sets with the same symmetrals. Geom. Dedicata **20**, 27–41 (1986)
15. Longinetti, M.: Una proprietà di massimo dei poligoni affinemente regolari. Rend. Circ. Mat. Palermo **24**, 448–459 (1985)
16. Longinetti, M.: Some questions of stability in the reconstruction of plane convex bodies from projections. Inverse Probl. **1**, 87–97 (1985)
17. Lopez, M.A., Reisner, S.: Efficient approximation of convex polygons. Int. J. Comput. Geom. Appl. **10**, 445–452 (2000)
18. Ludwig, M.: A characterization of affine length and asymptotic approximation of convex discs. Abh. Math. Semin. Univ. Hamburg **169**, 75–88 (1999)
19. Reisner, S., Schütt, C., Werner, E.: Dropping a vertex or a facet from a convex polytope. Forum Math. **13**, 359–378 (2001)
20. Renyi, A., Sulanke, R.: Über die konvexe Hülle von n zufällig gewählten Punkten. Z. Wahrscheinlichkeitsth. Verw. Geb. **2**, 75–84 (1963)
21. Volčič, A.: Well-posedness of the Gardner–McMullen reconstruction problem. In: Proceedings of Conference on Measure Theory, Oberwolfach, 1983. Lecture Notes in Mathematics, vol. 1089, pp. 199–210. Springer, Berlin (1984)

Improved Output-Sensitive Snap Rounding

John Hershberger

Abstract This paper presents new algorithms for snap rounding an arrangement $\mathcal{A}$ of line segments in the plane. Snap rounding defines a set of *hot pixels*, which are unit squares centered on the integer grid points closest to the vertices of $\mathcal{A}$. Snap rounding simplifies $\mathcal{A}$ by replacing every input segment by a piecewise linear curve connecting the centers of the hot pixels the segment intersects. Let $\mathcal{H}$ be the set of all hot pixels, and for each $h \in \mathcal{H}$ let $is(h)$ be the number of segments with an intersection or endpoint inside h. If $\mathcal{A}$ contains n input segments, the running time of the first new algorithm is $O(\sum_{h\in\mathcal{H}} is(h) \log n)$. This improves previous input- and output-sensitive algorithms by a factor of $\Theta(n)$ in the worst case. The second algorithm has an even better running time of $O(\sum_{h\in\mathcal{H}} ed(h) \log n)$; here $ed(h)$ is the description complexity of the crossing pattern in h, which may be substantially less than $is(h)$ and is never greater.

Keywords Snap rounding · Robust geometric computation

1 Introduction

Arrangements of line segments in the plane are a central tool in computational geometry [6, 19]. They are often used as building blocks in more complex algorithms, and so the arrangement vertices induced by intersections of the line segments may be used as the basis of further computation. This may lead to robustness difficulties. If two segments are represented with a certain finite precision, approximately double that precision (twice as many bits) will be required to represent the intersection of the segments accurately. If this precision-doubling cascades through several levels of algorithmic building blocks, accurately representing the results of a computation

J. Hershberger (✉)
Mentor Graphics Corp., 8005 SW Boeckman Road, Wilsonville, OR 97070, USA
e-mail: john_hershberger@mentor.com

may require more precision than the machine arithmetic provides, forcing the use of a much slower software arithmetic implementation.

One approach to avoiding high-precision geometric computation is to *round* the vertices of the arrangement to some grid, which by suitable scaling one may take to be the integer grid. If this is done naïvely, it may lead to topological inconsistencies between the rounded and unrounded arrangements. A particularly successful approach to rounding an arrangement is the *snap rounding* scheme introduced by Greene [10] and Hobby [15], which is defined as follows: The plane is divided into unit square *pixels* centered on the integer grid points. Every pixel that contains a segment endpoint or an intersection of two segments is declared to be *hot*. Each segment is replaced by a polygonal path joining the centers of the hot pixels it intersects, in the order of intersection. The union of all the rounded segments defines the rounded arrangement. The rounded arrangement can be regarded as a graph $\mathcal{G} = (\mathcal{H}, \mathcal{E})$, where the nodes are identified with the set of hot pixels $\mathcal{H}$, and the *arcs* of the graph link nodes whose hot pixels are joined by rounded segments.

Each arc of $\mathcal{G}$ may correspond to multiple unrounded segments that in the original arrangement pass in parallel from one hot pixel to the next. Depending on the application of the snap rounded arrangement, it may be important to know the set of original segments associated with each arc of the rounded arrangement.

Guibas and Marimont [11] have shown that snap rounding has many desirable properties: every arc of the rounded arrangement has integer grid points as endpoints, every rounded segment is within half a pixel distance of the corresponding unrounded segment, and the rounded and unrounded arrangements are "topologically equivalent up to the collapsing of features." (As noted by Halperin and Packer [14], rounded arcs may still pass very near hot pixel centers; recent work by Packer [18] shows how to avoid this near-degeneracy while preserving the approximation guarantees of snap rounding.)

This paper focuses on methods to compute a snap rounded arrangement efficiently. The running times of these algorithms depend on several parameters: I is the number of pairwise intersections among all the input segments, $\mathcal{H}$ is the set of hot pixels, $|\mathcal{H}|$ is its size, and for any $h \in \mathcal{H}$, the number of original segments that intersect h is $|h|$. The size of $\mathcal{G}$, not counting edge multiplicities, is $O(|\mathcal{H}|)$, because it is a planar graph. The algorithm of Hobby [15] runs in time $O((n + I)\log n + \sum_{h\in\mathcal{H}} |h|)$. The algorithm of Guibas and Marimont [11] runs in time $O(n \log n + I + \sum_{h\in\mathcal{H}} |h| \log |h|)$, plus another term of usually-smaller magnitude that is specific to their approach. Goodrich et al. [9] avoid the dependence on the input size I with an algorithm that runs in time $O((n + \sum_{h\in\mathcal{H}} |h|)\log n)$; they claim optimality to within a logarithmic factor for their algorithm. However, as Halperin and co-authors have noted [5, 13, 14], the definition of $\sum_{h\in\mathcal{H}} |h|$ as the output size is misleading, because it can be as large as $\Theta(n^3)$, even though the size of the snap rounded arrangement, $\Theta(|\mathcal{H}|)$, can never be larger than $O(n^2)$. See Fig. 1. The discrepancy arises because $\sum_{h\in\mathcal{H}} |h|$ counts each arc of $\mathcal{G}$ with its multiplicity (the number of original segments that round to it). An algorithm of de Berg, Halperin, and Overmars [5] runs in time $O((n + I)\log n)$ and produces $\mathcal{G}$ without representing the segments associated with each arc explicitly. Unfortunately, that algorithm performs poorly on the star arrangement shown in Fig. 2; it runs in $O(n^2 \log n)$ time, whereas the algorithm of Goodrich et al. [9] needs only $O(n \log n)$ time.

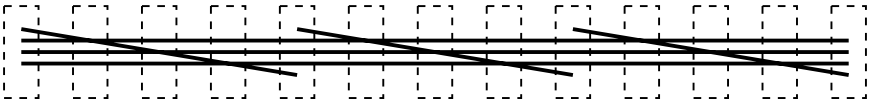
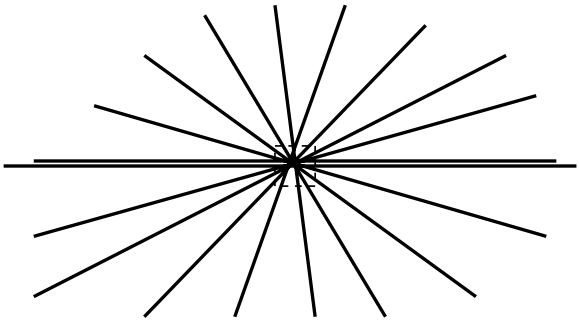


Fig. 1 The algorithm of [9] runs in $\Theta(n^3 \log n)$ on this arrangement because $\sum_{h \in \mathcal{H}} |h| = \Theta(n^3)$, even though $|\mathcal{G}|$ is only $O(n^2)$. (The arrangement is stretched vertically for clarity)

Fig. 2 The algorithm of [5] runs in $\Theta(n^2 \log n)$ on this arrangement, even though $|\mathcal{G}|$ is only $O(n)$, because $I = \Theta(n^2)$



The output an algorithm is required to produce strongly affects its minimum running time. If an algorithm is required to report every segment's intersections with all hot pixels, then the algorithm of [9] is within a logarithmic factor of optimal. Indeed, this is the measure used in that paper's claim of optimality. However, if only the embedded planar graph $\mathcal{G}$ is required, then [9] is far from optimal, as is [5]. See Figs. 1 and 2. Intermediate between these is the goal of producing $\mathcal{G}$ in a data structure that allows a client to report the set of segments associated with any arc of $\mathcal{G}$ efficiently. A further possible requirement on an algorithm is that the set should be spatially ordered and support spatial searches within the set (since no segments in the set intersect between the two hot pixel endpoints of the arc, this is possible).

This paper improves the running time of all previous algorithms. It presents two algorithms, whose running times depend on two new characterizations of the segments incident to a hot pixel. The *intersecting segment count* $is(h)$ is the number of segments that have an endpoint or an intersection inside h. The algorithm of Sect. 3 is very simple to describe and runs in time $O(\sum_{h \in \mathcal{H}} is(h) \log n)$. In particular, it runs in time $O(n^2 \log n)$ and $O(n \log n)$, resp., on the examples of Figs. 1 and 2. The algorithm of Sect. 4 depends on the *edit distance* $ed(h)$, which represents the complexity of the crossover between the segments passing through a hot pixel h. It is always less than $O(is(h))$, and sometimes significantly so. The algorithm of Sect. 4 is more complex than that of Sect. 3, but it runs in $O(\sum_{h \in \mathcal{H}} ed(h) \log n)$ time. Figure 5 in Sect. 4 shows an example for which the two algorithms differ in running time by a factor of $\Theta(\sqrt{n})$.

Both of the algorithms produce the snap rounded arrangement in a form that allows reporting of the original segments associated with each rounded arc. Each arc points to a data structure that records the full set of original segments that round to it, correctly ordered according to the original segment positions; the data structure supports point location among the segments corresponding to each arc. The segment sets are represented using persistent data structures [7], which allow different sets to share

portions of their representations. Although the total size of the sets is $\Theta(\sum_{h\in\mathcal{H}}|h|)$—as large as $\Theta(n^3)$ in some cases—the space needed to represent them is no greater than the algorithm's running time, which is at most $O(n^2 \log n)$.

For convenience the algorithms are described assuming exact arithmetic. The necessary primitives can be implemented using fixed precision as in Hobby's algorithm [15].

2 Preliminaries

The input to a snap rounding algorithm is a collection S of n *ursegments* ("ur" can be taken to refer either to "**u**n**r**ounded" or to the German for "original"). The arrangement of the ursegments, denoted $\mathcal{A}$, has complexity $|\mathcal{A}|$ proportional to n plus the number of intersections between ursegments of S.

A snap rounded arrangement is an embedded planar graph $\mathcal{G} = (\mathcal{H}, \mathcal{E})$, where the nodes are identified with the *hot pixels*, a set of unit squares centered on integer grid points. Each pixel is closed at the left and bottom, and open at the top and right. The hot pixels are exactly those pixels that contain a vertex of $\mathcal{A}$. Each ursegment $s \in S$ is snap rounded to a polygonal path that connects the centers of the hot pixels s intersects, in the order of intersection. There is an arc in $\mathcal{G}$ between two hot pixels $h_1, h_2 \in \mathcal{H}$ if and only if there is an ursegment $s \in S$ whose snap rounded polygonal path visits h_1 and h_2 consecutively.

Note that many ursegments of S may map to the same arc of $\mathcal{G}$. The set of ursegments associated with an arc e is denoted by $segs(e)$. No ursegment crossings or endpoints occur outside hot pixels, so the members of $segs(e)$ can be ordered, which may be important for some applications.

A *sweepline* is a data structure that maintains the intersections of an arrangement of segments with a vertical line as the line sweeps over the arrangement from left to right [1, 2, 6]. The sweepline stores a collection of *active segments* in their order of intersection with the vertical line. Each active segment has a *next event*, which is either its right endpoint or its intersection with the segment below it in the sweepline, if that intersection exists to the right of the sweepline. The segments of a sweepline are stored at the leaves of a balanced binary tree, such as a red-black tree [3]. Internal nodes of the tree implement a min-queue on the x-coordinates of the next events in their subtrees: if a node v has children u and w, the value v stores is the minimum of the values stored at u and w. The sweepline tree supports insert, delete, search, split, and concatenate operations in $O(\log n)$ time.

3 Crossing-Segment Sensitivity

This section presents a simple sweepline algorithm for computing $\mathcal{G}$ that is an order of magnitude faster than all previous algorithms, at least on worst-case inputs. The new algorithm can be viewed as a modification of the sweepline algorithm of Goodrich et al. [9]. That earlier algorithm sweeps a vertical line over $\mathcal{A}$, looking for ursegment endpoints or intersections. Whenever it detects one of these *critical points*, it creates a new hot pixel h and performs surgery on the ursegments of S. Every ursegment s

that intersects h is cut into fragments at its crossings with the boundary of h, and the fragment of s inside h is deleted. Four new unit-length segments are inserted on the boundary of h. The fragments of ursegments outside hot pixels terminate on these newly added boundary segments, and all the vertices of $\mathcal{A}$ (all the critical points) are removed by the surgery. The complexity of the modified arrangement is proportional to the number of intersections of ursegments of S with hot pixels in $\mathcal{H}$, and it can be computed using time only a logarithmic factor greater. It is straightforward to extract $\mathcal{G}$ from the modified arrangement in linear time.

Although the algorithm of Goodrich et al. avoids processing potentially costly ursegment intersections by erasing the part of the arrangement inside the hot pixels of $\mathcal{H}$, it also erases parts of ursegments that are intersection-free, leading to the problem illustrated in Fig. 1. A simple remedy suggests itself: Why not erase only the ursegments that are known to have intersections? The new algorithm develops that simple idea. The algorithm computes $\mathcal{G}$ in two phases: it first computes the set of hot pixels $\mathcal{H}$, then computes the arcs that join the hot pixels.

3.1 Computing the Hot Pixels

As in the algorithm of Goodrich et al., the basis of the crossing-sensitive computation of $\mathcal{H}$ is a Bentley-Ottmann sweep [1, 2, 6] over $\mathcal{A}$. The algorithm assumes no ursegment is vertical, although this can be enforced if necessary by an infinitesimal symbolic rotation of vertical ursegments. A vertical sweepline passes over the ursegments of S, and the algorithm maintains a sorted list of the ursegments intersecting the sweepline in vertical order. A priority queue maintains the next event to the right of the sweepline—as noted in Sect. 2, the sweepline data structure itself (a balanced binary tree) serves as a priority queue for the next event involving an ursegment in the current active set. Events have four types: ursegment left and right endpoints, ursegment intersections, and ursegment *re-insertions*. The first three are standard, but the fourth is a feature of the algorithm: an ursegment that has an intersection inside a hot pixel h is removed from the sweepline and scheduled for re-insertion at the point where it crosses out of h. In essence, this modifies the set of segments swept over by the sweepline so that the fragments derived from any ursegment have at most one intersection per hot pixel.

The algorithm uses a subroutine $trim(s, h)$ that operates on an ursegment s and a hot pixel h it intersects. On entry to $trim(s, h)$, ursegment s is present in the sweepline. The subroutine removes s from the sweepline, then computes the intersections of s with the boundary of h. If s has a fragment that lies to the right of its intersection with the interior of h, then $trim(s, h)$ schedules s for re-insertion into the sweepline at the left endpoint of that fragment. Recall that the bottom boundary of h is contained in h (because h is closed on its left and bottom sides). Thus if s exits h through the bottom, the re-insertion happens infinitesimally after the crossing, so that s is re-inserted *after* it leaves h.

Here is the algorithm to compute the hot pixel set $\mathcal{H}$. For purposes of this algorithm, a hot pixel is represented as an (x, y) pair of integers denoting the center of the pixel, and the algorithm stores pixels in a set *HPSet*. The algorithm obtains integers from the real coordinates of intersections and endpoints using a function $round(r) \equiv \lfloor r + \frac{1}{2} \rfloor$. Thus if $\bar{r} = round(r)$, $r \in [\bar{r} - \frac{1}{2}, \bar{r} + \frac{1}{2})$.

Algorithm FINDHOTPIXELS:

```
HPSet ← ∅;
Initialize the event queue for the Bentley-Ottmann sweep.
while the event queue is nonempty do
  Remove the next event e from the queue.
  Let (x_e, y_e) be the coordinates of e.
  Advance the sweepline to x_e.
  If e is not a re-insertion then
    HPSet ← HPSet ∪ (round(x_e), round(y_e))
  If e is a right endpoint of ursegment s then
    Remove s from the sweepline.
  Else if e is a left endpoint or a re-insertion of ursegment s then
    Insert s into the sweepline.
  Else {e is an intersection of ursegments s_1 and s_2}
    h = (round(x_e), round(y_e));
    trim(s_1, h); trim(s_2, h);
```

As in a standard Bentley-Ottmann sweep, any modification of the sweepline contents (insertion or deletion of a segment) causes the next events for those segments and their neighbors to change, and the modified x-coordinates of those events are propagated up the tree, so that each node records the x-value of the current leftmost event in its subtree. The following lemmas establish the correctness and runtime performance of FINDHOTPIXELS:

Lemma 3.1 *Algorithm* FINDHOTPIXELS *correctly computes all hot pixels in* $\mathcal{G}$.

Proof Because the algorithm recognizes a hot pixel only for ursegment endpoints or intersections, the set *HPSet* it computes is a subset of the hot pixels in $\mathcal{G}$. To argue that every hot pixel is added to *HPSet*, note that subsegments of an ursegment s are removed by $trim(s, h)$ only inside a known hot pixel h. The Bentley-Ottmann sweep algorithm detects all intersections between untrimmed ursegments. If a hot pixel contains no ursegment endpoints (i.e., it is made hot only by ursegment intersections), then at least one of the ursegment intersections it contains will be detected, because the participating ursegments will not be trimmed before the pixel is known to be hot. □

Define the *intersecting segment count* $is(h)$ to be the number of ursegments of S that have an endpoint or an intersection inside a pixel h. Note that $is(h)$ is *not* the number of intersections inside h. In fact, the number of ursegment intersections inside h may be as large as $\binom{|h|}{2} = \Theta(|h|^2)$, while $is(h)$ is never larger than $|h|$. Furthermore, $is(h)$ may be much less than $|h|$, which lends significance to the following result:

Lemma 3.2 *The running time of* FINDHOTPIXELS *is* $O(\sum_{h \in \mathcal{H}} is(h) \log n)$.

Proof An ursegment s is trimmed by $trim(s, h)$ after the first intersection that FINDHOTPIXELS detects for s inside h. Therefore the algorithm processes at most one intersection for each ursegment/pixel pair. The subsequent re-insertion event can be

charged to the $trim(s, h)$ operation that scheduled it. The algorithm performs work for an ursegment s only if it has an intersection or an endpoint inside h, and the number of operations for each ursegment is $O(1)$ per pixel. Each operation involves $O(1)$ standard binary tree operations on the sweepline and the event queue, which take $O(\log n)$ time apiece. □

3.2 Computing the Arcs of $\mathcal{G}$

This section presents an algorithm for computing the arcs of $\mathcal{G}$. The algorithm runs in the same time as FINDHOTPIXELS, and represents $segs(e)$ for each arc e using a persistent data structure. The core of the algorithm is a slight modification of the method of de Berg, Halperin, and Overmars [5]; a clean-up phase takes care of one special case that algorithm does not fully handle.

The algorithm of [5] assumes that the hot pixels have already been identified and computes the arcs between them using two Bentley-Ottmann sweeps. The first sweep processes ursegments with nonnegative slopes, and the second (symmetric) sweep processes those with negative slopes. The algorithm assumes that no ursegment is vertical, though this restriction is easy to enforce using a symbolic perturbation, if necessary.

This section presents the algorithm of [5] in some detail, because variations on this algorithm are important both here and in Sect. 4.3. The algorithm is based on a Bentley-Ottmann sweep over the ursegments of S with nonnegative slopes. The sequence of ursegments intersecting the sweepline is divided into subsequences (*bundles*) defined by the hot pixels that the ursegments intersect immediately to the left of the sweepline. A bundle is a maximal subsequence of ursegments with a single hot pixel predecessor. Bundles are recorded compactly in the tree representing the sweepline as follows (this detail differs from the algorithm in [5]): Call a node in the tree *pure* if all of its leaf descendants have the same predecessor hot pixel. Each maximal pure node (the node is pure, but its parent is not) is labelled with the identity of the hot pixel predecessor. Thus only $O(\log n)$ nodes in the tree are involved in labelling each bundle, and all these nodes are children of a path of length $O(\log n)$ in the tree.

The algorithm processes the hot pixels left-to-right, grouped in vertically aligned columns, from bottom-to-top within each column. The sweepline, instead of being a single vertical line, is a staircase consisting of at most five segments: below the current hot pixel h it coincides with the right side of the column, above h it coincides with the left side of the column, and within h a vertical segment sweeps from left to right. See Fig. 3. For each hot pixel h the algorithm FINDARCS performs the following steps:

1. Find the subsequence of ursegments in the sweepline that intersect h. The segments of the sweepline are properly ordered along the staircase profile, so this is a simple tree search.
2. For every bundle that hits h, create an arc e of $\mathcal{G}$ joining h to the bundle's predecessor hot pixel, and use persistence [7] to record the subset of the bundle that hits h (a subsequence in the current sweepline) as $segs(e)$. This can be done in $O(\log n)$ time per arc created.

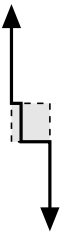


Fig. 3 The sweepline in the algorithm of [5] is a staircase with five segments

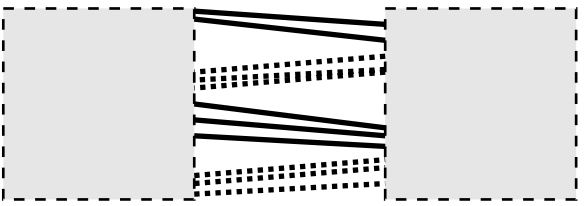


Fig. 4 The positive- and negative-slope ursegments for the arc joining these two hot pixels are stored in two separate sequences, one shown as dashed segments and one shown solid. Their interleaving is undetermined

3. Of the bundles that hit h, at most two may hit it with only a fraction of the bundle segments. One of these two partially passes below h, and the other above; the two rôles may even be filled by the same bundle. Split these (up to) two bundles and label the portion of each that misses h with the same label (the same predecessor) that it had originally.
4. Propagate the ursegments that hit h through the hot pixel, and insert or delete any ursegments with endpoints inside h. In the original paper [5] this propagation is done using a standard Bentley-Ottmann sweep in $O(int(h)\log n)$ time, where $int(h)$ is the number of ursegment intersections and endpoints inside h. It is straightforward to replace this step by a sweep that calls $trim(s, h)$ at the first intersection of an ursegment s, thereby reducing the time to $O(is(h)\log n)$. Note that no effort is needed to propagate the ursegments that do not hit the hot pixels. Their order is the same on both sides of the column.
5. Label all the ursegments that exit h on its top and right sides as a single bundle.

This algorithm finds all the arcs of $\mathcal{G}$ in $O(\sum_{h\in\mathcal{H}} is(h)\log n)$ time, and for every arc produces at most two sequences of ursegments (one for each sweep) that contain all the ursegments that belong to the arc. If the arc is not horizontal or vertical, it is discovered by only one of the two sweeps, and its sequence contains the ursegments in the order they appear in $\mathcal{A}$. If the arc is horizontal or vertical, the nonnegative-slope and negative-slope ursegments that define it are discovered in two separate sweeps, and recorded in two separate sequences. Each sequence is correctly ordered, but their possible interleaving is undetermined. See Fig. 4.

If it is important to represent every arc by a single properly ordered sequence of ursegments, this can be accomplished using two more sweeps, one horizontal and one vertical. Each sweep is responsible for creating bundles for the arcs perpendicular to the sweepline. The horizontal sweep, for example, passes a vertical sweepline over the arrangement as in algorithm FINDHOTPIXELS. When the sweep reaches the right side of a hot pixel h, it labels the bundle of ursegments that emerge from the right side of h with its predecessor h, as in algorithm FINDARCS. Because the algorithm is

not interested in ursegments that emerge from the tops or bottoms of h—they cannot contribute to a horizontal arc of $\mathcal{G}$—it is able to perform the labelling in a single logarithmic-time operation per hot pixel. When the sweep reaches the left side of a hot pixel h, it checks whether any of the ursegments that hit the left side of h emanate from a hot pixel h' straight left of h. If so (a logarithmic-time test), the subsequence that originates at h' and hits h can be identified and recorded using persistence as $segs(e)$, for $e = (h', h)$, in $O(\log n)$ time.

This completes the proof of the following theorem:

Theorem 3.3 *Given a set S of n ursegments, its snap-rounded arrangement $\mathcal{G} = (\mathcal{H}, \mathcal{E})$ can be computed in time $O(\sum_{h\in\mathcal{H}} is(h) \log n)$, where $is(h)$ is the number of ursegments of S that have endpoints or intersections inside a hot pixel h. The ursegment sequences associated with the arcs of $\mathcal{E}$ can be computed and recorded within a matching time and space bound.*

4 Edit Distance Sensitivity

Although the algorithm of the preceding section is a substantial improvement over both [5] and [9], it still seems suboptimal for some inputs. Consider the example shown in Fig. 5. In the figure, there are $\sqrt{2n}$ bundles containing $\sqrt{n/2}$ parallel ursegments apiece, with the bundles arranged in a grid. The total number of hot pixels is $2\sqrt{2n} + n/2$, and the total number of ursegment intersections is $n^2/4$. For each hot pixel h determined by ursegment intersections, $is(h) = \sqrt{2n}$ and $int(h) = n/2$. For this input, the algorithm of [5] runs in $O(n^2 \log n)$ time, and both [9] and the algorithm of Sect. 3 run in $O(n\sqrt{n} \log n)$ time. Nevertheless, it seems that the algorithms are missing an opportunity for efficiency, because the intersection pattern in each hot pixel is particularly simple. If one could take advantage of this simplicity, one could reduce the processing time further. The improved algorithm presented in this section achieves a running time of $O(n \log n)$ for the example in Fig. 5.

4.1 Edit Distance

Our intuition tells us that the crossover inside each hot pixel of Fig. 5 is simple. This section formalizes that intuition using the notion of *edit distance*, in particular

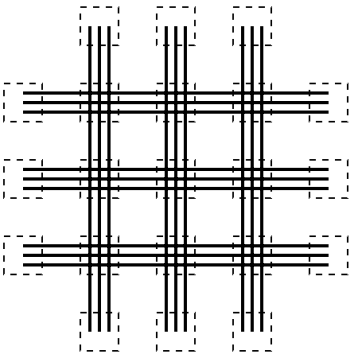


Fig. 5 In this grid of bundles, $\sum_{h\in\mathcal{H}} is(h) = \Theta(n\sqrt{n})$, although the crossover in each hot pixel is very simple

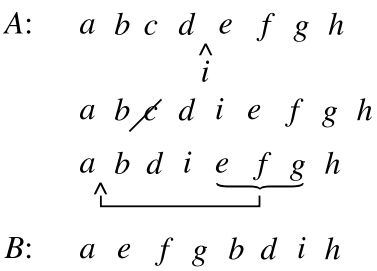


Fig. 6 A is transformed into B by three editing operations

edit distance with moves [4, 20]. Consider two sequences of symbols A and B, such that each symbol appears at most once in each of A and B. (That is, the sequences are *nonrepeating*.) The sequence A can be transformed into B by a series of *editing operations* of the following three types: insert a symbol at any position, delete a symbol at any position, and move a subsequence of symbols from any position in the sequence to some other position. See Fig. 6 for an example.

If the sequence is stored in a doubly linked list and pointers to the locations of operations are provided, each of these operations takes $O(1)$ time; if it is stored in a balanced binary tree, each takes $O(\log(|A| + |B|))$; if it is stored in a finger search tree [12, 16], then insert/delete take $O(1)$ amortized time and the move operation takes $O(\log(\ell + d))$, where ℓ is the length of the subsequence and d is the distance it moves.

The number of editing operations needed to transform one sequence into another is the edit distance between the sequences. However, this quantity may be difficult to calculate; an equivalent but easier-to-compute metric is the *neighbor difference distance*, defined to be the number of symbols in A and B whose neighbors are not identical in the two sequences. In Fig. 6 the neighbor difference distance is 8, because only f has its neighbors unchanged.

Lemma 4.1 *The edit distance and the neighbor difference distance between two nonrepeating sequences are equal to within a constant factor.*

Proof Each editing operation affects $O(1)$ neighbors. Thus there is a constant c such that the neighbor difference distance is at most c times the edit distance. On the other hand, if the neighbor difference distance between two sequences is d, then the sequences can be partitioned into $O(d)$ subsequences and singleton elements that can be rearranged with $O(d)$ editing operations to transform one sequence into the other. □

Because the edit distance and the neighbor difference distance are constant-factor equivalent, this paper uses the more euphonious name *edit distance* to refer to the easier-to-compute *neighbor difference distance*. That is, the edit distance $ed(A, B)$ between sequences A and B is actually computed as the neighbor difference distance.

The concept of edit distance for sequences carries over easily to the setting of a sweepline. Two different positions x_1 and x_2 of the sweepline induce two different sequences of ursegment intersections. Each ursegment is viewed as a symbol in the sequences, and the edit distance between sweepline positions x_1 and x_2, denoted

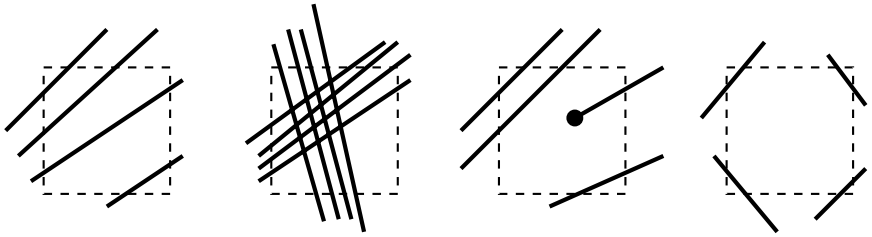


Fig. 7 The edit distances of these pixels in left-to-right order are 0, 4, 3, and 4(!)

$ed(x_1, x_2)$, is the number of ursegments whose neighbors along the sweepline differ at the two positions. If x_1 and x_2 are the left and right sides of a column of pixels C, then $ed(C) \equiv ed(x_1, x_2)$.

Extending the concept of edit distance to the sequence of ursegments crossing the boundary of a pixel is trickier, but possible. The complication arises because the sequence of ursegment intersections with the boundary is circular, and there is only one sequence, not two. Each ursegment that intersects a pixel h without having an endpoint inside it crosses the boundary of h twice. (An ursegment that is tangent to the boundary of h is defined to have zero or two intersections, depending on whether the boundary is open or closed at the point(s) of tangency.) The two intersections of such a *pass-through* ursegment s with the boundary of h are labelled s and s'. Intuitively, the edit distance of a pixel with no intersections or endpoints inside it should be zero. If a group of parallel ursegments passes through a pixel h without intersecting, then for every triple abc of ursegment boundary crossings that appears in counterclockwise order, the sequence must also contain the triple $c'b'a'$. Therefore the *edit distance of a pixel* h, $ed(h)$, is defined to be the number of ursegments that intersect h and either (a) have an endpoint inside, or (b) have neighbor pairs at the two boundary crossings that are not mirror reflections of each other. See Fig. 7. Ursegments with exactly one endpoint inside h cause a contribution of type (b) to $ed(h)$, so $ed(h)$ could also be defined in terms of the sequence at the boundary of h plus a term for the trivial ursegments fully contained in h.

Note that in the last example of Fig. 7, the edit distance is 4 even though there are no intersections or endpoints inside the pixel. This is somewhat of an anomaly, but it can be justified because such a configuration arises only if there are intersections or endpoints of the implicated ursegments in an adjacent pixel. The following lemmas characterize the pixel edit distance:

Lemma 4.2 *If a pixel h has no ursegment endpoints or intersections inside it, then* $ed(h) = O(1)$.

Proof The ursegments crossing h partition the interior of h into faces. Each face is a convex polygon, and all its vertices are either corners of h or intersections between ursegments and the boundary of h. The boundary of a face consists of an alternating sequence of ursegments and portions of the boundary of h. Each side of h appears at most once on the boundary of each face. An ursegment s contributes to $ed(h)$ if and only if at least one of the two faces it bounds has more than two ursegments on its

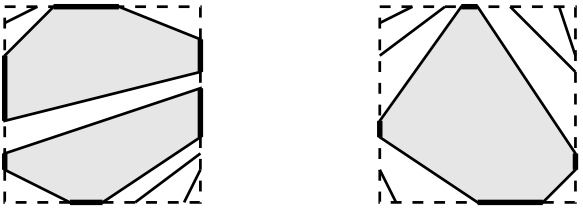


Fig. 8 A pixel containing no ursegment endpoints or intersections has edit distance at most 6

boundary. If a face has more than two ursegments on its boundary, it must also have portions of at least as many sides of h. Because the faces are interior-disjoint there can be at most two faces with three ursegments on their boundaries, or at most one face with four ursegments on its boundary. This proves the lemma. See Fig. 8. □

Lemma 4.3 *For any hot pixel h, $ed(h) = O(is(h))$.*

Proof Let T be the set of ursegments counted in $is(h)$. Estimate $ed(h)$ by removing all ursegments in T, then adding them back one by one. After all ursegments in T are removed from h, $ed(h) = O(1)$, by Lemma 4.2. Adding the ursegments back one by one increases $ed(h)$ by at most $O(1)$ per addition, because each new ursegment has at most four neighbors on the boundary of h. After all ursegments have been added, $ed(h) = O(1 + |T|) = O(is(h))$. □

Lemma 4.4 *For any hot pixel h, $ed(h) > 0$.*

Proof If any ursegment has an endpoint inside, the ursegment contributes to $ed(h)$, by definition. Otherwise, if there is an intersection inside h, then either the intersecting ursegments are neighbors on the boundary, or by induction on the number of ursegments separating them along the boundary there exists some pair of intersecting ursegments that *are* neighbors on the boundary. These two ursegments clearly have different neighbors on the opposite sides of h, and therefore contribute to $ed(h)$. □

The concept of edit distance for pixels can be generalized to convex regions, though the generalization of Lemma 4.2 holds only for constant-complexity convex polygons. The edit distance of a convex region R, $ed(R)$, is the number of ursegments that intersect R and either (a) have an endpoint inside, or (b) have neighbor pairs at the two boundary crossings that are not mirror reflections of each other. If there are no ursegments fully contained in a column C, then the convex region definition of edit distance is equivalent to $ed(C)$. The following lemma helps relate $ed(C)$ to $ed(h)$ for the hot pixels $h \in C$:

Lemma 4.5 *Let R be a convex region that is partitioned by a line ℓ into two convex fragments R' and R''. By convention assume that $\ell \cap R$ is included in at least one of R' and R''. Then*

$$ed(R) \leq ed(R') + ed(R'') + O(1).$$

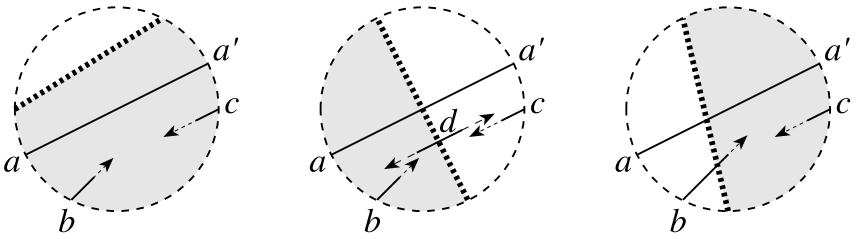


Fig. 9 Ursegment a is counted in the edit distance for at least one of the fragments R' and R'' created by the partitioning line

Proof It suffices to show that all but $O(1)$ segments that are counted in $ed(R)$ are counted in at least one of $ed(R')$ and $ed(R'')$. This is clear for segments with an endpoint in R, because the endpoint lies in R' or R''.

Consider a pass-through segment a, and without loss of generality suppose that the counterclockwise neighbor of crossing a is b and the clockwise neighbor of crossing a' is $c \neq b'$. Further assume that ℓ does not separate either of the pairs (a, b) or (c, a'). See Fig. 9. Without loss of generality suppose that R' contains (a, b). If R' also contains (c, a'), then a is counted in $ed(R')$. If (c, a') belongs to R'', then consider the intersections of ursegments a and b with ℓ. If crossings with b and a do not occur along ℓ in counterclockwise order in R', then a is counted in $ed(R')$; if they do, then the configuration of a in R is replicated in R'', and a is counted in $ed(R'')$. Because at most four ursegment crossings are separated from their neighbors along the boundary of R by ℓ, the lemma holds. □

The sequence edit distance for columns is related to the pixel edit distance as follows:

Lemma 4.6 *Let C be a column of pixels containing at least one hot pixel, and let $\mathcal{H} \cap C$ denote the set of hot pixels in C. Then*

$$ed(C) = O\Bigg(\sum_{h \in (\mathcal{H} \cap C)} ed(h)\Bigg).$$

Proof The proof of Lemma 4.2 extends trivially to show that if a rectangle R has no ursegment endpoints or intersections inside it, then $ed(R) = O(1)$. Let R_C be the convex hull of all the pixels in C that are crossed by ursegments. As observed earlier, $ed(C) = O(ed(R_C))$. Now partition R_C at the boundaries of the hot pixels into $|\mathcal{H} \cap C|$ hot pixels and at most $|\mathcal{H} \cap C| + 1$ rectangles containing no ursegment endpoints or intersections. By Lemma 4.5 and the extension of Lemma 4.2 to rectangles, $ed(R_C) = O(\sum_{h \in (\mathcal{H} \cap C)} ed(h)) + O(|\mathcal{H} \cap C|)$. By Lemma 4.4 this is $O(\sum_{h \in (\mathcal{H} \cap C)} ed(h))$. □

4.2 Edit-Distance-Sensitive Sequence Transformation

This section tells how to transform one sequence into another in time proportional to the edit distance times a logarithmic factor, given the availability of certain primitive operations that are easy to implement in the sweepline setting.

Consider the problem of transforming a sequence A into a second sequence B, assuming that A is stored at the leaves of a balanced binary tree, and that the transformation is implemented by destructive surgery on the tree. Define $n = \max(|A|, |B|)$, and assume the existence of a comparison operator $<_B$, which when applied to two elements $a, b \in B$ returns true if and only if a appears to the left of b in B. Further suppose that the input to the problem identifies all elements that appear in only one of A and B, and all consecutive pairs $a, b \in A$ such that $b <_B a$. The following algorithm dismantles A and constructs B from the resulting fragments:

Algorithm EDITDISTTRANSFORM:
Delete from A all elements in $A \setminus B$.
Break A into subsequences $A_1, A_2, \ldots, A_k$ by splitting at all the deletion sites and between every neighbor pair a, b such that $b <_B a$.
{Each subsequence is correctly ordered in both A and B, and the ends of each subsequence are counted in the edit distance $ed(A, B)$.}
Set $T \leftarrow \emptyset$, an empty sequence.
for $i \leftarrow 1$ to k do
$T \leftarrow merge(T, A_i)$; {Now $\forall a, b \in T, a <_B b$}
Insert into T all elements in $B \setminus A$.
return T;

The subroutine $merge(T, A_i)$ merges two sorted sequences into one in time $O(\log n)$ times the number of fragments into which A_i is subdivided. It is an unsophisticated variant on an algorithm of Hwang and Lin [17] and can be expressed recursively as follows:

$merge(U, V)$
if $empty(U)$ then return V;
else if $empty(V)$ then return U;
else if $head(V) <_B head(U)$ then
return $merge(V, U)$;
{Now $head(U) <_B head(V)$}
Split U into U' and U'' at $head(V)$.
return $concat(U', merge(V, U''))$;

The split and concatenate operations take $O(\log n)$ time apiece and all other operations take constant time. The total number of operations is proportional to the number of contiguous fragments into which the input sequences U and V are decomposed.

Lemma 4.7 *Algorithm* EDITDISTTRANSFORM *runs in time* $O(ed(A, B) \log n)$.

Proof The number of operations the algorithm performs outside the $merge()$ calls is proportional to $ed(A, B)$. Each operation is a standard binary search tree operation and takes $O(\log n)$ time. Each call to $merge(T, A_i)$ takes time $O(t \log n)$, where t is the number of fragments into which A_i is split by the $merge()$ subroutine. If A_i is split into t fragments, that means that $t - 1$ pairs of adjacent symbols in A_i are separated in B, and therefore the total time spent in $merge()$ is $O(ed(A, B) \log n)$. □

Note that EDITDISTTRANSFORM is essentially an adaptive algorithm for sorting a partially pre-sorted sequence, where the measure of disorder is the edit distance. Other such adaptive sorting algorithms are described in [8].

4.3 Computing the Hot Pixels

This section shows how to provide the input data and comparison operator required by the algorithm EDITDISTTRANSFORM for the ursegment sequences determined by two positions of the sweepline. It follows that it is possible to advance the sweepline from position x to another position x' in time $O(ed(x, x') \log n)$. It seems more difficult to perform a similar operation for the ursegments incident to a hot pixel, largely because it is hard to determine which ursegments are entering and which are exiting; indeed, ursegments may both enter and exit (during a left-to-right sweep) through the top and bottom of a hot pixel. Therefore the algorithm presented here finesses the issue, reducing the problem to a series of cases in which algorithm EDITDISTTRANSFORM *is* applicable.

Hot pixels determined by ursegment endpoints are easy to detect; the algorithm focuses on finding hot pixels determined only by intersections. These fall into two classes: (a) *same-side* pixels in which two ursegments cross the same side of the pixel and intersect inside and (b) *cruciform* pixels in which every intersecting pair consists of one ursegment crossing the top and bottom of the pixel and another crossing the left and right sides. The two classes of hot pixels are detected separately.

Lemma 4.8 *If a hot pixel h contains no ursegment endpoints and there exist two ursegments that cross the same edge of h and intersect inside h, then there are two ursegments that intersect inside h and are adjacent along the specified edge.*

Proof The proof is by induction on the number of ursegments that separate the two chosen ursegments along the boundary of h. Suppose that ursegments a and b cross an edge e of h and intersect inside h. Ursegments a, b, and the edge e bound a triangle inside h. If a and b are not adjacent along e, then there exists some ursegment s that crosses e between them and intersects either a or b (say a) on the triangle boundary. Then a and s both cross e, intersect inside h, and are closer along e than a and b. By induction the lemma follows. □

As noted in Sect. 2, the binary tree implementing the sweepline stores a next-event x-value for each ursegment crossing the sweepline, and the nodes of the tree implement a tournament on these x-values. Thus it is possible to find in $O(t \log n)$ time all t active segments whose next scheduled event occurs left of any desired x-value. (Note that if an ursegment s has event x-value $\bar{x}$, that does not necessarily imply that the first event involving s occurs at $\bar{x}$; that claim is true only for the ursegment with the leftmost event x-value. What is true is that an ursegment with x-value $\bar{x}$ will have an event at or before $\bar{x}$, assuming the neighbor defining the event is not deleted first.) The t ursegments with events left of a given x-value $\bar{x}$ partition the sweepline into at most $t + 1$ subsequences with the property that each subsequence, considered in isolation, has no events left of $\bar{x}$. That is, the vertical order of each subsequence is the

same at $\bar{x}$ as at the current sweepline position. This partition into event-free subsequences is part of the input needed by algorithm EDITDISTTRANSFORM; the other part, the comparison operator, simply checks the intersection order of two ursegments with the vertical line $x = \bar{x}$.

The following algorithm finds all same-side pixels whose defining pair of ursegments crosses the left pixel boundary:

Algorithm EDITDISTSAMESIDE:
HPSet $\leftarrow \emptyset$;
Initialize the endpoint queue and the sweepline.
while the endpoint queue is nonempty do
Let $\bar{x}$ be the x-coordinate of the next event (either an endpoint or an intersection).
Let $x_h = round(\bar{x})$.
{The sweepline currently holds ursegments in proper order for $x_h - \frac{1}{2}$.}
Find all ursegments in the sweepline with a scheduled event before $x_h + \frac{1}{2}$.
For each scheduled event e with $x_e < x_h + \frac{1}{2}$ do
HPSet $\leftarrow$ *HPSet* $\cup$ $(x_h, round(y_e))$;
Apply EDITDISTTRANSFORM to advance the sweepline from $x_h - \frac{1}{2}$ to $x_h + \frac{1}{2}$, updating the endpoint queue as necessary.

It follows from Lemmas 4.6, 4.7, and 4.8 that EDITDISTSAMESIDE finds all hot pixels determined by intersecting ursegments that enter through the left side in time $O(\sum_{h \in \mathcal{H}} ed(h) \log n)$. Applying the algorithm four times, once for each cardinal direction, finds all hot pixels except the cruciform pixels. In fact, it is enough to apply the algorithm twice: algorithm EDITDISTTRANSFORM detects every ursegment whose sweepline neighbors at $x_h - \frac{1}{2}$ differ from its neighbors at $x_h + \frac{1}{2}$, and this means that one invocation of EDITDISTSAMESIDE can detect all same-side hot pixels determined by ursegments crossing either their left or right sides.

The algorithm for detecting cruciform pixels is loosely based on the algorithm of [5] for arc computation described in Sect. 3.2. If the arc computation is omitted, that algorithm can be used to find intersections between pairs of positive-slope ursegments and pairs of negative-slope ursegments, but not between pairs with one positive and one negative slope. The extension described here depends on the observation that it is not the slope of the ursegments that is important for the algorithm, but how the ursegments cross the boundary of a hot pixel. In particular, ursegments are allowed to enter a hot pixel only through its left and bottom sides, and allowed to exit only through its right and top sides. This is automatically true for positive-slope ursegments, but it can be enforced for ursegments of all slopes by modifying the algorithm. In brief, the algorithm EDITDISTCRUCIFORM described below processes all the ursegments together, but gives special treatment to ursegments that enter a hot pixel h through its top boundary. Such ursegments are not processed inside h, but are instead merged directly into the sequence at the right boundary of the current column.

The algorithm EDITDISTCRUCIFORM processes hot pixels column-by-column from left to right. A hot column is identified by an ursegment endpoint or by an ursegment intersection event detected by the sweepline. Within each column C of hot pixels, the algorithm repeatedly identifies and processes the lowest unprocessed

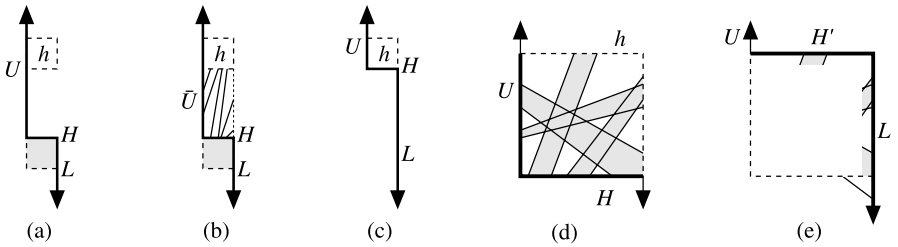


Fig. 10 **a** Configuration at the start of each step of EDITDISTCRUCIFORM. **b** Let $\bar{U}$ be the portion of U below the next hot pixel h. Separate the non-crossing ursegments in $\bar{U} \cup H$ into those that hit h and those that do not (those that cross the thin dashed segment in the figure). **c** Configuration before processing hot pixel h. **d** Split the sequence of ursegments entering h through its bottom and left sides into subsequences with no intersections inside or below h. Further split each subsequence into ursegments that cross the top of H and those that do not. Note that some ursegments from U may cross through the bottom of h. **e** Merge the subsequences at the top of h to create H'; merge the other subsequences into L

hot pixel. At the beginning of each step the sweepline profile is a stairstep, with one horizontal segment at the top of the previously processed hot pixel in C. See Fig. 10a. The sweepline is partitioned into three portions U, H, and L. U and L are the upper and lower vertical portions, with U at the left side of C and L at the right; H is the horizontal segment joining U and L. All ursegments associated with H have positive slope (they pass from below H to above it). As the algorithm runs, ursegments are removed from U and added to L; ursegments are both added to and removed from H, and the y-coordinate associated with H increases.

The ursegments crossing H (the top of the previous hot pixel) and U below the next hot pixel h form a contiguous subsequence on the sweepline. These ursegments have no events scheduled in the region between H and h (else h would be lower). Thus they are intersection-free in that region, and a single tree search separates the subsequence that hits the bottom of h from the one that reaches the right edge of C. (See Fig. 10b.) The portion that reaches the right edge of C is merged into L, and the portion that reaches h becomes the new contents of H; H moves up to the bottom of h (Fig. 10c). To process h, the algorithm examines the subsequence of $U \cup H$ that crosses the left and bottom sides of h, and splits it into subsequences with no scheduled events in or below h. Each of these subsequences consists of ursegments that pass through h without any neighbor intersections, starting from the bottom and left sides of h. Each subsequence is further split into one piece that hits the top of h and one that does not (Fig. 10d). The portions that hit the top are merged to create a new sequence H' to replace H; the portions that do not hit the top are merged into L. In the process, H is emptied and the part of U that crosses the left edge of h is deleted (Fig. 10e). Note that ursegments from U that hit the bottom of h receive no further processing in C below h. Any intersections those segments have in C below h are not detected. The propagation of U and H through h is very similar to the processing of Algorithm EDITDISTTRANSFORM, except applied on a pixel-by-pixel basis. The key to applying that algorithm is the separation between the source sequences (U and H) and the destination sequences (H' and L).

To find the lowest pixel above H containing a scheduled event, and to support the splitting of H and U into subsequences with no scheduled event inside or below a pixel h, the algorithm uses a quantized and enhanced version of the sweepline

described in Sect. 2. Each leaf node (representing an ursegment) has an associated event location (x_e, y_e), as before, but internal nodes store both coordinates of an event. Among the events in its subtree with minimal $round(x_e)$ values, each node stores the one with minimum y_e. That is, the events are grouped into columns of pixels, and a node selects the lowest event in the leftmost event-containing column. This value can be computed in constant time at each node based on the values stored at the node's children.

Observation 4.9 *The quantized sweepline data structure described above stores at its root the lowest scheduled event in the leftmost column of pixels containing an event.*

Lemma 4.10 *If all the ursegment events stored in a quantized sweepline lie in or to the right of a column of pixels C, then the sweepline can be used to find all scheduled events lying in or below a pixel $h \in C$ in $O(\log n)$ time apiece.*

Proof The search rule is simple: if and only if an internal node being visited stores an event in the query region, visit the node's children. Because each node stores an event belonging to one of its leaves, it is clear that if the children are visited, then the subtree contains an event of interest. But conversely, if the children are not visited, then the subtree contains no event of interest: if a node v's event location is (x_v, y_v) and the center of h is (x_h, y_h), then if $round(x_v) > x_h$ all descendants of v are right of C, and if $y_v \geq y_h + \frac{1}{2}$ then all descendants of v in C are above h. For each event found, the algorithm visits all the ancestors of the leaf containing the event and all the ancestors' siblings, for a total cost of $O(\log n)$ per event reported. □

Lemma 4.11 *If a cruciform pixel h contains a nonnegative-slope ursegment that passes through h from bottom to top, the algorithm* EDITDISTCRUCIFORM *detects an event inside h.*

Proof By a slight modification of the proof of Lemma 4.8 one can show that there exists a pair of ursegments that are adjacent along U and H (one on U and one on H) that intersect inside h. This pair defines an intersection event inside h that will be detected no later than the processing of the hot pixel below h. □

Note that EDITDISTCRUCIFORM does *not* detect all hot pixels. In particular, all ursegments that enter a pixel from the top are merged directly into the sweepline L, and so any cruciform pixels determined only by such ursegments will not be detected. See Fig. 11.

Lemma 4.12 *The total running time of algorithm* EDITDISTCRUCIFORM *is $O(\sum_{h \in \mathcal{H}} ed(h) \log n)$.*

Proof The algorithm spends $O(\log n)$ time to detect each hot pixel that it finds and to process the part of C between successive hot pixels. Within each hot pixel h the algorithm spends $O(ed(h) \log n)$ time to propagate H and the relevant part of U across h and into H' and L, as in Lemma 4.7. A further term of $O(ed(C) \log n)$

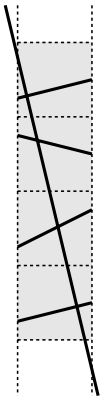


Fig. 11 Each of the shaded pixels is hot, but only the top one will be found by the positive-slope instance of EDITDISTCRUCIFORM. The others are found by the negative-slope instance

covers the cost of merging into L the portions of U that exit through the bottom of h for all hot pixels $h \in C$. Because $ed(C) = O(\sum_{h\in(\mathcal{H}\cap C)} ed(h))$ by Lemma 4.6, this term is dominated by the sum of the per-pixel costs. □

By Lemma 4.11, two applications of Algorithm EDITDISTCRUCIFORM, one each for positive and negative slopes, suffice to find all cruciform hot pixels. Combining this with two applications of EDITDISTSAMESIDE finds all hot pixels in a total of $O(\sum_{h\in\mathcal{H}} ed(h)\log n)$ time.

4.4 Computing the Arcs of $\mathcal{G}$

A relatively straightforward extension of the algorithms of Sect. 3.2 computes the arcs of $\mathcal{G}$ and their associated ursegment sets in $O(\sum_{h\in\mathcal{H}} ed(h)\log n)$ time. The step in the algorithm of de Berg, Halperin, and Overmars that processes a hot pixel h is modified, as in the previous subsection, to advance a quantized sweepline over h in time $O(ed(h)\log n)$. The algorithm is somewhat simpler than that in Sect. 4.3, however, because it does not need to handle negative-slope ursegments. Likewise, if the clean-up phase needs to perform additional sweeps to compute the ursegment sets of horizontal and vertical arcs, it uses the method of algorithm EDITDISTSAMESIDE to sweep over each column C of hot pixels in time $O(ed(C)\log n)$.

These observations, plus the algorithms of Sect. 4.3, establish the following theorem:

Theorem 4.13 *Given a set S of n ursegments, its snap-rounded arrangement $\mathcal{G} = (\mathcal{H}, \mathcal{E})$ can be computed in time $O(\sum_{h\in\mathcal{H}} ed(h)\log n)$, where $ed(h)$ is the edit distance of a pixel h. The ursegment sequences associated with the arcs of $\mathcal{E}$ can be computed and recorded within a matching time and space bound.*

The edit distance of every hot pixel where bundles cross in Fig. 5 is $O(1)$, and so the running time of this algorithm applied to that set of ursegments is $O(\sqrt{n}\log n)$ per endpoint-containing pixel and $O(\log n)$ per intersection-containing pixel, for a total time of $O(n\log n)$.

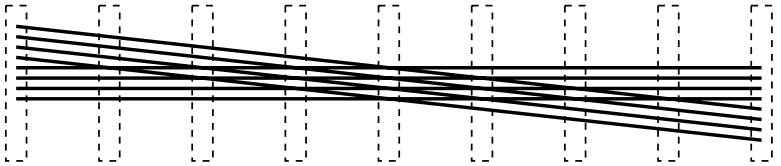


Fig. 12 $\sum_{h\in\mathcal{H}} ed(h) = \Theta(n^2)$, although $\mathcal{G}$ has linear size. (The arrangement is stretched vertically for clarity)

5 Conclusion

The algorithm of Sect. 3 is a practical addition to the available algorithms for computing snap rounded arrangements of line segments. It avoids the excessive running times of previous algorithms by a simple idea that leads to a simple algorithm.

The algorithm of Sect. 4 is primarily of theoretical interest because it requires several independent sweeps over the ursegments. Nevertheless, it points the way toward a new class of snap rounding algorithms that depend on the edit distance of the hot pixels. If an algorithm is required to produce output that represents the ursegments associated with each arc of $\mathcal{G}$ in a sorted sequence, bounds depending on the pixel edit distance are arguably the best possible. Reordering the sequence of ursegments entering a pixel to obtain the sequence of ursegments exiting seems to require a number of operations proportional to the edit distance.

If the order of ursegments associated with arcs of $\mathcal{G}$ is unimportant, some improvement may still be possible. Consider the arrangement of ursegments shown in Fig. 12, in which most of the hot pixels have edit distance $\Theta(n)$, and $\sum_{h\in\mathcal{H}} ed(h) = \Theta(\sum_{h\in\mathcal{H}} is(h)) = \Theta(\sum_{h\in\mathcal{H}} |h|) = \Theta(I) = \Theta(n^2)$. For this arrangement all the known algorithms run in time $\Omega(n^2 \log n)$, even though $\mathcal{G}$ has size $O(n)$ and the unordered value of $segs(e)$ is the same for every arc in $\mathcal{G}$.

References

1. Bentley, J.L., Ottmann, T.A.: Algorithms for reporting and counting geometric intersections. IEEE Trans. Comput. **C-28**(9), 643–647 (1979)
2. Brown, K.Q.: Comments on "Algorithms for reporting and counting geometric intersections". IEEE Trans. Comput. **C-30**, 147–148 (1981)
3. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 2nd edn. MIT Press, Cambridge (2001)
4. Cormode, G., Muthukrishnan, S.: The string edit distance matching problem with moves. In: Proc. 13th ACM-SIAM Sympos. Discrete Algorithms, pp. 667–676, 2002
5. de Berg, M., Halperin, D., Overmars, M.: An intersection-sensitive algorithm for snap rounding. Comput. Geom. Theory Appl. **36**, 159–165 (2007)
6. de Berg, M., van Kreveld, M., Overmars, M., Schwarzkopf, O.: Computational Geometry: Algorithms and Applications, 2nd edn. Springer, Berlin (2000)
7. Driscoll, J.R., Sarnak, N., Sleator, D.D., Tarjan, R.E.: Making data structures persistent. J. Comput. Syst. Sci. **38**, 86–124 (1989)
8. Estivill-Castro, V., Wood, D.: A survey of adaptive sorting algorithms. ACM Comput. Surv. **24**, 441–476 (1992)
9. Goodrich, M., Guibas, L.J., Hershberger, J., Tanenbaum, P.: Snap rounding line segments efficiently in two and three dimensions. In: Proc. 13th Annu. ACM Sympos. Comput. Geom., pp. 284–293, 1997
10. Greene, D.H.: Integer line segment intersection (unpublished manuscript)

11. Guibas, L., Marimont, D.: Rounding arrangements dynamically. In: Proc. 11th Annu. ACM Sympos. Comput. Geom., pp. 190–199, 1995
12. Guibas, L.J., McCreight, E., Plass, M., Roberts, J.: A new representation for linear lists. In: Proc. 9th Annu. ACM Sympos. Theory Comput., pp. 49–60, 1977
13. Halperin, D.: Problem 1: Output sensitive algorithm for snap rounding, June 2005. Open Problem Session: 21st Annu. ACM Sympos. Comput. Geom.
14. Halperin, D., Packer, E.: Iterated snap rounding. Comput. Geom. Theory Appl. **23**, 209–225 (2002)
15. Hobby, J.D.: Practical segment intersection with finite precision output. Comput. Geom. Theory Appl. **13**(4), 199–214 (1999)
16. Huddleston, S., Mehlhorn, K.: A new data structure for representing sorted lists. Acta Inform. **17**, 157–184 (1982)
17. Hwang, F.K., Lin, S.: A simple algorithm for merging two disjoint linearly ordered sets. SIAM J. Comput. **1**(1), 31–39 (1972)
18. Packer, E.: Iterated snap rounding with bounded drift. In: Proc. 22nd Annu. ACM Sympos. Comput. Geom., pp. 367–376, 2006
19. Sharir, M., Agarwal, P.K.: Davenport–Schinzel Sequences and Their Geometric Applications. Cambridge University Press, New York (1995)
20. Tichy, W.F.: The string-to-string correction problem with block moves. ACM Trans. Comput. Syst. **2**(4), 309–321 (1984)

Generating All Vertices of a Polyhedron Is Hard

Leonid Khachiyan · Endre Boros · Konrad Borys · Khaled Elbassioni · Vladimir Gurvich

Abstract We show that generating all negative cycles of a weighted graph is a hard enumeration problem, in both the directed and undirected cases. More precisely, given a family of negative (directed) cycles, it is an NP-complete problem to decide whether this family can be extended or there are no other negative (directed) cycles in the graph, implying that (directed) negative cycles cannot be generated in polynomial output time, unless $P = \text{NP}$. As a corollary, we solve in the negative two well-known generating problems from linear programming: (i) Given an infeasible system of linear inequalities, generating all minimal infeasible subsystems is hard. Yet, for generating maximal feasible subsystems the complexity remains open. (ii) Given a feasible system of linear inequalities, generating all vertices of the corresponding polyhedron is hard. Yet, in the case of bounded polyhedra the complexity

Communicated by Günter M. Ziegler.

This research was partially supported by the National Science Foundation (Grant IIS-0118635), and by DIMACS, the National Science Foundation's Center for Discrete Mathematics and Theoretical Computer Science. An extended abstract of this paper appears in the *Proceedings of the ACM–SIAM Symposium on Discrete Algorithms*, Miami, Florida, January 22–24, 2006.

Our friend and colleague, Leo Khachiyan, passed away with tragic suddenness while we were preparing this manuscript.

E. Boros (✉) · K. Borys · V. Gurvich
RUTCOR, Rutgers University, 640 Bartholomew Road, Piscataway, NJ 08854-8003, USA
e-mail: boros@rutcor.rutgers.edu

K. Borys
e-mail: kborys@rutcor.rutgers.edu

V. Gurvich
e-mail: gurvich@rutcor.rutgers.edu

K. Elbassioni
Department 1, Max-Planck-Institut für Informatik, 66123 Saarbrücken, Germany
e-mail: elbassio@mpi-sb.mpg

remains open. Equiva lently, the complexity of generating vertices and extreme rays of polyhedra remains open.

1 Introduction and Main Results

Let $G = (V, E)$ be a directed graph (digraph) and let $w\colon E \to \mathbb{R}$ be a real-valued weight function defined on its arcs. We call such a pair a *weighted digraph* and denote it by (G, w). For every subset of arcs $F \subseteq E$ its weight is defined as the total weight of all its arcs, $w(F) = \sum_{e \in F} w(e)$. We call a simple directed cycle a *circuit*. A circuit is called *negative* if its weight is negative. Finally, we denote by $\mathcal{C}^- = \mathcal{C}^-(G, w)$ the family of negative circuits of (G, w), i.e., $\mathcal{C}^- = \{C \subseteq E \mid C$ is a circuit with $w(C) < 0\}$.

First we consider the problem of generating exhaustively all negative circuits of a given weighted directed graph (G, w), in other words the problem of enumerating the family $\mathcal{C}^-(G, w)$. Since the number of negative circuits may be exponential in the size of the input description, i.e., the size of G and w, the efficiency of such enumeration algorithms is measured customarily in both the input and output sizes (see, e.g., [28, 32, 43]). More precisely, such an enumeration problem is said to be solvable in *polynomial total time* if the output can be generated in time polynomial in the input and output sizes. It is easy to see that for *self-reducible* (see, e.g., [29]) problems a family $\mathcal{C}$ is enumerable in polynomial total time if and only if for each subfamily $\mathcal{X} \subseteq \mathcal{C}$, the problem of deciding $\mathcal{X} \neq \mathcal{C}$; if yes, finding $C \in \mathcal{C} \backslash \mathcal{X}$, is solvable in time polynomial in size(G, w) and $|\mathcal{X}|$. On the other hand, when this decision problem is NP-hard, the enumeration problem is called NP-hard, too (see [32]). Thus, NP-hard enumeration problems are unlikely to have total polynomial time enumeration algorithms, unless $P = \text{NP}$.

Our main result claims that enumerating negative circuits of a weighted directed graph is a hard enumeration problem.

Theorem 1 *Given a weighted digraph $G = (V, E)$, $w\colon E \to \mathbb{R}$, and a family $\mathcal{X} \subseteq \mathcal{C}^-$ of its negative circuits, it is an NP-complete problem to decide whether $\mathcal{X} \neq \mathcal{C}^-$, even if w takes only two different values.*

We add that the analogous hardness result can be shown for undirected graphs, as well. In this case we also call a simple cycle a circuit and we denote by $\mathcal{C}^- = \mathcal{C}^-(G, w)$ the family of all negative circuits of an undirected graph $G = (V, E)$.

Theorem 2 *Given a weighted undirected graph $G = (V, E)$, $w\colon E \to \mathbb{R}$, and a family $\mathcal{X} \subseteq \mathcal{C}^-(G, w)$ of its negative circuits, it is an NP-complete problem to decide whether $\mathcal{X} \neq \mathcal{C}^-$, even if w takes only two different values.*

We remark that all circuits of a directed or undirected graph can be enumerated efficiently, e.g., by a simple backtracking algorithm [37].

Note that if w takes the same value for all edges (arcs), then negative circuits either do not exist or all circuits are negative. Thus, the enumeration problems for both directed and undirected graphs can be solved efficiently, as we noted earlier. Furthermore, when w takes only two different values, those can be assumed to be

integers, and hence by edge (arc) splitting the input can be transformed to one in which all edges (arcs) have weight ± 1. Though this transformation may increase the size of the input in a nonpolynomial way, in the case of the specific constructions we provide in the proofs of the above two theorems, it is a polynomial transformation, implying that generating all negative circuits is NP-hard even if all edges (arcs) have weights ± 1.

We derive several consequences of the above results, including the hardness of generating all vertices of a (possibly unbounded) polyhedron, generating all minimal infeasible subsystems of a system of linear inequalities, etc. We prove Theorems 1 and 2 in Sects. 2 and 3, respectively.

1.1 Negative Circuits and Minimal Infeasible Subsystems

We first note that deciding the existence and finding a negative circuit in a weighted directed graph are polynomially solvable tasks. Gallai [25] proved that (G, w) has no negative circuit if and only if by a potential transformation all edge weights can be changed to nonnegative values. Furthermore, a negative circuit can be found in $O(|V|^3)$ time, if the graph has negative circuits [23, 44]. We use Gallai's approach to reformulate the problem and derive some interesting consequences.

To a weighted digraph (G, w), where $G = (V, E)$ and $w\colon E \to \mathbb{R}$, we associate a polyhedron $P(E, w)$ defined by

$$P(E, w) = \left\{ x \in \mathbb{R}^V \;\middle|\; x_v - x_u \leq w(u, v) \text{ for all arcs } (u, v) \in E \right\}. \tag{1}$$

Note that every vector $x \in P(E, w)$ is a potential in the sense Gallai [25] defined it, proving that G is negative circuit free. Namely, defining $w'(u, v) = w(u, v) + x_u - x_v$ for all arcs $(u, v) \in E$ we get another weighting of the arcs of G, such that $w'(C) = w(C)$ for all directed circuits $C \subseteq E$, and for which $w'(u, v) \geq 0$ for all arcs $(u, v) \in E$, according to the definition of $P(E, w)$. This latter shows that G is indeed negative cycle free.

Thus applying Gallai's result to subgraphs of G we obtain that $P(E', w) = \emptyset$ for some $E' \subseteq E$ if and only if the subgraph $G' = (V, E')$ contains a negative cycle with respect to the weight function w. Therefore, the minimal infeasible subsystems of the system of linear inequalities (1) correspond in a one-to-one way to the negative circuits of (G, w). Hence, Theorem 1 implies the following result.

Corollary 1 *Enumerating all minimal infeasible subsystems of a system of linear inequalities is an NP-hard enumeration problem, even if we restrict the input to linear systems involving at most two variables in each inequality.*

The problems of finding minimal infeasible subsystems of a system of linear inequalities, sometimes called *Irreducible Inconsistent Subsystems* (IIS) or *Helly systems*, and its natural dual of finding maximal feasible subsystems received ample attention in the literature, see, e.g., [5, 35, 38]. The optimization versions of these problems, i.e., finding a maximum cardinality feasible subsystem, and finding a minimum cardinality infeasible subsystem are known to be NP-hard, see, e.g., [14, 27, 35].

1.2 Minimal Infeasible Subsystems and Vertex Enumeration

Recall that the infeasibility of a system of linear inequalities is well characterized by the Farkas Lemma: either the system $Ax \geq b$ has a solution, or there exists a nonnegative vector $y \geq 0$ such that $y^T A = 0$ and $y^T b > 0$, but not both (see [21]). Using this claim, Gleeson and Ryan [26] associated to a system of linear inequalities $Ax \geq b$, $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, a so-called *alternative polyhedron* defined as $Q = \{y \in \mathbb{R}^m_+ \mid y^T A = 0,\ y^T b = 1\}$, and observed that minimal infeasible subsystems of $Ax \geq b$ are in a one-to-one correspondence with vertices of Q. Indeed, for every vector $y \in Q$ we consider the subsystem of $Ax \geq b$ corresponding to the support set $S(y) = \{i \mid y_i \neq 0\}$. By the Farkas Lemma, we have that these corresponding subsystems are indeed infeasible. Conversely, if S is the index set of an infeasible subsystem of $Ax \geq b$, then again by Farkas's lemma we have a vector $y \in Q$ for which $S(y) \subseteq S$. Thus, minimal infeasible subsystems correspond to vectors $y \in Q$ with minimal support sets, and hence those are indeed vertices of Q.

This observation, coupled with Corollary 1, implies the hardness of enumerating the vertices of polyhedra.

Corollary 2 *Enumerating all vertices of a rational polyhedron, given as the intersection of finitely many closed half-spaces, is an NP-hard enumeration problem.*

Proof We consider an infeasible system of rational linear inequalities $Ax \geq b$, and its alternative polyhedron Q. We can write Q equivalently as $Q = \{y \in \mathbb{R}^m \mid y \geq 0,\ A^T y \geq 0,\ -A^T y \geq 0,\ b^T y \geq 1,\ -b^T y \geq -1\}$, i.e., as the intersection of $m + 2n + 2$ closed half-spaces. Thus, by the above observation, enumerating the vertices of this rational polyhedron would also enumerate all minimal infeasible subsystems of $Ax \geq b$, which is an NP-hard enumeration problem according to Corollary 1. □

Vertex enumeration is a fundamental problem in computational geometry and polyhedral combinatorics (see, e.g., [19] for a list of applications), and has many equivalent formulations. Most notably for bounded polyhedra, vertex enumeration is equivalent with *facet generation*, i.e., enumerating the facets of a polytope given by an explicit list of its vertices (see, e.g., the so-called polytope–polyhedron problem in [31]).

We add that in this paper we consider polyhedra which have vertices. This condition is easy to check in polynomial time and does not restrict generality. We emphasize that whenever the system of equations $A^T y = 0$, $b^T y = 0$ has a nontrivial solution for which $y \geq 0$, then Q in Corollary 2 is an unbounded polyhedron. Thus, our reduction through Theorem 1 yields in general unbounded polyhedra, and hence does not imply the hardness of vertex generation for bounded polyhedra, which remains an open problem. Furthermore, and equivalently, the complexity of enumerating together vertices and extreme rays of polyhedra is also an open problem (any unbounded polyhedron P can be projectively transformed into a bounded polyhedron, by adding one "far face," whose vertices correspond to the extreme rays of P, see, e.g., [35]).

Numerous algorithmic ideas have been introduced in the literature (either for vertex or for facet enumeration, see e.g., [1, 3, 4, 6, 9, 10, 12, 15, 16, 18, 19, 33, 34, 36, 41, 42]). Efficient algorithms (typically linear in the number of vertices) were

proposed for several special cases, including simple polyhedra, i.e., in which every vertex is incident with exactly n facets [3], simplicial polyhedra, which are the dual of simple polyhedra [9], network polytopes [36], polytopes with zero–one vertices [10], and polyhedra in which every facet defining inequality involves at most two nonzero coefficients [1]. Furthermore, for fixed dimension both vertices and rays of a polyhedron can be enumerated efficiently [13]. However, no method proved to be efficient (yet) for the general case. In fact, several publications [2, 11, 24] analyzed the proposed general-purpose methods for vertex/facet enumeration, and showed that all of the known algorithms may require in the worst case superpolynomial time in the output size. Along the same lines, Corollary 2 shows that vertex enumeration is indeed a hard enumeration problem for unbounded polyhedra (unless of course $P = \text{NP}$).

In analyzing the reasons why backtracking methods are not efficient for vertex enumeration, in general, Fukuda et al. [24] noted that such methods require repeatedly solving decision problems, which turn out to be NP-hard. In particular, they showed that for a given rational polyhedron P and an open rational half-space $H = \{x \in \mathbb{R}^n \mid \alpha^T x > \beta\}$, it is NP-hard to decide if P has a vertex in H. We note that the same decision problem for bounded polyhedra is much easier, since it can be decided by maximizing $\alpha^T x$ over P, which is a linear programming problem, known to be polynomially solvable, see Khachiyan [30]. We can show, as a next corollary of Theorem 1, that the enumerative version of this decision problem is hard for bounded polyhedra.

To arrive at this claim, we recall that the vertices of the *circulation polytope*

$$P(G) = \left\{ y \in \mathbb{R}^E \;\middle|\; \begin{array}{ll} \sum_{v:\,(u,v)\in E} y_{uv} - \sum_{w:\,(w,u)\in E} y_{wu} = 0, & \forall u \in V, \\ \sum_{(u,v)\in E} y_{uv} = 1, & \\ 0 \le y_{uv}, & \forall (u,v) \in E \end{array} \right\}$$

of a directed graph $G = (V, E)$ correspond to circuits of G, namely for every vertex y of $P(G)$ its support set $S(y) = \{(u, v) \in E \mid y_{uv} \neq 0\}$ is a circuit in G.

We remark that $P(G)$ frequently occurs in the optimization literature under various names, e.g., as the trans-shipment or flow polyhedron, or simply as the set of feasible circulations, or feasible solutions to a trans-shipment problem, etc. (see, e.g., Chaps. 11–13 in [40]). The vertices and facial structure of $P(G)$ are well studied and understood. In particular, the vertices of $P(G)$ can be generated in linear (output) time by cycle enumeration [37].

Associating further to a rational weight function $w\colon E \to \mathbb{R}$ an open rational half-space defined by

$$H = \left\{ y \in \mathbb{R}^E \;\middle|\; \sum_{(u,v)\in E} w(u,v) y_{uv} < 0 \right\},$$

we get that the support sets of vertices of $P(G)$ belonging to H are exactly the negative circuits of the weighted directed graph (G, w). Thus, Theorem 1 readily implies the following claim.

Corollary 3 *Given a rational polyhedron P and an open rational half-space H, it is NP-hard to enumerate all vertices of P which belong to H, even if P is bounded.*

Many applications (see, e.g., [19]) call for the enumeration of all those basic feasible solutions to a linear programming problem (i.e., vertices of the corresponding polyhedron), the corresponding objective function value of which is above a given threshold. Corollary 3 indicates that unfortunately such enumeration problems are difficult in general, unless $P = \text{NP}$.

A further consequence of Theorem 1 is that enumerating all vertices of a bounded polyhedron P which do not belong to a given face of P is also hard, in general.

Corollary 4 *Given a bounded polyhedron P and a proper face F of it, it is NP-hard to enumerate the vertices of P which do not belong to F.*

Proof Let $\bar{H} = \{y \in \mathbb{R}^E \mid \sum_{(u,v)\in E} w(u,v)y_{uv} \leq 0\}$. Note that $P' = P(G) \cap \bar{H}$ is a bounded polyhedron, for which $\bar{H}$ is facet defining. Denoting this facet by F, the vertices of P' outside F correspond in a one-to-one way to the negative circuits of the weighted graph (G, w) to which we associated H and $P(G)$. Thus, the claim follows from Theorem 1. □

By Corollary 2 unless $P = \text{NP}$ there exists no algorithm that outputs in incremental (or total) polynomial time, the vertices and then the extreme directions of a polyhedron, in that order. In contrast we have the following statement.

Proposition 1 *If there exists an algorithm which enumerates all vertices of a bounded polyhedron in incremental polynomial time, then we can enumerate all extreme rays and then all vertices (in this order) of a polyhedron in incremental polynomial time.*

Proof Let $P = \{x \in \mathbb{R}^n\colon a_i^T x \leq b_i, i = 1, \ldots, m\}$ be an unbounded polyhedron and let V and R denote the set of vertices and the set of extreme rays of P, respectively. As before, we can assume that V contains at least one vertex v. Let $a = \sum_{i\colon a_i^T v = b_i} a_i$. Then $P' = \{x\colon a_i^T x \leq b_i, i = 1, \ldots, m, a^T x = -M\}$ is a bounded polyhedron whose vertices correspond to R, where M is an appropriately large constant. Furthermore, $P'' = \{x\colon a_i^T x \leq b_i, i = 1, \ldots, m, a^T x \geq -M\}$ is a bounded polyhedron whose vertices correspond to $V \cup R$.

Assuming the existence of an algorithm **A** that can enumerate all vertices of a bounded polyhedron in incremental polynomial time, it follows that for any given subset W of the vertices of that bounded polyhedron, we can decide if this subset contains all vertices or, if not, can generate a vertex not belonging to W, in time, polynomial in the size of the input description of the polyhedron and the set W of given vertices. This can be accomplished simply by running **A** until it stops, or it outputs $|W| + 1$ vertices, whichever happens earlier.

Thus, by first applying **A** to P' we can generate the set R incrementally efficiently. Furthermore, since R is a subset of the vertices of P'', we can continue by applying **A** to P'' and extend in this way the set R incrementally efficiently to $V \cup R$, as we described earlier. Hence, we can enumerate the set $V \cup R$ in the stated order, first R and then V, incrementally efficiently. □

1.3 Four Geometric Enumeration Problems

We finally recall four strongly related geometric enumeration problems. Let $\mathcal{A} \subseteq \mathbb{R}^n$ be a given subset of vectors in $\mathbb{R}^n$, fix a point $z \in \mathbb{R}^n$ called the *center*, and consider the following four definitions:

- A *simplex* is a minimal subset $X \subseteq \mathcal{A}$ containing the center in its convex hull, i.e., $z \in \text{conv}(X)$.
- An *anti-simplex* is a maximal subset $X \subseteq \mathcal{A}$ not containing the center in its convex hull, i.e., $z \notin \text{conv}(X)$.
- A *body* is a minimal (full-dimensional) subset $X \subseteq \mathcal{A}$ containing the center in the interior of its convex hull, i.e., $z \in \text{int}(\text{conv}(X))$.
- An *anti-body* is a maximal subset $X \subseteq \mathcal{A}$ not containing the center in the interior of its convex hull, i.e., $z \notin \text{int}(\text{conv}(X))$.

Equivalently, a simplex (body) is a minimal collection of the given vectors not contained in an *open* (*closed*) half-space through the center, while an anti-simplex (anti-body) is a maximal collection of vectors contained in an open (closed) half-space through the center. It can be seen easily that $|X| \leq n + 1$ for a simplex, and that $n + 1 \leq |X| \leq 2n$ for a body.

In what follows we assume that the center is at the origin, i.e., $z = 0$. For a given point set $\mathcal{A} \subseteq \mathbb{R}^n$ we denote, respectively, by $\mathcal{S}$ and $\mathcal{B}$ the hypergraphs on the base set $\mathcal{A}$, consisting of all simplices, and all bodies of $\mathcal{A}$. The corresponding families of maximal independent sets of these two hypergraphs are, respectively, all anti-simplices and anti-bodies of $\mathcal{A}$, denoted respectively by $\mathcal{S}^*$ and $\mathcal{B}^*$, i.e.,

$$\mathcal{S}^* = \{X \subseteq \mathcal{A} \mid X \text{ is maximal such that } X \not\supseteq S,\ \forall S \in \mathcal{S}\},$$

$$\mathcal{B}^* = \{Y \subseteq \mathcal{A} \mid Y \text{ is maximal such that } Y \not\supseteq B,\ \forall B \in \mathcal{B}\}.$$

Simplices, anti-simplices, bodies, and anti-bodies can naturally be related to minimal infeasible or maximal feasible subsystems of certain linear systems of inequalities. Namely, we denote by $A \in \mathbb{R}^{m \times n}$, where $m = |\mathcal{A}|$, the matrix whose row vectors are the vectors of $\mathcal{A}$, and we let $e \in \mathbb{R}^m$ denote the m-dimensional vector of all ones.

It follows from the above definitions that simplices and anti-simplices are in a one-to-one correspondence, respectively, with the minimal infeasible and maximal feasible subsystems of the linear system of inequalities:

$$Ax \geq \mathbf{e}, \quad x \in \mathbb{R}^n. \tag{2}$$

Similarly, it follows that bodies and anti-bodies correspond in a one-to-one way, respectively, to the minimal infeasible and maximal feasible subsystems of the system:

$$Ax \geq 0, \quad x \neq 0. \tag{3}$$

As for the complexity of these enumeration problems, it is known that the generation of anti-bodies is a hard problem:

Proposition 2 [7] *Given a set of vectors $\mathcal{A} \subseteq \mathbb{R}^n$, and a partial list $\mathcal{X} \subseteq \mathcal{B}^*$ of the anti-bodies of $\mathcal{A}$, it is NP-hard to determine if the given list is incomplete, i.e.,*

$\mathcal{X} \neq \mathcal{B}^*$, or not. Equivalently, given an infeasible system (3), *and a partial list of its maximal feasible subsystems, it is NP-hard to determine if the given partial list is incomplete or not.*

Enumeration of bodies turns out to be at least as hard as the well-known *hypergraph transversal problem* [8] whose exact complexity is still an outstanding open problem [20]. The best currently known algorithm for the hypergraph transversal problem runs in incremental *quasi-polynomial* time [22].

Proposition 3 [7] *The problem of incrementally enumerating bodies, for a given set of $m + n$ points $\mathcal{A} \subseteq \mathbb{R}^n$, includes as a special case the problem of enumerating all minimal transversals for a given hypergraph $\mathcal{H}$ with n hyperedges on m vertices. Equivalently, generating minimal infeasible subsystems of* (3) *is at least as hard as hypergraph transversal generation.*

The problem of generating simplices turns out to be equivalent, in general, to the problem of enumerating the vertices of bounded polyhedra, or enumerating the vertices and extreme rays of possibly unbounded polyhedra. To see this, we consider a vector set $\mathcal{A} = \{a_1, \ldots, a_n, b\} \subseteq \mathbb{R}^d$ and associate to it a polyhedron $P = \{x \in \mathbb{R}^n \mid Ax = -b, x \geq 0\}$, where $A = [a_1, \ldots, a_n]$ is the matrix with columns $a_1, \ldots, a_n$.

Recall that for a vector $y \in \mathbb{R}^n$ we called the set $S(y) = \{i \mid y_i \neq 0\}$ its support set.

Proposition 4 *If $y \in P$ is a vertex of P, then the set $\{a_i \mid i \in S(y)\} \cup \{b\}$ is a simplex of $\mathcal{A}$, while if $y \in P$ is an extreme ray of P, then the set $\{a_i \mid i \in S(y)\}$ is a simplex of $\mathcal{A}$. Furthermore, every simplex of $\mathcal{A}$ corresponds in this way either to a vertex or to an extreme ray of P.*

Proof It is well known that the vertices of P are the solutions which have minimal support sets, and the extreme rays are those solutions of the homogenized system (replace b by 0) which have minimal support sets (see, e.g., Chap. 8 in [39]). Clearly, the minimality of support sets in both cases implies the first two claims, by the definition of a simplex of $\mathcal{A}$.

For the last claim, let $S \subseteq \mathcal{A}$ be a simplex, i.e., a minimal subset for which $0 \in \text{conv}(S)$. If $b \in S$, then we have for some $\lambda_a \geq 0$, $a \in S \backslash \{b\}$, and $\lambda_b \geq 0$, with $\lambda_b + \sum_{a \in S \backslash \{b\}} \lambda_a = 1$, that

$$-\lambda_b b = \sum_{a \in S \backslash \{b\}} \lambda_a a.$$

Since S is minimal, we must have all these coefficients positive, and thus

$$-b = \sum_{a \in S \backslash \{b\}} \frac{\lambda_a}{\lambda_b} a.$$

Thus, the vector $x \in \mathbb{R}^n$, defined by

$$x_i = \begin{cases} \lambda_{a_i}/\lambda_b & \text{if } a_i \in S \backslash \{b\}, \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, n$, is a vertex of P, again by the minimality of S. While if $b \notin S$, then we have

$$0 = \sum_{a \in S} \lambda_a a$$

for some positive coefficients $\lambda_a > 0$, $a \in S$, for which $\sum_{a \in S} \lambda_a = 1$, and thus the vector $x \in \mathbb{R}^n$, defined by

$$x_i = \begin{cases} \lambda_{a_i} & \text{if } a_i \in S, \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, n$, is an extreme ray of P, once more by the minimality of S. □

In particular, if $P = \{x \in \mathbb{R}^n \mid Ax = b,\ x \geq 0\}$ is a bounded polyhedron, i.e., if $Ax = 0$ has no nontrivial nonnegative solutions, then the vertices of P correspond in a one-to-one way to the simplices of the set $\mathcal{A}$ formed by the column vectors of A and b.

For the special case of vectors $\mathcal{A} \subseteq \mathbb{R}^n$ in general position, we have $\mathcal{B} = \mathcal{S}$, and consequently the problem of enumerating bodies of $\mathcal{A}$ turns into the problem of enumerating vertices of the bounded polyhedron $\{x \in \mathbb{R}^n \mid Ax = 0, \mathbf{e}^T x = 1, x \geq 0\}$, each vertex of which is nondegenerate and has exactly $n + 1$ positive components. For such kinds of *simple* bounded polyhedra there exist algorithms that generate all vertices with polynomial delay (see e.g., [15] and [3]).

We finally mention that, although the status of the problem of enumerating all maximal feasible subsystems of (2) is not known in general, the situation changes if we fix a consistent subfamily of inequalities, and ask for enumerating all its extensions to a maximal feasible subsystem. In fact, such a problem turns out to be NP-hard, even if we fix only nonnegativity constraints.

Proposition 5 [7] *Let $A \in \mathbb{R}^{m \times n}$ be an $m \times n$ matrix, let $b \in \mathbb{R}^m$ be an m-dimensional vector, and assume that the system*

$$Ax \geq b, \quad x \in \mathbb{R}^n, \tag{4}$$

has no solution $x \geq 0$. Let $\mathcal{F}$ be the family of all maximal subsystems of (4) *which can be satisfied by a nonnegative solution x. Then, given a partial list $\mathcal{X} \subseteq \mathcal{F}$, it is an NP-complete problem to determine if the list is incomplete, i.e., if $\mathcal{X} \neq \mathcal{F}$, even if b is a unit vector, and entries in A are either, -1, 1, or 0.*

We conclude with the observation that the problem of finding, for an infeasible system

$$A'x \geq b', \qquad A''x \geq b'', \tag{5}$$

all maximal feasible subsystems extending the feasible subsystem $A''x \geq b''$, naturally includes both problems of generating anti-simplices and simplices. Clearly, the former problem can be written in the form (5) by considering (2) and all maximal extensions of an empty subsystem. For the latter problem, note that the vertices of a bounded polyhedron $\{x \in \mathbb{R}^n \mid Ax = b,\ x \geq 0\}$, where $b \neq 0$, are in one-to-one correspondence with the maximal feasible extensions of the subsystem $Ax = b$, $x \geq 0$

in the infeasible system $Ax = b$, $x \geq 0$, $x \leq 0$. Although the general problem of generating maximal feasible extensions is NP-hard as stated above, the special cases of generating simplices and anti-simplices remain open.

2 Proof of Theorem 1

In this section we prove Theorem 1 by a reduction from satisfiability, a well-known NP-complete problem (see [17]).

We consider n propositional Boolean variables X_j, $j = 1, \ldots, n$, we denote by $\overline{X} = 1 - X$ the negation of X, we call variables and their negations *literals*, and elementary disjunctions of literals *clauses*. We next consider an arbitrary conjunctive normal form (CNF) $\phi = C_1 \wedge C_2 \wedge \cdots \wedge C_m$, i.e., where C_i, $i = 1, \ldots, m$, are clauses. A truth assignment to the variables is called *satisfying* for the CNF ϕ, if ϕ evaluates to true, i.e., if at least one literal evaluates to true in each of the clauses of ϕ.

In what follows we associate to ϕ a weighted directed graph (G, w) and a set $\mathcal{X}$ of negative circuits of G such that (G, w) has a negative circuit not belonging to $\mathcal{X}$ if and only if ϕ has a satisfying assignment. Because (G, w) and $\mathcal{X}$ are constructed from ϕ in $O(mn)$ time, and the weight function w uses only two different values (1 and -1), Theorem 1 follows readily from this construction. This is because the decision problem "*Is there a negative circuit in* (G, w) *which does not belong to* $\mathcal{X}$*?*" is in NP. To complete the proof of Theorem 1, we provide in the following a construction with these properties, such that every satisfying assignment to ϕ corresponds to a negative circuit of (G, w) not belonging to $\mathcal{X}$ and, vice versa, every negative circuit of (G, w) which does not belong to $\mathcal{X}$ corresponds to a satisfying assignment of ϕ (though the correspondence is not necessarily one-to-one).

To describe our construction, we denote for $j = 1, \ldots, n$, respectively by o_j and $\bar{o}_j$, the number of occurrences of literal X_j and its negation $\overline{X}_j$; we denote by x_j^k the kth occurrence of X_j, $k = 1, \ldots, o_j$, and by $\bar{x}_j^k$ the kth occurrence of $\overline{X}_j$, $k = 1, \ldots, \bar{o}_j$, and let L denote the set of all literal occurrences, i.e.,

$$|L| = \sum_{i=1}^{m} |C_i| = \sum_{j=1}^{n} (o_j + \bar{o}_j).$$

Since monotone variables, i.e., ones for which $o_j = 0$ or $\bar{o}_j = 0$, can be easily eliminated from a satisfiability problem, we can assume without any loss of generality that $o_j > 0$ and $\bar{o}_j > 0$ hold for all variables $j = 1, \ldots, n$.

For instance, if $n = 3$ and

$$\phi = \left(X_1 \vee X_2 \vee \overline{X}_3\right) \wedge \left(X_1 \vee \overline{X}_2 \vee X_3\right) \wedge \left(\overline{X}_1 \vee X_2 \vee \overline{X}_3\right), \tag{6}$$

then we have $o_1 = 2$, $\bar{o}_1 = 1$, $o_2 = 2$, $\bar{o}_2 = 1$, $o_3 = 1$, $\bar{o}_3 = 2$, and

$$L = \left\{x_1^1, x_2^1, \bar{x}_3^1, x_1^2, \bar{x}_2^1, x_3^1, \bar{x}_1^1, x_2^2, \bar{x}_3^2\right\}.$$

We define the vertex set of the graph $G = (V, E)$ associated to ϕ as

$$V = U \cup Q \cup \bigcup_{j=1}^{n} (Y_j \cup Z_j),$$

where U, Q, and Y_j and Z_j for $j = 1, \ldots, n$ are pairwise disjoint, defined as

$$U = \{u_k \mid k = 0, 1, \ldots, m+n\},$$

$$Q = \big\{a(\ell), b(\ell) \;\big|\; \ell \in L\big\},$$

$$Y_j = \{y_{jk} \mid k = 1, \ldots, o_j - 1\} \quad \text{for } j = 1, \ldots, n, \quad \text{and}$$

$$Z_j = \{z_{jk} \mid k = 1, \ldots, \bar{o}_j - 1\} \quad \text{for } j = 1, \ldots, n.$$

The graph itself has a ring structure, the skeleton of which is the set U. For every variable X_j of ϕ we have two parallel directed paths from u_{j-1} to u_j. The first path corresponding to X_j contains vertices Y_j (and some other vertices), while the second path, corresponding to $\overline{X}_j$, passes through vertices of Z_j ($j = 1, \ldots, n$). For convenience, we also introduce the notation

$$y_{j0} = z_{j0} = u_{j-1} \quad \text{and} \quad y_{j,o_j} = z_{j,\bar{o}_j} = u_j \tag{7}$$

for $j = 1, \ldots, n$. To every clause C_i of ϕ we associate $|C_i|$ parallel directed paths from u_{n+i-1} to u_{n+i}, one for each of the literals in C_i ($i = 1, \ldots, m$). Finally vertices $a(\ell)$ and $b(\ell)$ correspond exclusively to literal occurrence $\ell \in L$.

We consider next the weighted graph $H(a, b, p, q, r, s)$ (see Fig. 1) on six nodes a, b, p, q, r, and s, having six arcs, the weights of which are as follows:

$$\begin{aligned} &w(a, b) = w(b, a) = -2 \quad \text{and} \\ &w(p, a) = w(b, q) = w(r, b) = w(a, s) = 1. \end{aligned} \tag{8}$$

To every literal occurrence $\ell \in L$ we associate a disjoint copy of $H(a, b, p, q, r, s)$, and denote by $a(\ell)$, $b(\ell)$, etc., its nodes, and by E_ℓ its arc set. Note that each of these small subgraphs can be decomposed into two directed paths, each consisting of three arcs, $E_\ell = E_\ell^v \cup E_\ell^c$, where

$$E_\ell^v = \big\{\big(p(\ell), a(\ell)\big), \big(a(\ell), b(\ell)\big), \big(b(\ell), q(\ell)\big)\big\}, \quad \text{and}$$

$$E_\ell^c = \big\{\big(r(\ell), b(\ell)\big), \big(b(\ell), a(\ell)\big), \big(a(\ell), s(\ell)\big)\big\}.$$

Finally we set

$$E = E_0 \cup \bigcup_{\ell \in L} E_\ell,$$

where $E_0 = \{(u_{m+n}, u_0)\}$ with weight $w(u_{m+n}, u_0) = -1$.

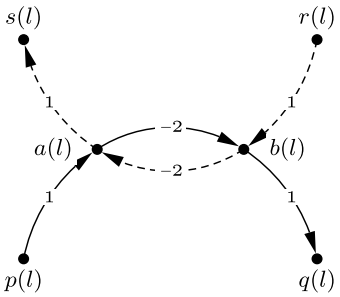


Fig. 1 The directed graph $H(a, b, p, q, r, s)$ associated with literal occurrences

In each of the subgraphs corresponding to the literal occurrences $\ell \in L$, we have the nodes $a(\ell)$ and $b(\ell)$ already introduced in $Q \subseteq V$, while the nodes $p(\ell)$, $q(\ell)$, $r(\ell)$, and $s(\ell)$ for $\ell \in L$ are corresponding to some other vertices of G, according to the following definitions:

$$p(\ell) = y_{j,k-1} \quad \text{and} \quad q(\ell) = y_{jk} \quad \text{if } \ell = x_j^k,$$
$$p(\ell) = z_{j,k-1} \quad \text{and} \quad q(\ell) = z_{jk} \quad \text{if } \ell = \bar{x}_j^k, \quad \text{and}$$
$$r(\ell) = u_{n+i-1} \quad \text{and} \quad s(\ell) = u_{n+i} \quad \text{if } \ell \in C_i.$$

In other words, for every literal occurrence ℓ of clause C_i the set E_ℓ^c forms a three-arc directed path from u_{n+i-1} to u_{n+i}. Furthermore, by (7) and by the above definitions, the sets E_ℓ^v for $\ell = x_j^1, x_j^2, \ldots, x_j^{o_j}$ form a directed path from u_{j-1} to u_j through the vertices of Y_j, consisting of $3o_j$ arcs, for every variable X_j. Similarly, the sets E_ℓ^v for $\ell = \bar{x}_j^1, \bar{x}_j^2, \ldots, \bar{x}_j^{\bar{o}_j}$ form another directed path from u_{j-1} to u_j through the vertices of Z_j, consisting of $3\bar{o}_j$ arcs.

In summary, $G = (V, E)$ consists of $|V| = 3|L| + m - n + 1$ vertices and $|E| = 6|L| + 1$ arcs, and the weight function w takes only values in $\{-2, -1, 1\}$. Note that we can split arcs of weight -2 to obtain a graph whose arcs all have weight ± 1.

Returning to the example CNF ϕ given in (6), the corresponding graph $G = (V, E)$ is shown in Fig. 2. To make the drawing of such a graph visually more clear, for every literal occurrence ℓ nodes $a(\ell)$ and $b(\ell)$ of G are represented by two separate points of the picture each, labeled as $a(\ell)$ and $a'(\ell)$, and as $b(\ell)$ and $b'(\ell)$, respectively. Similarly, node u_n is represented by two points in the figure, labeled u_n and u'_n. Arcs

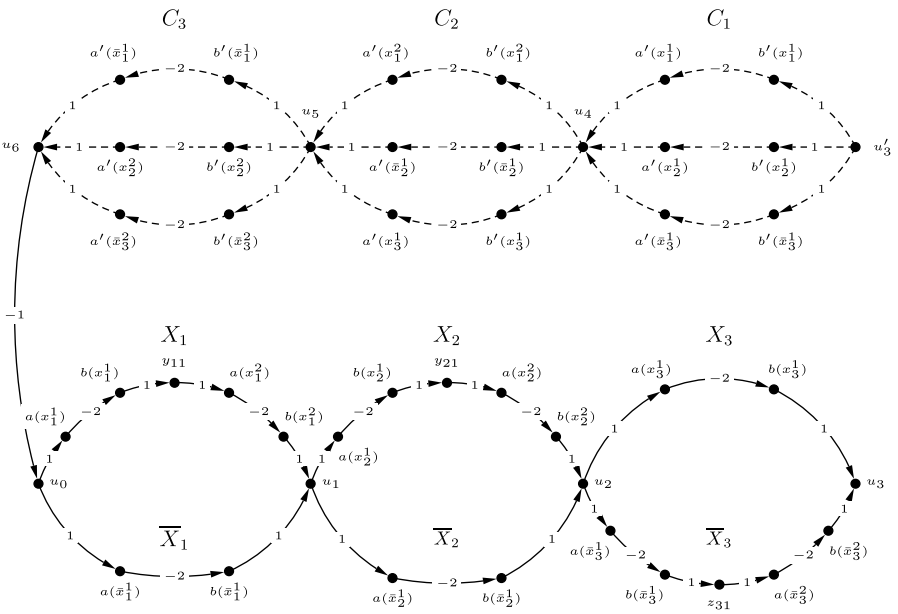


Fig. 2 G is obtained by identifying vertices $a(l)$, $a'(l)$, and $b(l)$, $b'(l)$, for each literal occurrence l, and u_3, u'_3 in the graph above. The *lower part* of the graph corresponds to the literals and the *upper part* corresponds to the clauses

in the sets E^c_ℓ for $\ell \in L$ are drawn as dashed lines, while those belonging to E^v_ℓ for $\ell \in L$ are drawn as solid lines.

Observe first that the arcs $(a(\ell), b(\ell))$ and $(b(\ell), a(\ell))$ form a circuit of total weight -4 for every literal occurrence $\ell \in L$. We denote by $\mathcal{X}$ the set of these circuits, i.e., $|\mathcal{X}| = |L|$, and we denote by $\mathcal{F}$ the set of all directed negative circuits of G.

We claim that from every satisfying assignment X of ϕ we can construct a directed negative circuit $D^X \in \mathcal{F}\backslash\mathcal{X}$ and, conversely, from every directed negative circuit $D \in \mathcal{F}\backslash\mathcal{X}$ we can construct a satisfying assignment X^D of ϕ. As we noted at the beginning of this section, this claim implies Theorem 1.

To see this claim, we first consider a satisfying assignment $X = (X_1, \ldots, X_n) \in \{0, 1\}^n$ of ϕ. Since X satisfies ϕ, we have a literal occurrence ℓ_i in every clause C_i, $i = 1, \ldots, m$, such that ℓ_i evaluates to true at X (i.e., $\ell_i(X) = 1$). We also denote by W the set of all those literal occurrences which evaluate to false at X, i.e., $W = \{\ell \in L \mid \ell(X) = 0\}$. Clearly, $\ell_i \notin W$ for $i = 1, \ldots, m$ by the above definitions. Then the set of arcs

$$D^X = \left(\bigcup_{i=1}^{m} E^c_{\ell_i}\right) \cup \left(\bigcup_{\ell \in W} E^v_\ell\right) \cup \left\{(u_{m+n}, u_0)\right\}$$

forms a circuit in G not belonging to $\mathcal{X}$. Since we have $w(E^c_\ell) = w(E^v_\ell) = 0$ for all literal occurrences $\ell \in L$, it follows by the above definitions that $w(D^X) = w(u_{m+n}, u_0) = -1$, i.e., $D^X \in \mathcal{F}\backslash\mathcal{X}$ as claimed.

We again return to the CNF ϕ given in (6). We consider the satisfying assignment $X = (1, 0, 0)$ of ϕ. We choose literal occurrences $\bar{x}^1_3 \in C_1$, $\bar{x}^2_1 \in C_2$, and $\bar{x}^2_3 \in C_3$ that evaluate to true at X. Figure 3 depicts the negative circuit $D^X = E^c_{\bar{x}^1_3} \cup E^c_{\bar{x}^1_2} \cup E^c_{\bar{x}^2_3} \cup E^v_{\bar{x}^1_1} \cup E^v_{\bar{x}^1_2} \cup E^v_{\bar{x}^2_2} \cup E^v_{\bar{x}^1_3} \cup (u_6, u_0)$.

Before proving the reverse direction of our main claim, we first observe some simple properties of our construction. To simplify notation, recall that $E_\ell = E^c_\ell \cup E^v_\ell$ for $\ell \in L$, and that the six-vertex subgraphs induced by the arc set E_ℓ have the same structure and weights, as in Fig. 1, for all $\ell \in L$. The following property of these subgraphs are instrumental in our proof.

Lemma 1 *Given a circuit $D \subseteq E$ of G, not belonging to $\mathcal{X}$, and given a literal occurrence $\ell \in L$, we have*

$$w(D \cap E_\ell) \in \{0, 2, 4\}.$$

Moreover, $w(D \cap E_\ell) = 0$ only if the set $D \cap E_\ell$ is one of the following three subsets of E_ℓ: E^c_ℓ, E^v_ℓ, or $\emptyset$.

Proof Since D is a circuit not belonging to $\mathcal{X}$, D cannot contain both arcs $(a(\ell), b(\ell))$ and $(b(\ell), a(\ell))$. Thus, denoting $A_\ell = \{(p(\ell), a(\ell)), (a(\ell), s(\ell))\}$ and $B_\ell = \{(r(\ell), b(\ell)), (b(\ell), q(\ell))\}$ we have that $D \cap E_\ell$ is one of the following six sets: $\emptyset$, A_ℓ, B_ℓ, $A_\ell \cup B_\ell$, E^c_ℓ, and E^v_ℓ. Since we have $w(\emptyset) = w(E^c_\ell) = w(E^v_\ell) = 0$, $w(A_\ell) = w(B_\ell) = 2$, and hence $w(A_\ell \cup B_\ell) = 4$, the statement follows. □

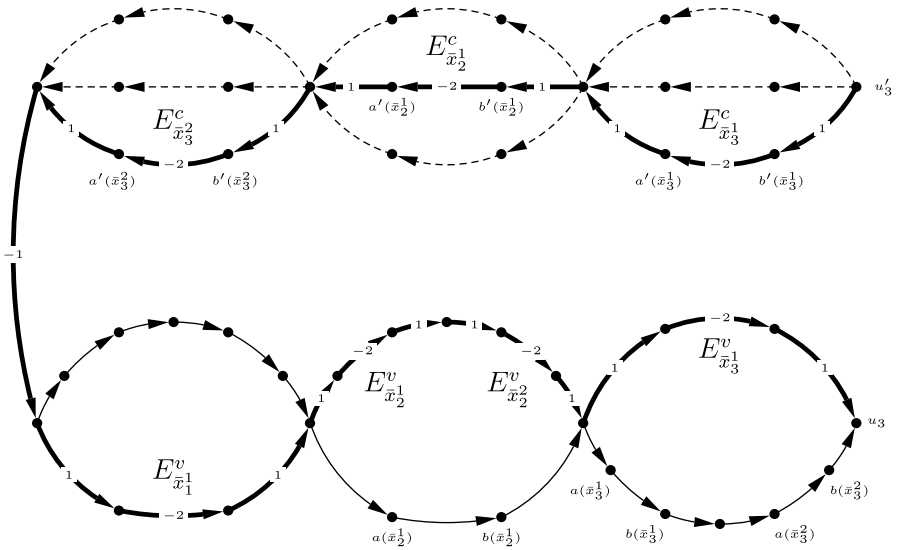


Fig. 3 *Thick lines* are edges of the negative circuit D^X corresponding to the satisfying assignment $X = (1, 0, 0)$. The vertices u_3, u'_3 are identified. Since D^X contains arcs $(b'(\bar{x}^1_3)a'(\bar{x}^1_3))$, $(b'(\bar{x}^1_2)a'(\bar{x}^1_2))$, and $(b'(\bar{x}^2_3)a'(\bar{x}^2_3))$ but it does not contain arcs $(a(\bar{x}^1_3)b(\bar{x}^1_3))$, $(a(\bar{x}^1_2)b(\bar{x}^1_2))$, and $(a(\bar{x}^2_3)b(\bar{x}^2_3))$, no circuit of $\mathcal{X}$ is contained in D^X

Returning to the reverse direction of our main claim, we consider a negative circuit $D \in \mathcal{F}\backslash\mathcal{X}$ of G. Since

$$w(D) = \sum_{\ell \in L} w(D \cap E_\ell) + w\big(D \cap \big\{(u_{m+n}, u_0)\big\}\big)$$

we must have by Lemma 1 that $(u_{m+n}, u_0) \in D$ and

$$w(D \cap E_\ell) = 0 \quad \text{for all } \ell \in L. \tag{9}$$

We show first that D passes through all vertices in U, includes exactly one of the two parallel paths between u_{j-1} and u_j for $j = 1, \ldots, n$, and exactly one of the parallel paths between u_{n+i-1} and u_{n+i} for all $i = 1, \ldots, m$.

As we observed above, we have u_0 as a vertex of D. Thus D must contain an arc leaving u_0, say it contains $(u_0, a_{x^1_1})$. Then, by (9) and by Lemma 1, we must have $E^v_{x^1_1} \subseteq D$, i.e., D must pass through vertex y_{11}. Since only $(y_{11}, a(x^2_1))$ is leaving y_{11}, by repeating the above argument we can conclude that we must also have $E^v_{x^2_1} \subseteq D$, etc., finally arriving at $E^v_{x^{o_1}_1} \subseteq D$, i.e., that D includes u_1 as a vertex. Repeating the same argument, we can prove by induction that for all indices $j = 1, \ldots, n$, if $E^v_{x^1_j} \subseteq D$, then we must have $E^v_{x^k_j} \subseteq D$ for all $k = 1, \ldots, o_j$, and that if $E^v_{\bar{x}^1_j} \subseteq D$, then we must also have $E^v_{\bar{x}^k_j} \subseteq D$ for all $k = 1, \ldots, \bar{o}_j$. We then define a truth assignment

X^D by

$$X_j^D = \begin{cases} 1 & \text{if } E^v_{\bar{x}_j^1} \subseteq D, \\ 0 & \text{if } E^v_{x_j^1} \subseteq D. \end{cases}$$

Furthermore, repeating a similar argument for vertices $u_n, u_{n+1}, \ldots, u_{n+m-1}, u_{n+m}$ we can also conclude that D must contain the set $E^c_{\ell_i}$ for exactly one of the literals $\ell_i \in C_i$, for each clause C_i of ϕ. Since D is a circuit in which no vertex $a(\ell)$ or $b(\ell)$ is repeated, we must have that $\ell_i(X^D) = 1$ for all $i = 1, \ldots, m$, i.e., that X^D is indeed a satisfying assignment of ϕ.

These observations prove the reverse direction of our main claim, and hence conclude the proof of Theorem 1. □

3 Proof of Theorem 2

We can repeat essentially the same proof as for the directed case, with the exception that we associate with every literal occurrence $\ell \in L$ a different subgraph denoted by E_ℓ: We now associate with $\ell \in L$ six nodes, $a = a(\ell), b = b(\ell), c = c(\ell), d = d(\ell), e = e(\ell)$, and $f = f(\ell)$, and the following ten edges:

$$E_\ell = \big\{(a,b), (b,c), (c,d), (d,e), (e,f), (a,f), (a,p), (b,q), (d,r), (e,s)\big\},$$

where nodes $p = p(\ell), q = q(\ell), r = r(\ell)$, and $s = s(\ell)$ are identified with the other nodes of G, in the same way as in the previous proof. To simplify notation, we omit the reference to ℓ whenever it is clear from the context which literal occurrence we are talking about. The weights of the edges of E_ℓ are defined as

$$w(a,p) = w(b,q) = w(d,r) = w(e,s) = \frac{5}{2}, \quad \text{and}$$
$$w(a,b) = w(b,c) = w(c,d) = w(d,e) = w(e,f) = w(a,f) = -1.$$

Note that in each of these subgraphs there is a negative circuit (see Fig. 4), formed by the six edges $D_\ell = \{(a,b), (b,c), (c,d), (d,e), (e,f), (a,f)\}$. We denote by $\mathcal{X} = \{D_\ell \mid \ell \in L\}$ the collection of these negative circuits, and let $\mathcal{F}$ denote the family of all negative circuits in G.

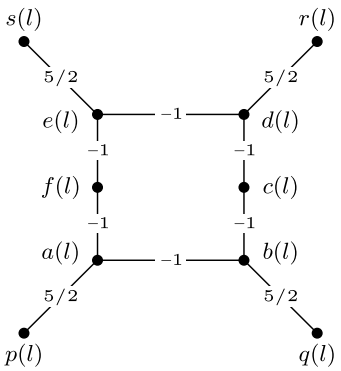


Fig. 4 The undirected graph associated with literal occurrences

By an analogous proof as in the previous section, we can show that there exists a negative circuit belonging to $\mathcal{F}\backslash\mathcal{X}$ if and only if ϕ has a satisfying assignment. The key observation in this case, the analogue of Lemma 1, is the following claim, which can easily be verified, e.g., by looking at Fig. 4.

Lemma 2 *For a circuit D of G not belonging to $\mathcal{X}$ and literal occurrence $\ell \in L$ we have*

$$w(D \cap E_\ell) \in \{0, 1, 2, 3, 4\}$$

and it is equal to 0 *only if $D \cap E_\ell$ is one of the following three sets*: $\emptyset$,

$$E^v_\ell = \big\{(b, c), (c, d), (d, e), (e, f), (a, f), (a, p), (b, q)\big\}, \quad \text{or}$$

$$E^c_\ell = \big\{(a, b), (b, c), (c, d), (e, f), (a, f), (d, r), (e, s)\big\}.$$

Remark The construction in Theorem 1 can be slightly modified to show that the NP-hardness result of Corollary 2 applies to polyhedra with 0/1-vertices (see arXiv:0801.3790v1 for more details).

Acknowledgements We thank the anonymous referees for their helpful remarks.

References

1. Abdullahi, S.D.: Vertex enumeration and counting for certain classes of polyhedra. Ph.D. thesis, Computing (Computer Algorithms), Leeds University (2003)
2. Avis, D., Bremner, B., Seidel, R.: How good are convex hull algorithms. Comput. Geom. Theory Appl. **7**, 265–302 (1997)
3. Avis, D., Fukuda, K.: A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra. Discrete Comput. Geom. **8**(3), 295–313 (1992)
4. Avis, D., Fukudam, K.: Reverse search for enumeration. Discrete Appl. Math. **65**(1–3), 21–46 (1996)
5. Amaldi, E., Pfetsch, M.E., Trotter, L.E.: On the maximum feasible subsystem problem, IISs and IIS-hypergraphs. Math. Program. **95**, 533–554 (2003)
6. Balinski, M.L.: An algorithm for finding all vertices of convex polyhedral sets. SIAM J. Appl. Math. **9**, 72–81 (1961)
7. Boros, E., Elbassioni, K., Gurvich, V., Khachiyan, L.: Enumerating minimal dicuts and strongly connected subgraphs and related geometric problems. In: Bienstock, D., Nemhauser, G. (eds.) Integer Programming and Combinatorial Optimization, 10th International IPCO Conference. Lecture Notes in Computer Science, vol. 3064, pp. 152–162. Springer, Berlin (2004). (An extended version is to appear in Algorithmica)
8. Berge, C.: Hypergraphs. Elsevier-North Holland, Amsterdam (1989)
9. Bremner, D., Fukuda, K., Marzetta, A.: Primal–dual methods for vertex and facet enumeration. Discrete Comput. Geom. **20**, 333–357 (1998)
10. Bussieck, M.R., Lübbecke, M.E.: The vertex set of a $0/1$ polytope is strongly $\mathcal{P}$-enumerable. Comput. Geom. Theory Appl. **11**(2), 103–109 (1998)
11. Bremner, D.: Incremental convex hull algorothms are not output sensitive. Discrete Comput. Geom. **21**, 57–68 (1999)
12. Charnes, A., Cooper, W.W., Henderson, A.: An Introduction to Linear Programming. Wiley, New York (1953)
13. Chazelle, B.: An optimal convex hull algorithm in any fixed dimension. Discrete Comput. Geom. **10**, 377–409 (1993)
14. Chakravarti, N.: Some results concerning post-infeasibility analysis. Eur. J. Oper. Res. **73**, 139–143 (1994)
15. Chvátal, V.: Linear Programming. Freeman, San Francisco (1983)

16. Chand, D.R., Kapur, S.S.: An algorithm for convex polytopes. J. Assoc. Comput. Mach. **17**(1), 78–86 (1970)
17. Cook, S.A.: The complexity of theorem proving procedures. In: Proceedings of the Third Annual ACM Symposium on Theory of Computing, pp. 151–158 (1971)
18. Dyer, M.E.: The complexity of vertex enumeration methods. Math. Oper. Res. **8**, 381–402 (1983)
19. Dyer, M.E., Proll, L.G.: An algorithm for determining all extreme points of a convex polytope. Math. Program. **12**, 81–96 (1977)
20. Eiter, T., Gottlob, G.: Identifying the minimal transversals of a hypergraph and related problems. SIAM J. Comput. **24**, 1278–1304 (1995)
21. Farkas, J.: Theorie der einfachen ungleichungen. J. Rein. Angew. Math. **124**, 1–27 (1901)
22. Fredman, M., Khachiyan, L.: On the complexity of dualization of monotone disjunctive normal forms. J. Algorithms **21**, 618–628 (1996)
23. Floyd, R.W.: Algorithm 97: Shortest path. Commun. Assoc. Comput. Mach. **5**, 345 (1962)
24. Fukuda, K., Liebling, Th.M., Margot, F.: Analysis of backtrack algorithms for listing all vertices and all faces of a convex polyhedron. CGTA **8**, 1–12 (1997)
25. Gallai, T.: Maximum-minimum Sätze über Graphen. Acta Math. Acad. Sci. Hung. **9**, 395–434 (1958)
26. Gleeson, J., Ryan, J.: Identifying minimally infeasible subsystems of inequalities. ORSA J. Comput. **2**(1), 61–63 (1990)
27. Johnson, D.S., Preparata, F.P.: The densest hemisphere problem. Theor. Comput. Sci. **6**, 93–107 (1978)
28. Johnson, D.S., Papadimitriou, Ch.H.: On generating all maximal independent sets. Inf. Process. Lett. **27**, 119–123 (1988)
29. Jerrum, M.R., Valiant, L.G., Vazirani, V.V.: Random generation of combinatorial structures from a uniform distribution. Theor. Comput. Sci. **44**, 169–188 (1986)
30. Khachiyan, L.: A polynomial algorithm in linear programming. Sov. Math. Dokl. **20**, 191–194 (1979)
31. Lovász, L.: Combinatorial optimization: some problems and trends. DIMACS Technical Report 92-53, Rutgers University (1992)
32. Lawler, E., Lenstra, J.K., Rinnooy Kan, A.H.G.: Generating all maximal independent sets: NP-hardness and polynomial-time algorithms. SIAM J. Comput. **9**, 558–565 (1980)
33. Mattheiss, T.H.: An algorithm for determining irrelevant constraints and all vertices in systems of linear inequalities. Oper. Res. **21**, 247–260 (1973)
34. Motzkin, T.S., Raiffa, H., Thompson, G.L., Thrall, R.M.: The double description method. In: H.W. Kuhn and A.W. Tucker (eds.) Contributions to the Theory of Games, vol. II, pp. 51–73 (1953)
35. Pfetsch, M.E.: The maximum feasible subsystem problem and vertex-facet incidences of polyhedra. Dissertation, TU Berlin (2002)
36. Provan, J.S.: Efficient enumeration of the vertices of polyhedra associated with network lp's. Math. Program. **63**(1), 47–64 (1994)
37. Read, R.C., Tarjan, R.E.: Bounds on backtrack algorithms for listing cycles, paths, and spanning trees. Networks **5**, 237–252 (1975)
38. Ryan, J.: IIS-hypergraphs. SIAM J. Discrete Math. **9**(4), 643–653 (1996)
39. Schrijver, A.: Theory of Linear and Integer Programming. Wiley, New York (1986)
40. Schrijver, A.: Combinatorial Optimization: Polyhedra and Efficiency, vol. A. Springer, Berlin (2003)
41. Seidel, R.: Output-size sensitive algorithms for constructive problems in computational geometry. Computer Science. Cornell University, Ithaka (1986)
42. Swart, G.: Finding the convex hull facet by facet. J. Algorithms **6**, 17–48 (1985)
43. Valiant, L.G.: The complexity of enumeration and reliability problems. SIAM J. Comput. **8**, 410–421 (1979)
44. Warshall, S.: A theorem on boolean matrices. J. Assoc. Comput. Mach. **9**, 11–12 (1962)

Pure Point Diffractive Substitution Delone Sets Have the Meyer Property

Jeong-Yup Lee · Boris Solomyak

Abstract We prove that a primitive substitution Delone set, which is pure point diffractive, is a Meyer set. This answers a question of J.C. Lagarias. We also show that for primitive substitution Delone sets, being a Meyer set is equivalent to having a relatively dense set of Bragg peaks. The proof is based on tiling dynamical systems and the connection between the diffraction and dynamical spectra.

1 Introduction

The discovery of quasicrystals in the 1980s inspired a lot of research in the area of "aperiodic order" and "mathematical quasicrystals." Roughly speaking, physical quasicrystals are aperiodic structures which exhibit sharp bright spots (called Bragg peaks) in their X-ray diffraction pattern. The presence of Bragg peaks indicates the presence of "long-range order" in the structure. A mathematical idealization of a large set of atoms is a discrete set in $\mathbb{R}^d$. The most general class of sets modeling solids is the class of *Delone sets*, that is, subsets of $\mathbb{R}^d$ which are relatively dense and uniformly discrete. Usually some additional assumptions are made. A Delone set Λ

The first author acknowledges support from the NSERC post-doctoral fellowship and thanks the University of Washington and the University of Victoria for being the host universities of the fellowship. The second author is grateful to the Weizmann Institute of Science where he was a Rosi and Max Varon Visiting Professor when this work was completed. He was also supported in part by NSF Grant DMS 0355187.

J.-Y. Lee (✉)
Department of Mathematics and Statistics, University of Victoria, P.O. Box 3045 STN CSC, Victoria, British Columbia, V8W 3P4, Canada
e-mail: jylee@math.uvic.ca

B. Solomyak
Department of Mathematics, University of Washington, P.O. Box 354350, Seattle, WA 98195, USA
e-mail: solomyak@math.washington.edu

is of *finite local complexity* (or "finite type") if $\Lambda - \Lambda$ is closed and discrete, which is equivalent to having finitely many local patterns, up to translations, see [12]. Another common assumption is *repetitivity*, which means that every pattern of a Delone set (and not just individual points) occurs relatively densely in space. This is still not enough for long-range order, since a repetitive Delone set of finite local complexity may fail to have any Bragg peaks. The Delone set Λ is said to be a *Meyer set* if $\Lambda - \Lambda$ is uniformly discrete. Meyer sets were introduced (under the name of "harmonious sets") in 1969–1970 by Meyer [19] in the context of harmonic analysis. In the last 10 years their importance in the theory of long-range aperiodic order has been revealed in many investigations, see, e.g., [20], [17], [15], and [2].

The mathematical concept of diffraction spectrum is based on the Fourier transform of the autocorrelation measure, see [7] and [8]. Under certain conditions, this Fourier transform is a measure (called *diffraction measure*) on $\mathbb{R}^d$, whose discrete component corresponds to the Bragg peaks. A Delone set Λ is said to be *pure point diffractive* (or "perfectly diffractive," or a "Patterson set" [13]) if the diffraction measure is pure point (pure discrete). There is another notion of spectrum, which comes from Ergodic Theory via a dynamical system associated with the Delone set. As shown by Dworkin [4] (see also [16], [6], and [1]), there is a close connection between the two notions of spectra.

In his survey on mathematical quasicrystals, Lagarias raised the following problem [13, Problem 4.10]. *Let Λ be a Delone set of finite type which is repetitive. If Λ is pure point diffractive, must it be a Meyer set?* We do not have an answer for this question, but we solve the following special case:

[13, Problem 4.11]. *Suppose that Λ is a primitive self-replicating Delone set of finite type. If Λ is pure point diffractive, must Λ be a Meyer set?*

At this point, we just mention that a primitive self-replicating Delone set, roughly speaking, corresponds to the set of "control points" of a self-affine tiling. In this paper we refer to it as a representable primitive substitution Delone set. Precise definitions on representable primitive substitution Delone sets are given in the next section.

Our main result, Theorem 4.11, answers this question affirmatively. This result is applicable to [17] and [15] in which the Meyer condition is additionally assumed to understand the structure of pure point diffractive point sets.

In fact, the condition of being pure point diffractive may be weakened. We only need the fact that the set of Bragg peaks is relatively dense in the entire space (this holds in the case of a pure point diffractive set). This condition turns out to be necessary and sufficient for the Meyer property on the class of substitution Delone sets (see Theorem 4.14).

The proof of the implication (for substitution Delone sets)

$$\text{relatively dense set of Bragg peaks} \;\Rightarrow\; \text{Meyer set}$$

relies on the theory of tiling dynamical systems developed in [23] and the connection between substitution Delone sets, substitution Delone set families, and self-affine tilings, studied in [14] and [17]. The second key ingredient is a generalization of classical results by Pisot in Diophantine approximation, due to Környei [11] and

Mauduit [18]. The relevance of PV-numbers (Pisot–Vijayaraghavan numbers) for the Meyer set property was already pointed out by Meyer [19]. We show that the expanding linear map associated with our substitution Delone set satisfies the "Pisot family" condition (this is essentially proved in [22] based on [23]), and we obtain some extra information about the set of translation vectors between tiles of the same type. The last ingredient is a generalization of the well-known "Garsia Lemma" [5, Lemma 1.51] (obtained independently by other authors as well), which implies that the set of polynomials of arbitrary degree with integer coefficients bounded by a uniform constant, evaluated at a PV-number, yields a uniformly discrete set.

Now we can state our main result.

Theorem 1.1 *If Λ is a representable primitive substitution Delone set of finite local complexity (FLC) such that the Bragg peaks are relatively dense in $\mathbb{R}^d$, then Λ is a Meyer set.*

We note that the converse is also true by a theorem of Strungaru [24]: if Λ is a Meyer set, then the Bragg peaks are relatively dense.

Corollary 1.2 *If Λ is a representable primitive substitution Delone set of FLC which is pure point diffractive, then Λ is a Meyer set.*

This resolves Problem 4.11 of [13] (it follows from the context of [13] that FLC is implicitly assumed).

2 Preliminaries

2.1 Substitution Delone Multisets and Tilings

A *multiset*[1] or *m-multiset* in $\mathbb{R}^d$ is a subset $\boldsymbol{\Lambda} = \Lambda_1 \times \cdots \times \Lambda_m \subset \mathbb{R}^d \times \cdots \times \mathbb{R}^d$ (m copies) where $\Lambda_i \subset \mathbb{R}^d$. We also write $\boldsymbol{\Lambda} = (\Lambda_1, \ldots, \Lambda_m) = (\Lambda_i)_{i \le m}$. Recall that a Delone set is a relatively dense and uniformly discrete subset of $\mathbb{R}^d$. We say that $\boldsymbol{\Lambda} = (\Lambda_i)_{i \le m}$ is a *Delone multiset* in $\mathbb{R}^d$ if each Λ_i is Delone and $\operatorname{supp}(\boldsymbol{\Lambda}) := \bigcup_{i=1}^m \Lambda_i \subset \mathbb{R}^d$ is Delone.

Although $\boldsymbol{\Lambda}$ is a product of sets, it is convenient to think of it as a set with types or colors, i being the color of points in Λ_i. A *cluster* of $\boldsymbol{\Lambda}$ is, by definition, a family $\mathbf{P} = (P_i)_{i \le m}$ where $P_i \subset \Lambda_i$ is finite for all $i \le m$. For a bounded set $A \subset \mathbb{R}^d$, let $A \cap \boldsymbol{\Lambda} := (A \cap \Lambda_i)_{i \le m}$. There is a natural translation $\mathbb{R}^d$-action on the set of Delone multisets and their clusters in $\mathbb{R}^d$. The translate of a cluster $\mathbf{P}$ by $x \in \mathbb{R}^d$ is $x + \mathbf{P} = (x + P_i)_{i \le m}$. We say that two clusters $\mathbf{P}$ and $\mathbf{P}'$ are translationally equivalent if $\mathbf{P} = x + \mathbf{P}'$, i.e., $P_i = x + P_i'$ for all $i \le m$, for some $x \in \mathbb{R}^d$. We write $B_R(y)$ for the closed ball of radius R centered at y.

[1] Caution: In [14] the word multiset refers to a set with multiplicities.

Definition 2.1 A Delone multiset $\mathbf{\Lambda}$ has *finite local complexity (FLC)* if for every $R > 0$ there exists a finite set $Y \subset \operatorname{supp}(\mathbf{\Lambda}) = \bigcup_{i=1}^{m} \Lambda_i$ such that

$$\forall x \in \operatorname{supp}(\mathbf{\Lambda}), \quad \exists\, y \in Y, \qquad B_R(x) \cap \mathbf{\Lambda} = (B_R(y) \cap \mathbf{\Lambda}) + (x - y).$$

In plain language, for each radius $R > 0$ there are only finitely many translational classes of clusters whose support lies in some ball of radius R.

Definition 2.2 A Delone set Λ is called a *Meyer set* if $\Lambda - \Lambda$ is uniformly discrete.

For a cluster $\mathbf{P}$ and a bounded set $A \subset \mathbb{R}^d$ denote

$$L_{\mathbf{P}}(A) = \sharp\{x \in \mathbb{R}^d \colon\, x + \mathbf{P} \subset A \cap \mathbf{\Lambda}\},$$

where $\sharp$ means the cardinality. In plain language, $L_{\mathbf{P}}(A)$ is the number of translates of $\mathbf{P}$ contained in A, which is clearly finite. For a bounded set $F \subset \mathbb{R}^d$ and $r > 0$, let $(F)^{+r} := \{x \in \mathbb{R}^d \colon \operatorname{dist}(x, F) \le r\}$ denote the r-neighborhood of F. A *van Hove sequence* for $\mathbb{R}^d$ is a sequence $\mathcal{F} = \{F_n\}_{n \ge 1}$ of bounded measurable subsets of $\mathbb{R}^d$ satisfying

$$\lim_{n \to \infty} 0((\partial F_n)^{+r})/0(F_n) = 0, \qquad \text{for all} \quad r > 0. \tag{2.1}$$

Definition 2.3 Let $\{F_n\}_{n \ge 1}$ be a van Hove sequence. The Delone multiset $\mathbf{\Lambda}$ has *uniform cluster frequencies* (UCF) (relative to $\{F_n\}_{n \ge 1}$) if for any nonempty cluster $\mathbf{P}$, the limit

$$\operatorname{freq}(\mathbf{P}, \mathbf{\Lambda}) = \lim_{n \to \infty} \frac{L_{\mathbf{P}}(x + F_n)}{0(F_n)} \ge 0$$

exists uniformly in $x \in \mathbb{R}^d$.

A linear map $Q \colon \mathbb{R}^d \to \mathbb{R}^d$ is *expansive* if its every eigenvalue lies outside the unit circle.

Definition 2.4 $\mathbf{\Lambda} = (\Lambda_i)_{i \le m}$ is called a *substitution Delone multiset* if $\mathbf{\Lambda}$ is a Delone multiset and there exist an expansive map $Q \colon \mathbb{R}^d \to \mathbb{R}^d$ and finite sets $\mathcal{D}_{ij}$ for $i, j \le m$ such that

$$\Lambda_i = \bigcup_{j=1}^{m} (Q\Lambda_j + \mathcal{D}_{ij}), \qquad i \le m, \tag{2.2}$$

where the unions on the right-hand side are disjoint.

For any given substitution Delone multiset $\mathbf{\Lambda} = (\Lambda_i)_{i \le m}$, we define $\Phi_{ij} = \{f \colon\, x \mapsto Qx + a \colon\, a \in \mathcal{D}_{ij}\}$. Then $\Phi_{ij}(\Lambda_j) = Q\Lambda_j + \mathcal{D}_{ij}$, where $i \le m$. We define Φ as an $m \times m$ array for which each entry is Φ_{ij}, and call Φ a *matrix function system (MFS)* for the substitution. For any $k \in \mathbb{Z}_+$ and $x \in \Lambda_j$ with $j \le m$, we let $\Phi^k(x) = \Phi^{k-1}((\Phi_{ij}(x))_{i \le m})$.

We say that the substitution Delone multiset $\mathbf{\Lambda}$ is *primitive* if the corresponding substitution matrix S, with $S_{ij} = \sharp(\mathcal{D}_{ij})$, is primitive, i.e., there is an $l > 0$ for which S^l has no zero entries.

We say that a Delone set Λ is a *substitution Delone set* if there is a substitution Delone multiset $\mathbf{\Lambda} = (\Lambda_i)_{i \le m}$ such that $\Lambda = \bigcup_{i=1}^m \Lambda_i$. The Delone set Λ is said to be primitive if the substitution Delone multiset $\mathbf{\Lambda}$ can be chosen primitive.

Next we briefly review the basic definitions of tilings and substitution tilings. We begin with a set of types (or colors) $\{1, \dots, m\}$, which we fix once and for all. A *tile* in $\mathbb{R}^d$ is defined as a pair $T = (A, i)$ where $A = \mathrm{supp}(T)$ (the support of T) is a compact set in $\mathbb{R}^d$ which is the closure of its interior, and $i = l(T) \in \{1, \dots, m\}$ is the type of T. We let $g + T = (g + A, i)$ for $g \in \mathbb{R}^d$. We say that a set P of tiles is a *patch* if the number of tiles in P is finite and the tiles of P have mutually disjoint interiors (strictly speaking, we have to say "supports of tiles," but this abuse of language should not lead to confusion). A tiling of $\mathbb{R}^d$ is a set $\mathcal{T}$ of tiles such that $\mathbb{R}^d = \bigcup\{\mathrm{supp}(T)\colon\ T \in \mathcal{T}\}$ and distinct tiles have disjoint interiors. Given a tiling $\mathcal{T}$, finite sets of tiles of $\mathcal{T}$ are called $\mathcal{T}$-patches.

We define FLC and UCF for tilings in the same way as the corresponding properties for Delone multisets.

We always assume that any two $\mathcal{T}$-tiles with the same color are translationally equivalent. (Hence there are finitely many $\mathcal{T}$-tiles up to translation.)

Definition 2.5 Let $\mathcal{A} = \{T_1, \dots, T_m\}$ be a finite set of tiles in $\mathbb{R}^d$ such that $T_i = (A_i, i)$; we call them *prototiles*. Denote by $\mathcal{P}_\mathcal{A}$ the set of patches made of tiles each of which is a translate of one of T_i's. We say that $\omega\colon\ \mathcal{A} \to \mathcal{P}_\mathcal{A}$ is a *tile-substitution* (or simply *substitution*) with expansive map Q if there exist finite sets $\mathcal{D}_{ij} \subset \mathbb{R}^d$ for $i, j \le m$, such that

$$\omega(T_j) = \{u + T_i\colon\ u \in \mathcal{D}_{ij},\ i = 1, \dots, m\} \qquad \text{for} \quad j \le m, \tag{2.3}$$

with

$$QA_j = \bigcup_{i=1}^m (\mathcal{D}_{ij} + A_i).$$

Here all sets in the right-hand side must have disjoint interiors; it is possible for some of the $\mathcal{D}_{ij}$ to be empty.

The substitution (2.3) is extended to all translates of prototiles by $\omega(x + T_j) = Qx + \omega(T_j)$, and to patches and tilings by $\omega(P) = \bigcup\{\omega(T)\colon\ T \in P\}$. The substitution ω can be iterated, producing larger and larger patches $\omega^k(T_j)$. To the substitution ω we associate its $m \times m$ substitution matrix S, with $S_{ij} := \sharp(\mathcal{D}_{ij})$. The substitution ω is called *primitive* if the substitution matrix S is primitive. We say that $\mathcal{T}$ is a fixed point of a substitution if $\omega(\mathcal{T}) = \mathcal{T}$.

For each primitive substitution Delone multiset $\mathbf{\Lambda}$ (2.2) one can set up an *adjoint system* of equations

$$QA_j = \bigcup_{i=1}^m (\mathcal{D}_{ij} + A_i), \qquad j \le m. \tag{2.4}$$

From Hutchinson's Theory (or rather, its generalization to the "graph-directed" setting), it follows that (2.4) always has a unique solution for which $\mathcal{A} = \{A_1, \ldots, A_m\}$ is a family of nonempty compact sets of $\mathbb{R}^d$ (see for example Proposition 1.3 of [3]). It is proved in Theorems 2.4 and 5.5 of [14] that if $\mathbf{\Lambda}$ is a primitive substitution Delone multiset, then all the sets A_i from (2.4) have nonempty interiors and, moreover, each A_i is the closure of its interior.

Definition 2.6 A Delone multiset $\mathbf{\Lambda} = (\Lambda_i)_{i \le m}$ is called *representable* (by tiles) for a tiling if there exists a set of prototiles $\mathcal{A} = \{T_i \colon i \le m\}$ so that

$$\mathbf{\Lambda} + \mathcal{A} := \{x + T_i \colon x \in \Lambda_i,\ i \le m\} \qquad \text{is a tiling of } \mathbb{R}^d, \tag{2.5}$$

that is, $\mathbb{R}^d = \bigcup_{i \le m} \bigcup_{x \in \Lambda_i} (x + A_i)$ where $T_i = (A_i, i)$ for $i \le m$, and the sets in this union have disjoint interiors. In the case that $\mathbf{\Lambda}$ is a primitive substitution Delone multiset we understand the term representable to mean relative to the tiles $T_i = (A_i, i)$, for $i \le m$, arising from the solution to the adjoint system (2.4). We call $\mathbf{\Lambda} + \mathcal{A}$ the associated tiling of $\mathbf{\Lambda}$.

Definition 2.7 Let $\mathbf{\Lambda}$ be a primitive substitution Delone multiset and let $\mathbf{P}$ be a cluster of $\mathbf{\Lambda}$. The cluster $\mathbf{P}$ is called *legal* if it is a translate of a subcluster of $\Phi^k(x_j)$ for some $x_j \in \Lambda_j$, $j \le m$, and $k \in \mathbb{Z}_+$.

Lemma 2.8 [17] *Let $\mathbf{\Lambda}$ be a primitive substitution Delone multiset such that every $\mathbf{\Lambda}$-cluster is legal. Then $\mathbf{\Lambda}$ is repetitive.*

Not every substitution Delone multiset is representable (see Exercise 3.12 of [17]), but the following theorem provides the sufficient condition for it.

Theorem 2.9 [17] *Let $\mathbf{\Lambda}$ be a repetitive primitive substitution Delone multiset. Then every $\mathbf{\Lambda}$-cluster is legal if and only if $\mathbf{\Lambda}$ is representable.*

Remark 2.10 In Lemma 3.2 of [14] it is shown that if $\mathbf{\Lambda}$ is a substitution Delone multiset, then there is a finite multiset (cluster) $\mathbf{P} \subset \mathbf{\Lambda}$ for which $\Phi^{n-1}(\mathbf{P}) \subset \Phi^n(\mathbf{P})$ for $n \ge 1$ and $\mathbf{\Lambda} = \lim_{n \to \infty} \Phi^n(\mathbf{P})$. We call such a multiset $\mathbf{P}$ a *generating multiset*. Note that, in order to check that every $\mathbf{\Lambda}$-cluster is legal, we only need to see if some cluster that contains a finite generating multiset for $\mathbf{\Lambda}$ is legal.

Let $\Xi(\mathcal{T})$ be the set of translation vectors between $\mathcal{T}$-tiles of the same type:

$$\Xi(\mathcal{T}) := \{x \in \mathbb{R}^d \colon \exists\, T, T' \in \mathcal{T},\ T' = x + T\}. \tag{2.6}$$

Since $\mathcal{T}$ has the inflation symmetry with the expansive map Q, we have that $Q\,\Xi(\mathcal{T}) \subset \Xi(\mathcal{T})$.

Remark 2.11 We should be careful to distinguish between substitution Delone *multisets* and substitution Delone *sets*. Lagarias [13] considers the latter under the name of *self-replicating sets*. Note that a substitution Delone set may arise from different substitution Delone multisets.

2.2 Diffraction and Dynamical Spectra on Delone Sets

We use the mathematical concept of diffraction measure developed by Hof [7], [8]. Given a translation-bounded measure ν on $\mathbb{R}^d$, let $\gamma(\nu)$ denote its autocorrelation (assuming it is unique), that is, the vague limit

$$\gamma(\nu) = \lim_{n\to\infty} \frac{1}{0(F_n)} \left(\nu|_{F_n} * \widetilde{\nu}|_{F_n}\right), \tag{2.7}$$

where $\{F_n\}_{n\geq 1}$ is a van Hove sequence.[2] The measure $\gamma(\nu)$ is positive definite, so by Bochner's theorem the Fourier transform $\widehat{\gamma(\nu)}$ is a positive measure on $\mathbb{R}^d$, called the *diffraction measure* for ν. We say that the measure ν has a *pure point diffraction spectrum*, if $\widehat{\gamma(\nu)}$ is a pure point or discrete measure. The point masses of the diffraction measure are called *Bragg peaks*. For a Delone set Λ let

$$\delta_\Lambda := \sum_{x\in\Lambda} \delta_x.$$

It is known that if Λ is a primitive substitution Delone set of finite local complexity, then δ_Λ has a unique autocorrelation measure $\gamma(\delta_\Lambda)$ (see [17]). We say that Λ is *pure point diffractive* if the diffraction measure $\widehat{\gamma(\delta_\Lambda)}$ is pure discrete.

Let $\mathbf{\Lambda}$ be a Delone multiset and let $X_{\mathbf{\Lambda}}$ be the collection of all Delone multisets each of whose clusters is a translate of a $\mathbf{\Lambda}$-cluster. We introduce a metric on Delone multisets in a simple variation of the standard way: for Delone multisets $\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_2 \in X_{\mathbf{\Lambda}}$,

$$d(\mathbf{\Lambda}_1, \mathbf{\Lambda}_2) := \min\{\tilde{d}(\mathbf{\Lambda}_1, \mathbf{\Lambda}_2), 2^{-1/2}\}, \tag{2.8}$$

where

$$\tilde{d}(\mathbf{\Lambda}_1, \mathbf{\Lambda}_2) = \inf\{\varepsilon > 0\colon\ \exists\, x, y \in B_\varepsilon(0),$$
$$B_{1/\varepsilon}(0) \cap (-x + \mathbf{\Lambda}_1) = B_{1/\varepsilon}(0) \cap (-y + \mathbf{\Lambda}_2)\}.$$

For the proof that d is a metric, see [16].

Observe that $X_{\mathbf{\Lambda}} = \overline{\{-h + \mathbf{\Lambda}\colon\ h \in \mathbb{R}^d\}}$ where the closure is taken in the topology induced by the metric d. The group $\mathbb{R}^d$ acts on $X_{\mathbf{\Lambda}}$ by translations which are obviously homeomorphisms, and we get a topological dynamical system $(X_{\mathbf{\Lambda}}, \mathbb{R}^d)$.

Let μ be an ergodic invariant Borel probability measure for the dynamical system $(X_{\mathbf{\Lambda}}, \mathbb{R}^d)$. We consider the associated group of unitary operators $\{U_g\}_{g\in\mathbb{R}^d}$ on $L^2(X_{\mathbf{\Lambda}}, \mu)$:

$$U_g f(\mathcal{S}) = f(-g + \mathcal{S}).$$

[2] Recall that if f is a function in $\mathbb{R}^d$, then $\tilde{f}$ is defined by $\tilde{f}(x) = \overline{f(-x)}$. If μ is a measure, $\tilde{\mu}$ is defined by $\tilde{\mu}(f) = \overline{\mu(\tilde{f})}$ for all $f \in \mathcal{C}_0(\mathbb{R}^d)$.

A vector $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{R}^d$ is said to be an eigenvalue for the $\mathbb{R}^d$-action if there exists an eigenfunction $f \in L^2(X_{\mathbf{\Lambda}}, \mu)$, that is, $f \not\equiv 0$ and

$$U_g f = e^{2\pi i g\cdot\alpha} f, \qquad \text{for all} \quad g \in \mathbb{R}^d.$$

The dynamical system $(X_{\mathbf{\Lambda}}, \mu, \mathbb{R}^d)$ is said to have a *pure discrete* (or pure point) *spectrum* if the linear span of the eigenfunctions is dense in $L^2(X_{\mathbf{\Lambda}}, \mu)$.

Let $X_{\mathcal{T}} = \overline{\{-g + \mathcal{T}\colon\ g \in \mathbb{R}^d\}}$, where $X_{\mathcal{T}}$ carries a well-known topology, given analogously to (2.8) for $X_{\mathbf{\Lambda}}$, relative to which it is compact (equivalent to FLC). We have a natural action of $\mathbb{R}^d$ on $X_{\mathcal{T}}$ which makes it a topological dynamical system. The set $\{-g + \mathcal{T}\colon\ g \in \mathbb{R}^d\}$ is the orbit of $\mathcal{T}$.

Recall that a topological dynamical system is *uniquely ergodic* if there is a unique invariant probability measure (which is then automatically ergodic). It is known (see, e.g., Theorem 2.7 of [16]) that for a Delone multiset $\mathbf{\Lambda}$ with FLC, the dynamical system $(X_{\mathbf{\Lambda}}, \mathbb{R}^d)$ is uniquely ergodic if and only if $\mathbf{\Lambda}$ has UCF.

Theorem 2.12 [16, Theorem 3.2] *Suppose that a Delone multiset* $\mathbf{\Lambda}$ *has FLC and UCF. Then the following are equivalent*:

(i) $\mathbf{\Lambda}$ *has a pure point dynamical spectrum.*
(ii) *The measure* $\nu = \sum_{i\le m} a_i \delta_{\Lambda_i}$ *has a pure point diffraction spectrum, for any choice of complex numbers* $(a_i)_{i\le m}$.
(iii) *The measures* δ_{Λ_i} *have pure point diffraction spectra, for* $i \le m$.

3 Jordan Canonical Form

Let Q be a linear map from $\mathbb{R}^d$ to $\mathbb{R}^d$. We can consider Q as a $(d \times d)$ matrix. We discuss the matrix analysis on Q that we use in this paper (see [9]). The matrix Q is similar to a matrix in the Jordan canonical form J, so that $Q = SJS^{-1}$ for some invertible matrix S over $\mathbb{C}$. Suppose that Q has r distinct eigenvalues $\lambda_1, \ldots, \lambda_r \in \mathbb{C}$. For each eigenvalue $\lambda_i, 1 \le i \le r$, there are Jordan blocks $J_{i1}(\lambda_i), \ldots, J_{im_i}(\lambda_i)$ corresponding to λ_i. We simply write J_{ij} for $J_{ij}(\lambda_i)$. We can decompose $J_{ij} = \lambda_i I + N$ with a matrix $\lambda_i I$ of diagonal entries and a matrix N of off-diagonal entries. For each Jordan block $J_{ij}, 1 \le j \le m_i$, we have vectors $e_{ij1}, \ldots, e_{ijk_{ij}} \in \mathbb{C}^d$ such that

$$Qe_{ij1} = \lambda_i e_{ij1} \quad \text{and} \quad Qe_{ijl} = e_{ij(l-1)} + \lambda_i e_{ijl} \qquad \text{for} \quad 2 \le l \le k_{ij}.$$

For each Jordan block J_{ij} and any $n \in \mathbb{Z}_+$, there is a simple general formula for $(J_{ij})^n$:

$$(J_{ij})^n = (\lambda_i I + N)^n = \sum_{k=0}^{n} \binom{n}{k} \lambda_i^{n-k} N^k.$$

We define $\binom{n}{k} = 0$ for $n < k$. Then for any $n \in \mathbb{Z}_+$,

$$(J_{ij})^n = \begin{bmatrix} \lambda_i^n & \binom{n}{1}\lambda_i^{n-1} & \binom{n}{2}\lambda_i^{n-2} & \cdots & \binom{n}{k_{ij}-1}\lambda_i^{n-k_{ij}+1} \\ 0 & \lambda_i^n & \binom{n}{1}\lambda_i^{n-1} & \cdots & \vdots \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & \ddots & & \binom{n}{2}\lambda_i^{n-2} \\ 0 & 0 & & \ddots & \binom{n}{1}\lambda_i^{n-1} \\ 0 & 0 & \cdots & \cdots & \lambda_i^n \end{bmatrix}$$

Note that $E := \{e_{ijl} \in \mathbb{C}^d\colon\ 1 \le i \le r, 1 \le j \le m_i, 1 \le l \le k_{ij}\}$ is a basis of $\mathbb{C}^d$. So for any $y \in \mathbb{R}^d$, we can write

$$y = \sum_{i=1}^{r}\sum_{j=1}^{m_i}\sum_{l=1}^{k_{ij}} a_{ijl}(y)e_{ijl}, \tag{3.1}$$

where $a_{ijl}(y) \in \mathbb{C}$.

Let $\langle x, y\rangle$ be the standard inner product of x, y in $\mathbb{C}^d$ and let $K := \max\{k_{ij} - 1\colon\ 1 \le i \le r, 1 \le j \le m_i\}$.

Lemma 3.1 *Let $\alpha \in \mathbb{R}^d$ and $Q\colon\ \mathbb{R}^d \to \mathbb{R}^d$ be a linear map. For any $n \in \mathbb{Z}_+$ and $w \in \mathbb{R}^d$ for which $w = \sum_{i=1}^{r}\sum_{j=1}^{m_i}\sum_{l=1}^{k_{ij}} a_{ijl}(w)e_{ijl}$ with $a_{ijl}(w) \in \mathbb{C}$,*

$$\left\langle \sum_{j=1}^{m_i}\sum_{l=1}^{k_{ij}} a_{ijl}(w)Q^n e_{ijl}, \alpha \right\rangle = (P_{\alpha,w})_i(n)\lambda_i^n \qquad \text{for} \quad 1 \le i \le r$$

and so

$$\langle Q^n w, \alpha\rangle = \sum_{i=1}^{r}(P_{\alpha,w})_i(n)\lambda_i^n,$$

where $(P_{\alpha,w})_i$ is a polynomial over $\mathbb{C}$ of degree less than or equal to K.

Proof This is standard; we provide a proof for completeness.

We extend the linear map Q from $\mathbb{R}^d$ to $\mathbb{C}^d$, i.e., $Q\colon\ \mathbb{C}^d \to \mathbb{C}^d$ (just use the same matrix). First note that for any $1 \le i \le r$ and $1 \le j \le m_i$,

$$\begin{aligned} &\left\langle \sum_{l=1}^{k_{ij}} a_{ijl}(w)Q^n e_{ijl}, \alpha \right\rangle \\ &\quad = \langle a_{ij1}(w)\lambda_i^n e_{ij1}, \alpha\rangle \\ &\qquad + \left\langle a_{ij2}(w)\left(\binom{n}{1}\lambda_i^{n-1}e_{ij1} + \lambda_i^n e_{ij2}\right), \alpha \right\rangle \end{aligned}$$

$$\vdots$$
$$+\left\langle a_{ijk_{ij}}(w)\left(\binom{n}{k_{ij}-1}\lambda_i^{n-k_{ij}+1}e_{ij1}+\cdots+\lambda_i^n e_{ijk_{ij}}\right),\alpha\right\rangle.$$

Rearranging the above equation,

$$\begin{aligned}
&\left\langle\sum_{l=1}^{k_{ij}}a_{ijl}(w)Q^n e_{ijl},\alpha\right\rangle\\
&\quad=\left(a_{ij1}(w)\lambda_i^0+\cdots+a_{ijk_{ij}}(w)\binom{n}{k_{ij}-1}\lambda_i^{-k_{ij}+1}\right)\langle e_{ij1},\alpha\rangle\lambda_i^n\\
&\qquad+\left(a_{ij2}(w)\lambda_i^0+\cdots+a_{ijk_{ij}}(w)\binom{n}{k_{ij}-2}\lambda_i^{-k_{ij}+2}\right)\langle e_{ij2},\alpha\rangle\lambda_i^n\\
&\qquad\vdots\\
&\qquad+\left(a_{ijk_{ij}}(w)\lambda_i^0\right)\langle e_{ijk_{ij}},\alpha\rangle\lambda_i^n\,.
\end{aligned}$$

Thus we get

$$\left\langle\sum_{l=1}^{k_{ij}}a_{ijl}(w)Q^n e_{ijl},\alpha\right\rangle=(P_{\alpha,w})_{ij}(n)\lambda_i^n,$$

where $(P_{\alpha,w})_{ij}$ is a polynomial over $\mathbb{C}$ of degree at most $k_{ij}-1$. Then for each $1\le i\le r$, we can write

$$\left\langle\sum_{j=1}^{m_i}\sum_{l=1}^{k_{ij}}a_{ijl}(w)Q^n e_{ijl},\alpha\right\rangle=(P_{\alpha,w})_i(n)\lambda_i^n,\tag{3.2}$$

where $(P_{\alpha,w})_i=\sum_{j=1}^{m_i}(P_{\alpha,w})_{ij}$ is a polynomial over $\mathbb{C}$ of degree $\le K$. Furthermore,

$$\langle Q^n w,\alpha\rangle=\left\langle Q^n\left(\sum_{i=1}^{r}\sum_{j=1}^{m_i}\sum_{l=1}^{k_{ij}}a_{ijl}(w)e_{ijl}\right),\alpha\right\rangle=\sum_{i=1}^{r}(P_{\alpha,w})_i(n)\lambda_i^n.$$

□

4 Proof of the Meyer Property

The result of the following lemma is taken from [10].

Lemma 4.1 *Suppose that L is a finitely generated free Abelian group in $\mathbb{R}^d$ such that L spans $\mathbb{R}^d$ and $QL\subset L$ with a linear map Q. Then all eigenvalues of Q are algebraic integers.*

Proof Let $\{v_1,\ldots,v_n\}$ be a set of generators for L. Consider the $(d\times n)$ matrix $N=[v_1,\ldots,v_n]$. Since L spans $\mathbb{R}^d$, the rank of N is d. Thus $N^Tx=\mathbf{0}$ has a unique

trivial solution. From the assumption of $QL \subset L$, for each $1 \le i \le n$, we can write

$$Qv_i = \sum_{j=1}^{n} a_{ij} v_j \qquad \text{for some} \quad a_{ij} \in \mathbb{Z}.$$

Let $M = (a_{ij})_{n\times n}$. Then $QN = NM^T$ and so $MN^T = N^T Q^T$. For any eigenvalue λ of Q^T and the corresponding eigenvector x,

$$M(N^T x) = N^T(Q^T x) = N^T \lambda x = \lambda(N^T x).$$

Since x is nonzero, $N^T x$ is nonzero and so λ is an eigenvalue of M. Since M is an integer matrix, λ is an algebraic integer. Since Q^T and Q have the same eigenvalues, all eigenvalues of Q are algebraic integers. □

Corollary 4.2 *Suppose that $\mathcal{T}$ is a fixed point of a primitive substitution with expansive map Q which has FLC. Then all eigenvalues of Q are algebraic integers.*

Proof Let L be an Abelian group generated by $\Xi(\mathcal{T})$. Since $\mathcal{T}$ has FLC, L is a finitely generated free Abelian group. From $Q\Xi(\mathcal{T}) \subset \Xi(\mathcal{T})$ we have $QL \subset L$. By Lemma 4.1, all eigenvalues of Q are algebraic integers. □

The following is a generalization of Pisot's theorem, due to Környei [11]. A similar result was obtained by Mauduit [18]. The theorem is about two equivalent conditions, but we state only one direction which we use later, in the special case we need. For $x \in \mathbb{R}$, let $\|x\|$ denote the distance from x to the nearest integer.

Theorem 4.3 [11, Theorem 1] *Let $\lambda_1, \ldots, \lambda_r$ be distinct algebraic numbers such that $|\lambda_i| \ge 1$, $i = 1, \ldots, r$, and let $P_1, \ldots, P_r$ be nonzero polynomials with complex coefficients. If $\sum_{i=1}^{r} P_i(n)\lambda_i^n$ is real for all n and*

$$\lim_{n\to\infty} \left\| \sum_{i=1}^{r} P_i(n)\lambda_i^n \right\| = 0,$$

then the following assertions are true:

(a) *The coefficients of P_i are elements of the algebraic extension $\mathbb{Q}(\lambda_i)$.*
(b) *If λ_s and λ_t are conjugate elements over $\mathbb{Q}$, and the corresponding polynomials have the form*

$$P_s(x) = \sum_{k=0}^{K_s} c_{s,k} x^k, \qquad P_t(x) = \sum_{k=0}^{K_t} c_{t,k} x^k,$$

then P_s and P_t have the same degree, $c_{s,k}$ and $c_{t,k}$ are conjugate elements over $\mathbb{Q}$, and for any isomorphism τ which is the identical mapping on $\mathbb{Q}$ and for which $\tau(\lambda_s) = \lambda_t$, we have

$$\tau(c_{s,k}) = c_{t,k}, \qquad \text{for any} \quad 0 \le k \le K_s = K_t.$$

(c) *All the conjugates of the λ_i's not occurring in the sum $\sum_{i=1}^{r} P_i(n)\lambda_i^n$ have absolute value less than one. In other words, if λ' is a conjugate of λ_i for some $i \le r$ and $|\lambda'| \ge 1$, then $\lambda' = \lambda_j$ for some $j \le r$.*

Definition 4.4 [21] Let $\mathcal{T}$ be a fixed point of a primitive substitution with expansive map Q. For each $\mathcal{T}$-tile T, fix a tile γT in the patch $\omega(T)$; choose γT with the same relative position for all tiles of the same type. This defines a map $\gamma\colon \mathcal{T} \to \mathcal{T}$ called the *tile map*. Then define the *control point* for a tile $T \in \mathcal{T}$ by

$$\{c(T)\} = \bigcap_{n=0}^{\infty} Q^{-n}(\gamma^n T).$$

The control points have the following properties:

(a) $T' = T + c(T') - c(T)$, for any tiles T, T' of the same type.
(b) $Q(c(T)) = c(\gamma T)$, for $T \in \mathcal{T}$.

Control points are also fixed for tiles of any tiling $\mathcal{S} \in X_{\mathcal{T}}$: they have the same relative position as in $\mathcal{T}$-tiles.

For $n \ge 1$ let $\mathcal{T}^n := \{Q^n T\colon\ T \in \mathcal{T}\}$. By definition, if $T = (A, i)$, then $Q^n T = (Q^n A, i)$. Thus we consider $Q^n T$ as a tile and $\mathcal{T}^n$ as a tiling. The tiles of $\mathcal{T}^n$ are called *supertiles of level n* and $\mathcal{T}^n$ is called a *supertiling*. Since $\mathcal{T}$ is a fixed point of the substitution ω with expansion Q, we recover $\mathcal{T}$ by subdividing the tiles of $\mathcal{T}^n$ n times. The control points are determined for the tiles of supertilings by $c(QT) = Qc(T)$. For each $T \in \mathcal{T}$ let $T^{(n)}$ be the unique supertile of level n such that $\operatorname{supp}(T) \subset \operatorname{supp}(T^{(n)})$.

Recall that our tile-substitution ω is primitive, that is, for some $k \in \mathbb{N}$, the kth power of the substitution matrix has strictly positive entries. Then we can replace ω by ω^k and assume that the substitution matrix itself is strictly positive (this does not lead to loss of generality since a fixed point of ω is also a fixed point of ω^k). This means that the patch $\omega(T)$ contains tiles of all types for every $T \in \mathcal{T}$. We can then define control points for $\mathcal{T}$-tiles choosing the tile map $\gamma\colon \mathcal{T} \to \mathcal{T}$ so that for any $T \in \mathcal{T}$, the tile γT has the same tile type in $\mathcal{T}$. Then for any $T, S \in \mathcal{T}$,

$$c(\gamma T) - c(\gamma S) \in \Xi(\mathcal{T}).$$

Since $Qc(T) = c(\gamma T)$ for any $T \in \mathcal{T}$,

$$Q(c(T) - c(S)) \in \Xi(\mathcal{T}) \qquad \text{for any} \quad T, S \in \mathcal{T}. \tag{4.1}$$

The next lemma is very close to Theorem 1.5 of [21] and Lemma 6.5 of [23] (however, in [23] FLC was assumed); we provide a direct proof for completeness.

Lemma 4.5 *Let $\mathcal{T}$ be a fixed point of a substitution with expansive map Q and a strictly positive substitution matrix, and suppose that the control points satisfy* (4.1). *Then there exists a finite set U in $\mathbb{R}^d$ for which $QU \subset \Xi(\mathcal{T})$ and $0 \in U$ so that for*

any $T, S \in \mathcal{T}$ *there exist* $N \in \mathbb{N}$ *and* $u(n), w(n) \in U,\ 0 \le n \le N$, *such that*

$$c(T) - c(S) = \sum_{n=0}^{N} Q^n(u(n) + w(n)).$$

Proof Fix any $T, S \in \mathcal{T}$ and consider the sequences of supertiles $T = T^{(0)} \subset T^{(1)} \subset \cdots$ and $S = S^{(0)} \subset S^{(1)} \subset \cdots$ defined above (to be more precise, we should write inclusions for supports). Fix any patch P with the origin in the interior of its support. Then there exists $N \in \mathbb{N}$ such that $T, S \in \omega^N(P)$. Fix such an N. Observe that $T^{(N)} = Q^N T'$ and $S^{(N)} = Q^N S'$ for some $T', S' \in P$. We have

$$\begin{aligned} c(T) - c(S) = &\sum_{n=0}^{N-1} \{c(T^{(n)}) - c(T^{(n+1)})\} + c(T^{(N)}) - c(S^{(N)}) \\ &- \sum_{n=0}^{N-1} \{c(S^{(n)}) - c(S^{(n+1)})\}. \end{aligned}$$

Note that $c(T^{(N)}) - c(S^{(N)}) = Q^N(c(T') - c(S'))$ and

$$\begin{aligned} c(T^{(n)}) - c(T^{(n+1)}) &= Q^n c(T_n'') - Q^n c(\gamma T_n''') \\ &= Q^n(c(T_n'') - c(\gamma T_n''')) \end{aligned}$$

for some $\mathcal{T}$-tiles T_n'', T_n''' such that $T_n'' \in \omega(T_n''')$. Similarly,

$$c(S^{(n)}) - c(S^{(n+1)}) = Q^n(c(S_n'') - c(\gamma S_n'''))$$

for some $\mathcal{T}$-tiles S_n'', S_n''' such that $S_n'' \in \omega(S_n''')$. Thus,

$$c(T) - c(S) = \sum_{n=0}^{N-1} Q^n\{c(T_n'') - c(\gamma T_n''') - (c(S_n'') - c(\gamma S_n'''))\} + Q^N(c(T') - c(S')).$$

Observe that there are finitely many possibilities for $c(T_n'') - c(\gamma T_n''')$, $c(S_n'') - c(\gamma S_n''')$, and $c(T') - c(S')$ (for the first two differences it suffices to consider all the cases for which T_n''' and S_n''' are prototiles, $T_n'' \in \omega(T_n''')$ and $S_n'' \in \omega(S_n''')$). Thus, we obtained the desired representation, in view of (4.1). □

Theorem 4.6 [23, Theorem 4.3] *Let* $\mathcal{T}$ *be a repetitive fixed point of a primitive substitution with expansive map* Q *which has FLC. If* $\alpha \in \mathbb{R}^d$ *is an eigenvalue for* $(X_{\mathcal{T}}, \mathbb{R}^d, \mu)$, *then for any* $x \in \Xi(\mathcal{T})$ *we have* $\|\langle Q^n x, \alpha\rangle\| \xrightarrow{n\to\infty} 0$.

In [23] it was assumed that the expansive map Q is diagonalizable over $\mathbb{C}$, but the proof works in full generality.

Let $\mathcal{M} = \{(c(T) - c(S)) - (c(T') - c(S')):\ T, S, T', S' \in \mathcal{T}\} \subset \mathbb{R}^d$. Combining Lemma 4.5 and Theorem 4.6, we obtain the following corollary.

Corollary 4.7 *Let $\mathcal{T}$ be a fixed point of a substitution with expansive map Q and a strictly positive substitution matrix which has FLC. Suppose that the control points satisfy* (4.1). *Let $\alpha \in \mathbb{R}^d$ be an eigenvalue for $(X_{\mathcal{T}}, \mathbb{R}^d, \mu)$. Then there exists a finite subset W in $\mathbb{R}^d$ independent of the choice of α for which*

$$\|\langle Q^n w, \alpha\rangle\| \xrightarrow{n\to\infty} 0 \qquad \text{for any} \quad w \in W,$$

and for any $y \in \mathcal{M}$, there exist $N \in \mathbb{N}$ and $w(n) \in W$, $0 \le n \le N$, such that

$$y = \sum_{n=0}^{N} Q^n w(n).$$

Proposition 4.8 *Let $\mathcal{T}$ be a fixed point of a substitution with expansive map Q and a strictly positive substitution matrix which has FLC. Suppose that the control points satisfy* (4.1) *and the set of eigenvalues for $(X_{\mathcal{T}}, \mathbb{R}^d, \mu)$ is relatively dense. Then $\{c(T) - c(S)\colon\ T, S \in \mathcal{T}\}$ is uniformly discrete, that is, $\{c(T)\colon\ T \in \mathcal{T}\}$ is a Meyer set.*

Proof Since the set of eigenvalues is relatively dense, there exist eigenvalues $\alpha_1, \ldots, \alpha_d$ for $(X_{\mathcal{T}}, \mathbb{R}^d, \mu)$ such that for any $0 \ne y \in \mathbb{R}^d$,

$$\langle y, \alpha_t\rangle \ne 0 \qquad \text{for some} \quad 1 \le t \le d.$$

We define a norm $|||\cdot|||$ on $\mathbb{R}^d$ in terms of the expansion (3.1):

$$|||y||| = \left|\left|\left| \sum_{i=1}^{r}\sum_{j=1}^{m_i}\sum_{l=1}^{k_{ij}} a_{ijl}(y) e_{ijl} \right|\right|\right| = \sum_{t=1}^{d}\left(\sum_{i=1}^{r}\sum_{j=1}^{m_i}\sum_{l=1}^{k_{ij}} |a_{ijl}(y)\langle e_{ijl}, \alpha_t\rangle| \right).$$

For any $\alpha \in \{\alpha_1, \ldots, \alpha_d\}$ we have

$$\langle y, \alpha\rangle = \sum_{i=1}^{r} T_{y,\alpha,i}, \qquad \text{where} \quad T_{y,\alpha,i} = \sum_{j=1}^{m_i}\sum_{l=1}^{k_{ij}} a_{ijl}(y)\langle e_{ijl}, \alpha\rangle. \tag{4.2}$$

Clearly,

$$|||y||| \ge \sum_{i=1}^{r} |T_{y,\alpha,i}|. \tag{4.3}$$

From Corollary 4.7 we know that any $y \in \mathcal{M}$ can be represented in terms of elements of W so that $y = \sum_{n=0}^{N} Q^n w(n)$ for some positive integer N, where $w(n) \in W$ for any $0 \le n \le N$. Let $w_1, \ldots, w_R$ be all the elements of W. We can rearrange the sum to write

$$y = \sum_{p=1}^{R}\sum_{n\in\mathcal{N}_p} Q^n w_p, \tag{4.4}$$

where $\{\mathcal{N}_1, \ldots, \mathcal{N}_R\}$ is a partition of $\{0, 1, \ldots, N\}$ such that $w(n) = w_p$ if and only if $n \in \mathcal{N}_p$ for $0 \le n \le N$. For any $\alpha \in \{\alpha_1, \ldots, \alpha_d\}$ and $w = \sum_{i=1}^{r} \sum_{j=1}^{m_i} \sum_{l=1}^{k_{ij}} a_{ijl}(w) e_{ijl} \in W$, by Lemma 3.1 we have

$$\left\langle \sum_{j=1}^{m_i} \sum_{l=1}^{k_{ij}} a_{ijl}(w) Q^n e_{ijl}, \alpha \right\rangle = (P_{\alpha,w})_i(n) \lambda_i^n \qquad \text{for} \quad 1 \le i \le r \tag{4.5}$$

and

$$\langle Q^n w, \alpha \rangle = \sum_{i=1}^{r} (P_{\alpha,w})_i(n) \lambda_i^n, \tag{4.6}$$

where $(P_{\alpha,w})_i$ is a polynomial over $\mathbb{C}$ of degree less than or equal to K. Comparing (4.2) and (4.4) and noting that $Q^n e_{ijl}$ is in the subspace of $\mathbb{C}^d$ spanned by $e_{ij1}, \ldots, e_{ijk_{ij}}$, we obtain

$$T_{y,\alpha,i} = \sum_{p=1}^{R} \sum_{n \in \mathcal{N}_p} \left\langle \sum_{j=1}^{m_i} \sum_{l=1}^{k_{ij}} a_{ijl}(w_p) Q^n e_{ijl}, \alpha \right\rangle \qquad \text{for} \quad 1 \le i \le r.$$

From (4.5), we get

$$T_{y,\alpha,i} = \sum_{p=1}^{R} \sum_{n \in \mathcal{N}_p} (P_{\alpha,w_p})_i(n) \lambda_i^n \qquad \text{for} \quad 1 \le i \le r. \tag{4.7}$$

Note that $\|\langle Q^n w, \alpha \rangle\| \xrightarrow{n \to \infty} 0$ by Corollary 4.7, and for any $1 \le i \le r$, λ_i is an algebraic integer by Corollary 4.2, with $|\lambda_i| > 1$ by the expansiveness of Q. Therefore, by Theorem 4.3, for any $w \in W$ and $1 \le i \le r$ we have

$$(P_{\alpha,w})_i(n) = \sum_{k=0}^{K} (c_{\alpha,w,i,k}) n^k, \tag{4.8}$$

where $c_{\alpha,w,i,k} \in \mathbb{Q}(\lambda_i)$, and every conjugate λ of λ_i, with $|\lambda| \ge 1$, occurs in the right-hand side of (4.6), that is, $\lambda = \lambda_j$ for some $j \le r$. Moreover, in this case

$$c_{\alpha,w,j,k} = \tau_{ij}(c_{\alpha,w,i,k}) \qquad \text{for any} \quad 0 \le k \le K,$$

where $\tau_{ij}\colon \mathbb{Q}(\lambda_i) \to \mathbb{Q}(\lambda_j)$ is an isomorphism which is identical on $\mathbb{Q}$ such that $\tau_{ij}(\lambda_i) = \lambda_j$. Since all λ_i are algebraic integers, we have

$$\mathbb{Q}(\lambda_i) = \mathbb{Q}[\lambda_i] = \{a_0 + a_1 \lambda_i + \cdots + a_{s_i - 1} \lambda_i^{s_i - 1} \colon a_n \in \mathbb{Q},\ 0 \le n \le s_i - 1\},$$

where s_i is the degree of the minimal polynomial of λ_i over $\mathbb{Q}$. There are finitely many numbers $c_{\alpha,w,i,k}$, so we can find a positive integer b such that

$$b c_{\alpha,w,i,k} \in \mathbb{Z}[\lambda_i], \qquad \forall \alpha \in \{\alpha_1, \ldots, \alpha_d\}, \quad \forall w \in W, \quad \forall i \le r, \quad \forall k \le K.$$

That is, there exist polynomials $g_{\alpha,w,i,k}(x)$ with integer coefficients such that

$$bc_{\alpha,w,i,k} = g_{\alpha,w,i,k}(\lambda_i) \tag{4.9}$$

and

$$\lambda_i,\ \lambda_j \text{ are conjugates } \Rightarrow\ g_{\alpha,w,i,k}(x) = g_{\alpha,w,j,k}(x).$$

Let

$$C_1 := \max\{|g_{\alpha,w,i,k}(x)|\colon\ |x| \le 1,\ \alpha \in \{\alpha_1, \dots, \alpha_d\},\ w \in W,\ i \le r,\ k \le K\}. \tag{4.10}$$

Note that $C_1 < \infty$.

Now fix $0 \ne y \in \mathcal{M}$ and choose $\alpha \in \{\alpha_1, \dots, \alpha_d\}$ such that $\langle y, \alpha\rangle \ne 0$. Then fix $1 \le i \le r$ such that $T_{y,\alpha,i} \ne 0$, see (4.2). Consider a polynomial $S(x) = S_{y,\alpha,i}(x) \in \mathbb{Z}[x]$ given by

$$S(x) = \sum_{p=1}^{R} \sum_{n \in \mathcal{N}_p} \sum_{k=0}^{K} g_{\alpha,w,i,k}(x) n^k x^n. \tag{4.11}$$

In view of (4.7), (4.8), (4.9), and (4.11),

$$S(\lambda_i) = bT_{y,\alpha,i}. \tag{4.12}$$

Let $\mathcal{H}_i = \{\text{all conjugates } \lambda \text{ of } \lambda_i\colon\ |\lambda| \ge 1\}$ and $\mathcal{G}_i = \{\text{all conjugates } \lambda \text{ of } \lambda_i\}$. By Theorem 4.3(c) we have $\mathcal{H}_i \subset \{\lambda_1, \dots, \lambda_r\}$ and

$$\lambda_j \in \mathcal{H}_i\ \Rightarrow\ S(\lambda_j) = \tau_{ij}(S(\lambda_i)).$$

On the other hand, for any $\lambda \in \mathcal{G}_i \backslash \mathcal{H}_i$,

$$|S(\lambda)| \le C_1 \sum_{k=0}^{K} \sum_{n=0}^{\infty} |n^k \lambda^n|,$$

where C_1 was defined in (4.10). Since $\sum_{n=0}^{\infty} n^k \lambda^n$ converges absolutely for any $|\lambda| < 1$ and $0 \le k \le K$, there exists a constant $C_2 > 0$, independent of y, α, i, such that

$$|S(\lambda)| < C_2 \qquad \text{for any} \quad \lambda \in \mathcal{G}_i \backslash \mathcal{H}_i.$$

Now observe that

$$\Phi := \prod_{\lambda \in \mathcal{G}_i} S(\lambda) \in \mathbb{Z},$$

since S is a polynomial over $\mathbb{Z}$ and the product is symmetric under permutations of the conjugates of λ_i. On the other hand, $\Phi \ne 0$, since $S(\lambda_i) = bT_{y,\alpha,i} \ne 0$ and therefore, $S(\lambda) = \tau(S(\lambda_i)) \ne 0$ where $\tau\colon\ \mathbb{Q}(\lambda_i) \to \mathbb{Q}(\lambda)$ is an isomorphism satisfying $\tau(\lambda_i) = \lambda$, for $\lambda \in \mathcal{G}_i$. Therefore, $|\Phi| \ge 1$, hence

$$\prod_{\lambda \in \mathcal{H}_i} |S(\lambda)| \ge \frac{1}{\prod_{\lambda \in \mathcal{G}_i \backslash \mathcal{H}_i} |S(\lambda)|}. \tag{4.13}$$

Note that

$$\prod_{\lambda\in\mathcal{G}_i\setminus\mathcal{H}_i} |S(\lambda)| \le (C_2)^L \qquad \text{where} \quad L = \#(\mathcal{G}_i\setminus\mathcal{H}_i).$$

Let $H = \#\mathcal{H}_i$. We obtain

$$\left(\sum_{\lambda\in\mathcal{H}_i} |S(\lambda)|\right)^H \ge \prod_{\lambda\in\mathcal{H}_i} |S(\lambda)| \ge (C_2)^{-L},$$

and, in view of (4.3) and (4.12),

$$|||y||| \ge \sum_{\lambda\in\mathcal{H}_i} |T_{y,\alpha,i}| = \frac{1}{b}\sum_{\lambda\in\mathcal{H}_i} |S(\lambda)| \ge \frac{1}{b}(C_2)^{-L/H}. \tag{4.14}$$

Thus, $\{|||y|||\colon\ y\in\mathcal{M}, y\neq 0\}$ has a uniform positive lower bound. Since all norms in $\mathbb{R}^d$ are equivalent, the set $\{c(T)-c(S)\colon\ T,S\in\mathcal{T}\}$ is uniformly discrete in the Euclidean norm. This completes the proof of the proposition. □

Corollary 4.9 *Let $\mathbf{\Lambda}$ be a primitive substitution Delone multiset with expansion Q for which every $\mathbf{\Lambda}$-cluster is legal and $\mathbf{\Lambda}$ has FLC. If the set of eigenvalues for $(X_{\mathbf{\Lambda}},\mathbb{R}^d,\mu)$ is relatively dense, then $\Lambda=\bigcup_{i\le m}\Lambda_i$ is a Meyer set.*

Proof Since $\mathbf{\Lambda}$ is representable by Theorem 2.9, we have that $\mathcal{T} := \mathbf{\Lambda}+\mathcal{A}$ is a repetitive tiling which has FLC and is a fixed point of a primitive substitution ω with expansion Q. Since $(X_{\mathbf{\Lambda}},\mathbb{R}^d,\mu)$ and $(X_{\mathcal{T}},\mathbb{R}^d,\mu)$ are topologically conjugate (see Lemma 3.10 of [17]), the set of eigenvalues for $(X_{\mathcal{T}},\mathbb{R}^d,\mu)$ is relatively dense. The substitution ω is primitive, so we can find $k\in\mathbb{N}$ such that ω^k has a strictly positive substitution matrix. Then we can consider $\mathcal{T}$ as a fixed point of ω^k with expansive map Q^k. We can choose control points for $\mathcal{T}$ to satisfy (4.1), with Q replaced by Q^k. Then Proposition 4.8 applies, and we obtain that $\mathcal{L}-\mathcal{L}$ is uniformly discrete, where $\mathcal{L} := \{c(T)\colon\ T\in\mathbf{\Lambda}+\mathcal{A}\}$.

Then for each $i\le m$, $\Lambda_i\subset a_i+\mathcal{L}$ for some $a_i\in\mathbb{R}^d$ and $\Lambda=\bigcup_{i\le m}\Lambda_i\subset F+\mathcal{L}$ for some finite set F of $\mathbb{R}^d$. So $\Lambda-\Lambda\subset(F-F)+\mathcal{L}-\mathcal{L}$. Since $(F-F)+\mathcal{L}-\mathcal{L}$ is uniformly discrete, Λ is a Meyer set. □

Lemma 4.10 *Let $\mathbf{\Lambda}$ be a Delone multiset in $\mathbb{R}^d$. Suppose that $(X_{\mathbf{\Lambda}},\mathbb{R}^d,\mu)$ has a pure point dynamical spectrum. Then the eigenvalues for the dynamical system $(X_{\mathbf{\Lambda}},\mathbb{R}^d,\mu)$ span $\mathbb{R}^d$.*

Proof Suppose that there is a nonzero $x\in\mathbb{R}^d$ such that $\langle x,\alpha\rangle=0$ for any eigenvalue α for $(X_{\mathbf{\Lambda}},\mathbb{R}^d,\mu)$. We take $x\in\mathbb{R}^d$ with small norm so that $a+x\notin\Lambda$ for all $a\in\Lambda=\bigcup_{i\le m}\Lambda_i$. For an eigenfunction f_α corresponding to the eigenvalue α,

$$f_\alpha(\mathbf{\Lambda}'-x) = e^{2\pi i\langle x,\alpha\rangle} f_\alpha(\mathbf{\Lambda}') = f_\alpha(\mathbf{\Lambda}'), \qquad \text{for } \mu\text{-a.e.} \quad \mathbf{\Lambda}'\in X_{\mathbf{\Lambda}}.$$

For any $f \in L^2(X_{\mathbf{\Lambda}}, \mu)$, $f = \sum_{n=1}^{\infty} f_{\alpha_n}$, where f_{α_n}'s are eigenfunctions. We denote the norm in $L^2(X_{\mathbf{\Lambda}}, \mu)$ by $\|\cdot\|_2$. For any $\epsilon > 0$, there is $N \in \mathbb{N}$ such that

$$\begin{aligned}\|f(\cdot - x) - f\|_2 &\leq \left\| f(\cdot - x) - \sum_{n=1}^{N} f_{\alpha_n}(\cdot - x) \right\|_2 + \left\| \sum_{n=1}^{N} f_{\alpha_n}(\cdot - x) - f \right\|_2 \\ &\leq \left\| f(\cdot - x) - \sum_{n=1}^{N} f_{\alpha_n}(\cdot - x) \right\|_2 + \left\| \sum_{n=1}^{N} f_{\alpha_n} - f \right\|_2 \\ &\leq 2\epsilon.\end{aligned}$$

So $f(\mathbf{\Lambda}' - x) = f(\mathbf{\Lambda}')$ for μ-a.e. $\mathbf{\Lambda}' \in X_{\mathbf{\Lambda}}$. Note that $\mathbf{\Lambda} \neq \mathbf{\Lambda} - x$ by the choice of x. Therefore, we can choose $\epsilon > 0$ such that the ϵ-neighborhood of $\mathbf{\Lambda}$ and its translation by x are disjoint, by the continuity of the action. Consider f to be the characteristic function of the ϵ-neighborhood of $\mathbf{\Lambda}$. We have $f(\mathbf{\Lambda}') = 1$ but $f(\mathbf{\Lambda}' - x) = 0$ for all $\mathbf{\Lambda}'$ in this neighborhood, which is a contradiction. □

Noticing that every integral linear combination of the eigenvalues for $(X_{\mathbf{\Lambda}}, \mathbb{R}^d, \mu)$ is also an eigenvalue for the dynamical system, from Corollary 4.9 and Lemma 4.10 we get the following theorem.

Theorem 4.11 *Let $\mathbf{\Lambda}$ be a primitive substitution Delone multiset with expansion Q for which every $\mathbf{\Lambda}$-cluster is legal and $\mathbf{\Lambda}$ has FLC. Suppose that $(X_{\mathbf{\Lambda}}, \mathbb{R}^d, \mu)$ has a pure point dynamical spectrum. Then $\Lambda = \bigcup_{i \leq m} \Lambda_i$ is a Meyer set.*

Theorem 4.12 [24] *If Λ is a Meyer set and its autocorrelation exists with respect to a van Hove sequence, then the set of Bragg peaks is relatively dense.*

Lemma 4.13 *Let $\mathbf{\Lambda}$ be a Delone multiset for which $\mathbf{\Lambda}$ has FLC and UCF. If the union of the Bragg peaks of the sets Λ_j, $1 \leq j \leq m$, is relatively dense, then the set of eigenvalues for $(X_{\mathbf{\Lambda}}, \mathbb{R}^d, \mu)$ is relatively dense.*

Proof This follows from Lemma 3.4 of [16], which was essentially taken from [4], [8]. We refer to [16] for more details.

It is enough to show that every Bragg peak of any set Λ_j is an eigenvalue for $(X_{\mathbf{\Lambda}}, \mathbb{R}^d, \mu)$. Let $\gamma = \gamma(\delta_{\Lambda_j})$ denote the autocorrelation of δ_{Λ_j} given by (2.7). Let $\omega \in \mathcal{C}_0(\mathbb{R}^d)$, that is, ω is continuous and has compact support. We define

$$f_{j,\omega}(\mathbf{\Lambda}') := (\omega * \delta_{\Lambda_j'})(0) \qquad \text{for} \quad \mathbf{\Lambda}' = (\Lambda_i')_{i \leq m} \in X_{\mathbf{\Lambda}}.$$

Denote by $\gamma_{\omega,\Lambda_j}$ the autocorrelation of $\omega * \delta_{\Lambda_j}$. Then $\gamma_{\omega,\Lambda_j} = (\omega * \widetilde{\omega}) * \gamma$ and, therefore, $\widehat{\gamma_{\omega,\Lambda_j}} = |\widehat{\omega}|^2 \widehat{\gamma}$. By Lemma 3.4 in [16] we note that

$$\sigma_{f_{j,\omega}} = \widehat{\gamma_{\omega,\Lambda_j}},$$

where $\sigma_{f_{j,\omega}}$ is the spectral measure corresponding to $f_{j,\omega}$. (In Lemma 3.4 of [16] we considered the measure $\nu = \sum_{i \leq m} a_i \delta_{\Lambda_i}$; here we take $a_i = \delta_{ij}$.) If α is a Bragg peak

of Λ_j, then $\widehat{\gamma}(\alpha) > 0$. We can certainly find $\omega \in \mathcal{C}_0(\mathbb{R}^d)$ such that $\widehat{\omega}(\alpha) \neq 0$, and then $\sigma_{f_{j,\omega}}(\alpha) > 0$. Thus, the spectral measure corresponding to some L^2 function has a point mass at α, and this implies that α is an eigenvalue for the group of unitary operators (see, e.g., [25]); we conclude that α is an eigenvalue for $(X_{\mathbf{\Lambda}}, \mathbb{R}^d, \mu)$. □

Combining the results above we obtain the following equivalences.

Theorem 4.14 *Let* $\mathbf{\Lambda}$ *be a primitive substitution Delone multiset with expansion* Q *for which every* $\mathbf{\Lambda}$*-cluster is legal and* $\mathbf{\Lambda}$ *has FLC. Then the following are equivalent*:

(i) *The set of Bragg peaks for each* Λ_j *is relatively dense.*
(ii) *The union of Bragg peaks of* Λ_j, $1 \leq j \leq m$, *is relatively dense.*
(iii) *The set of eigenvalues for* $(X_{\mathbf{\Lambda}}, \mathbb{R}^d, \mu)$ *is relatively dense.*
(iv) $\Lambda = \bigcup_{j \leq m} \Lambda_j$ *is a Meyer set.*

Proof (i) ⇒ (ii) is trivial; (ii) ⇒ (iii) is Lemma 4.13, (iii) ⇒ (iv) is Corollary 4.9. Finally, (iv) ⇒ (i) follows by Strungaru's Theorem 4.12. Note that each Λ_j is a Meyer set, since $\Lambda_j - \Lambda_j$ is uniformly discrete and Λ_j is a Delone set. We apply Theorem 4.12 to each Λ_j. (It is known that a primitive substitution Delone multiset for which every $\mathbf{\Lambda}$-cluster is legal has UCF, see, e.g., [17], hence for every Λ_j there exists unique autocorrelation.) □

This theorem readily shows Theorem 1.1 and Corollary 1.2 in the Introduction.

Acknowledgement We are grateful to the referees for many helpful comments.

References

1. Baake, M., Lenz, D.: Dynamical systems on translation bounded measures: pure point dynamical and diffraction spectra. Ergod. Theory Dyn. Syst. **24**(6), 1867–1893 (2004)
2. Baake, M., Lenz, D., Moody, R.V.: Characterization of model sets by dynamical systems. arXiv:math.DS/0511648 (2005)
3. Baake, M., Moody, R.V.: Self-similar measures for quasi-crystals. In: Baake, M., Moody, R.V. (eds.) Directions in Mathematical Quasicrystals. CRM Monograph Series, vol. 13, pp. 1–42. Am. Math. Soc., Providence (2000)
4. Dworkin, S.: Spectral theory and X-ray diffraction. J. Math. Phys. **34**, 2965–2967 (1993)
5. Garsia, A.: Arithmetic properties of Bernoulli convolutions. Trans. Am. Math. Soc. **102**, 409–432 (1962)
6. Gouéré, J.-B.: Diffraction et mesure de Palm des processus ponctuels (Diffraction and Palm measure of point processes). C.R. Math. Acad. Sci. Paris **336**(1), 57–62 (2003)
7. Hof, A.: On diffraction by aperiodic structures. Commun. Math. Phys. **169**(1), 25–43 (1995)
8. Hof, A.: Diffraction by aperiodic structures. In: Moody, R.V. (ed.) The Mathematics of Long-Range Aperiodic Order, Waterloo, 1995. NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci., vol. 489, pp. 239–268. Kluwer, Dordrecht (1997)
9. Horn, R.A., Johnson, C.R.: Topics in Matrix Analysis. Cambridge University Press, Cambridge (1991)
10. Kenyon, R.: Self-similar tilings. Ph.D. Thesis, Princeton University, Princeton (1990)
11. Környei, I.: On a theorem of Pisot. Publ. Math. (Debr.) **34**(3–4), 169–179 (1987)
12. Lagarias, J.C.: Geometric models for quasicrystals, I. Delone sets of finite type. Discrete Comput. Geom. **21**(2), 161–191 (1999)

13. Lagarias, J.C.: Mathematical quasicrystals and the problem of diffraction. In: Baake, M., Moody, R.V. (eds.) Directions in Mathematical Quasicrystals. CRM Monograph Series, vol. 13, pp. 61–93. Am. Math. Soc., Providence (2000)
14. Lagarias, J.C., Wang, Y.: Substitution Delone sets. Discrete Comput. Geom. **29**, 175–209 (2003)
15. Lee, J.-Y.: Substitution Delone sets with pure point spectrum are model sets. Preprint (2005)
16. Lee, J.-Y., Moody, R.V., Solomyak, B.: Pure point dynamical and diffraction spectra. Ann. Henri Poincaré **3**, 1003–1018 (2002)
17. Lee, J.-Y., Moody, R.V., Solomyak, B.: Consequences of pure point diffraction spectra for multiset substitution systems. Discrete Comput. Geom. **29**, 525–560 (2003)
18. Mauduit, C.: Caractérisation des ensembles normaux substitutifs. Invent. Math. **95**(1), 133–147 (1989)
19. Meyer, Y.: Algebraic Numbers and Harmonic Analysis. North-Holland Math. Library, vol. 2. North-Holland, Amsterdam (1972)
20. Moody, R.V.: Meyer sets and their duals. In: Moody, R.V. (ed.) The Mathematics of Long-Range Aperiodic Order, Waterloo, 1995. NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci., vol. 489, pp. 403–441. Kluwer, Dordrecht (1997)
21. Praggastis, B.: Numeration systems and Markov partitions from self-similar tilings. Trans. Amer. Math. Soc. **351**(8), 3315–3349 (1999)
22. Robinson, E.A. Jr.: Symbolic dynamics and tilings of $\mathbb{R}^d$. In: Symbolic Dynamics and Its Applications. Proc. Sympos. Appl. Math., vol. 60, pp. 81–119. Am. Math. Soc., Providence (2004)
23. Solomyak, B.: Dynamics of self-similar tilings. Ergod. Theory Dyn. Syst. **17**, 695–738 (1997). Corrections to "Dynamics of self-similar tilings", Ibid. **19**, 1685 (1999)
24. Strungaru, N.: Almost periodic measures and long-range order in Meyer sets. Discrete Comput. Geom. **33**(3), 483–505 (2005)
25. Weidmann, J.: Linear Operators in Hilbert Space. Graduate Texts in Mathematics. Springer, New York (1980)

Metric Combinatorics of Convex Polyhedra: Cut Loci and Nonoverlapping Unfoldings

Ezra Miller · Igor Pak

Abstract Let S be the boundary of a convex polytope of dimension $d+1$, or more generally let S be a *convex polyhedral pseudomanifold*. We prove that S has a polyhedral nonoverlapping unfolding into $\mathbb{R}^d$, so the metric space S is obtained from a closed (usually nonconvex) polyhedral ball in $\mathbb{R}^d$ by identifying pairs of boundary faces isometrically. Our existence proof exploits geodesic flow away from a source point $v \in S$, which is the exponential map to S from the tangent space at v. We characterize the *cut locus* (the closure of the set of points in S with more than one shortest path to v) as a polyhedral complex in terms of Voronoi diagrams on facets. Analyzing infinitesimal expansion of the wavefront consisting of points at constant distance from v on S produces an algorithmic method for constructing Voronoi diagrams in each facet, and hence the unfolding of S. The algorithm, for which we provide pseudocode, solves the discrete geodesic problem. Its main construction generalizes the source unfolding for boundaries of three-polytopes into $\mathbb{R}^2$. We present conjectures concerning the number of shortest paths on the boundaries of convex polyhedra, and concerning continuous unfolding of convex polyhedra. We also comment on the intrinsic nonpolynomial complexity of nonconvex manifolds.

The first author was partially supported by the National Science Foundation, and he acknowledges the Mathematisches Forschungsinstitut Oberwolfach for a stimulating research environment during the week-long program on Topological and Geometric Combinatorics (April, 2003). Most of this work was completed while the first author was at the Massachusetts Institute of Technology (Cambridge, MA) and the Mathematical Sciences Research Institute (Berkeley, CA). The second author was partially supported by the National Security Agency and the National Science Foundation. He thanks the organizers of the "Second Geometry Meeting dedicated to A.D. Aleksandrov" held at the Euler International Mathematical Institute in St. Petersburg (June, 2002), where these results were originally presented.

E. Miller (✉)
School of Mathematics, University of Minnesota, Minneapolis, MN 55455, USA
e-mail: ezra@math.umn.edu

I. Pak
Department of Mathematics, MIT, Cambridge, MA 02939, USA
e-mail: pak@math.mit.edu

Introduction

The past several decades have seen intense development in the combinatorics and geometry of convex polytopes [39]. Besides their intrinsic interest, the advances have been driven by applications to areas ranging as widely as combinatorial optimization, commutative algebra, symplectic geometry, theoretical physics, representation theory, statistics, and enumerative combinatorics. As a result, there is currently available a wealth of insight into (for example) algebraic invariants of the face posets of polytopes; arithmetic information connected to sets of lattice points inside polytopes; and geometric constructions associated with linear functionals, such as Morse-like decompositions and methods for locating extrema.

On the topological side, there are metric theories for polyhedral spaces, primarily motivated by differential geometry. In addition, there is a vast literature on general convexity. Nonetheless, there seems to be lacking a study of the interaction between the combinatorics of the boundaries of convex polytopes and their metric geometry in arbitrary dimension. This remains the case despite relations to a number of classical algorithmic problems in discrete and computational geometry.

The realization here is that convexity and polyhedrality together impose rich combinatorial structures on the collection of shortest paths in a metrized sphere. We initiate a systematic investigation of this *metric combinatorics* of convex polyhedra by proving the existence of polyhedral nonoverlapping unfoldings and analyzing the structure of the cut locus. The algorithmic aspect, which we include together with its complexity analysis, was for us a motivating feature of these results. That being said, we also show that our general methods are robust enough so that—with a few minor modifications—they extend to the abstract spaces we call 'convex polyhedral pseudomanifolds', whose sectional curvatures along low-dimensional faces are all positive. To conclude, we propose some directions for future research, including a series of precise conjectures on the number of combinatorial types of shortest paths, and on the geometry of unfolding boundaries of polyhedra.

Overview

Broadly speaking, the metric geometry of boundaries of three-dimensional polytopes is quite well understood, due in large part to the work of Aleksandrov [3, 4] and his school. For higher dimensions, however, less theory appears in the literature, partly because Aleksandrov's strongest methods do not extend to higher dimension. Although there do exist general frameworks for dealing with metric geometry in spaces general enough to include boundaries of convex polyhedra, such as [10], the special nature of polyhedral spaces usually plays no role.

The existing theory that does appear for polyhedral spaces is motivated from the perspective of Riemannian geometry, via metric geometry on simplicial complexes, and seems mainly due to Stone; see [34], for example. In contrast, our original motivation comes from two classical problems in discrete and computational geometry: the "discrete geodesic problem" [24] of finding shortest paths between points on polyhedral surfaces, and the problem of constructing nonoverlapping unfoldings of convex polytopes [28]. Both problems are well understood for the two-dimensional

boundaries of 3-polytopes, but have not been attempted in higher dimensions. We resolve them here in arbitrary dimension by a unified construction generalizing the "source unfolding" of three-dimensional convex polyhedra [33, 37].

Previous methods for source unfoldings have been specific to low dimension, relying for example on the fact that arcs of circles in the plane intersect polygons in finite sets of points. We instead use techniques based on differential geometry to obtain general results concerning cut loci on boundaries of polytopes in arbitrary dimension, namely Theorem 2.9 and Corollary 2.11, thereby producing polyhedral foldouts in Theorem 3.5. In more precise terms, our two main goals in this paper are to:

1. describe how the set of points on the boundary S of a convex polyhedron at given radius from a fixed *source point* changes as the radius increases continuously;
2. use this description of "wavefront expansion" to construct a polyhedral nonoverlapping unfolding of the d-dimensional polyhedral complex S into $\mathbb{R}^d$.

By "describe" and "construct" we mean to achieve these goals not just abstractly and combinatorially, but effectively, in a manner amenable to algorithmic computation. References such as [1, 5, 13, 20, 26, 27, 32], and [33], which have their roots and applications in computational geometry, carry this out in the $d = 2$ case of boundaries of 3-polytopes (and for the first goal, on any polyhedral surface of dimension $d = 2$). Here, in arbitrary dimension d, our Theorem 5.2 says precisely how past wavefront evolution determines the location in time and space of its next qualitative change. The combinatorial nature of Theorem 5.2 leads immediately to Algorithm 6.1 for effectively unfolding boundaries of polyhedra.

The results and proofs in Sections 1–6 for boundaries of convex polyhedra almost all hold verbatim in the more abstract setting of what we call d-dimensional *convex polyhedral pseudomanifolds*. The study of such spaces is suggested both by Stone's point of view in [34] and by the more general methods in [10]. Our Corollary 7.12 says that all convex polyhedral pseudomanifolds can be represented as quotients of Euclidean (usually nonconvex) polyhedral balls by identifying pairs of boundary components isometrically. The reader interested solely in this level of generality is urged to begin with Section 7, which gives a guide to Sections 1–6 from that perspective, and provides the slight requisite modifications where necessary. Hence the reader can avoid checking the proofs in the earlier sections twice.

The results in Section 7 on convex polyhedral pseudomanifolds are in many senses sharp, in that considering more general spaces would falsify certain conclusions. We substantiate this claim in Section 8, where we also discuss extensions of our methods that are nonetheless possible. For example, we present an algorithm to construct geodesic Voronoi diagrams on boundaries of convex polyhedra in Section 8.9.

The methods of this paper suggest a number of fundamental open questions about the metric combinatorics of convex polyhedra in arbitrary dimension, and we present these in Section 9. Most of them concern the notion of *vistal tree* in Definition 9.1, which encodes all of the combinatorial types of shortest paths (or equivalently, all bifurcations of the wavefront) emanating from a source point. The first two questions, Conjectures 9.2 and 9.4, concern the complexity of our unfolding algorithm and the behavior of geodesics in boundaries of polyhedra. Along these lines, we remark also on the complexity of nonconvex polyhedral manifolds, in Proposition 9.8. Our third question is about the canonical subdivision of the boundary of any convex polyhedron

determined by the sets of source points having isomorphic vistal trees (Definition 9.5 and Conjecture 9.6); it asks whether this *vistal subdivision* is polyhedral, and how many faces it has. Our final question asks how to realize unfoldings of polyhedral boundaries by embedded homotopies (Conjecture 9.12).

As a guide for the reader navigating this paper, the list of sections is as follows:

0 Methods

This section contains an extended overview of the paper, including background and somewhat informal descriptions of the geometric concepts involved.

Unfolding Polyhedra

While unfolding convex polytopes is easy [3], constructing a *nonoverlapping* unfolding is in fact a difficult task with a long history going back to Dürer in 1528 [31]. When cuts are restricted to ridges (faces of dimension $d-1$ in a polyhedron of dimension $d+1$), the existence of such unfoldings is open even for polytopes in $\mathbb{R}^3$ [28, 31]. It is known that nonconvex polyhedral surfaces need not admit such nonoverlapping unfoldings [7, 36].

In this paper we consider unfoldings of a more general nature: cuts are allowed to slice the interiors of facets. Nonoverlapping such unfoldings are known, but only for three-dimensional polytopes [1, 5, 13, 33]. In fact, two different (although strongly related) unfoldings appear in these and other references in the literature: the *Aleksandrov unfolding* (also known as the *star unfolding*) [3, 5], and the *source unfolding* [33, 37]. Unfortunately, the construction of Aleksandrov unfoldings fails in principle in higher dimension (Section 8.4). As we mentioned earlier, we generalize the source unfolding construction to prove that the boundary S of any convex polyhedron of dimension $d+1$, and more abstractly any convex polyhedral pseudomanifold S, has a nonoverlapping polyhedral unfolding $\overline{U}$ in $\mathbb{R}^d$. The second of the two foldouts of the cube in Fig. 1 is a $d=2$ example of a source unfolding. For clarity, we present the discussion below in the context of boundaries of polyhedra.

Cut Loci

The idea of the source unfolding in arbitrary dimension d is unchanged from the case $d=2$ of convex polyhedral surfaces. Pick a *source point* v interior to some facet

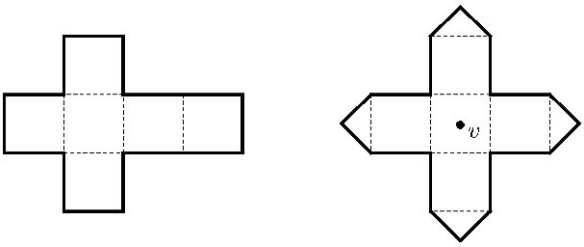


Fig. 1 An edge-unfolding and a source unfolding of a cube into $\mathbb{R}^2$.

(d-dimensional face) of S, so the tangent space T_v is well-defined. Then, treating S like a Riemannian manifold, define the *exponential map* from T_v to S by flowing along geodesics emanating from v. Our main unfolding result, Theorem 3.5, says that exponentiation takes a certain open polyhedral ball $U_v \subset T_v$ isometrically to a dense open subset of S consisting of points possessing a unique shortest path (length-minimizing geodesic) to v. The image of the closure $\overline{U_v}$ of the open ball U_v is all of S. The boundary $\overline{U_v} \backslash U_v$ maps onto the *cut locus* $\overline{K_v}$, which by definition is the closure of the set of points in S with more than one shortest path to the source point v. These properties characterize U_v.

In Riemannian geometry, when the manifold and the metric are both smooth, describing the cut locus for a source point is already an important and interesting problem (see [21] for an excellent introduction and numerous references), although of course the exponential map can only be an isometry, even locally, if the metric is flat. Extending the notion of cut locus from Riemannian geometry to the polyhedral context is just as easy as extending it to arbitrary metric spaces. However, showing that the open ball U_v is a polyhedral foldout requires strong conditions on the complement of the cut locus, such as metric flatness and polyhedrality. We prove these results in Sections 1 and 2 using methods based on the foundations of polyhedral geometry, and on Voronoi diagrams, culminating in Theorem 2.9 and Corollary 2.11. These conclusions depend crucially on convexity and do not hold in the nonconvex case.

Geometry of Wavefront Expansion

Our existence proof for polyhedral nonoverlapping source foldouts, even given their Voronoi characterization in Theorem 2.9, does not by itself provide a satisfactory combinatorial picture of the dynamics of wavefront expansion on polyhedra. For this, we must gain control over how the exponential map behaves as it interacts with *warped* points in S, namely those of nontrivial curvature,[1] or equivalently points on faces of dimension $d-2$ or less.

Imagine the picture kinetically: the source point v emits a signal, whose wavefront proceeds as a $(d-1)$-sphere of increasing radius—at least until the sphere

[1] In differential geometry, when polyhedra are expressed as limits of (sequences of) smooth Riemannian manifolds, all of the curvature is forced into decreasing neighborhoods of the $(d-2)$-skeleton. Consequently, the curvature actually tends to infinity near faces of dimension $d-2$ or less, even though in a polyhedral sense the curvature is finite.

hits the boundary of the facet containing v. At that stage, the wavefront folds over a *ridge*, or face of dimension $d-1$. Metrically, nothing has happened: points interior to ridges look to the wavefront just as flat as points interior to facets. However, later, as the wavefront encounters faces of lower dimension, it is forced to bifurcate around warped points and interfere with itself, as signals emitted originally in different directions from v curl around the nontrivial curvature and converge toward the cut locus.

The question becomes: What discrete structure governs evolution of the wavefront on polyhedra? The most obvious first step is to define a finite collection of "events," representing the points in time and space where the wavefront changes in some nontrivial way. If this is done properly, then it remains only to order the events according to the times at which they occur. However, in reality, the definition of an event is rather simple, while the geometry dictating time order of events is more complex.

Starting from scratch, one might be tempted (and we were) to mark an event every time the wavefront encounters a new warped face. Indeed, this works in dimension $d=2$ [26]: since the wavefront is a curve, its intersection with the set of edges is a finite set, and it is easy to detect when one of these points hits a vertex of S before another. However, because the geometry is substantially more complicated in higher dimensions, in the end we found it more natural to say an *event* has occurred every time the wavefront encounters a new facet through the relative interior of a *ridge* (see Definitions 2.3 and 4.14). This may seem counterintuitive, since the wavefront only interacts with and curls around faces of smaller dimension. However, wavefront collisions with warped points lead to intersections with ridge interiors infinitesimally afterward. In other words, the closest point (*event point*) on a facet to the source point v need not lie interior to a ridge, but can just as easily be warped.

Again think kinetically: once the wavefront has hit a new face (of small dimension, say), it begins to creep up each of the ridges containing that face. Although in a macroscopic sense the wavefront hits all of these ridges simultaneously, it creeps up their interiors at varying rates. Therefore the wavefront hits some of these ridges before others in an infinitesimal sense. The moral is that if one wants to detect curling of the wavefront around warped faces, it is simpler to detect the wake of this interaction infinitesimally on the interiors of neighboring ridges. Sufficiently refined tangent data along ridges then discretizes the finite set of events, thereby producing the desired "metric combinatorics" of wavefront expansion.

Source Poset

Making the above moral precise occupies Section 4. To single out a ridge whose interior is engulfed by the wavefront at a maximal rate (thereby making it *closer* to the source point) essentially is to find a ridge whose angle with the corresponding signal ray emitted from the source is minimal. When $d=2$, this means that we do not simply observe two signals hitting vertices simultaneously, but we notice also the angles at which they hit the edges containing those vertices. The edge forming the smallest angle with its signal ray is the earlier event, infinitesimally beating out other potential events. (That each angle must be measured inside some ambient facet is just one of the subtleties that we gloss over for now.)

To distinguish events in time macroscopically, only radii (distances from the source) are required. When $d = 2$, as we have just seen, a first derivative is enough to distinguish events infinitesimally. Generally, in dimension $d \geq 2$, one needs derivatives of order less than d or, more precisely, a directional derivative successively along each of $d - 1$ orthogonal directions inside a ridge. In Section 4 these derivatives are encoded not in single angles, but in *angle sequences* (Definition 4.2), which provide quantitative information about the goniometry of intersections between signal rays and the faces of varying dimension they encounter. More qualitative—and much more refined—data is carried by *minimal jet frames* (Definition 4.1), which record not just the sizes of the angles, but their directions as well.

The totality of the (finite amount of) radius and angle sequence data induces a partial order on events. The resulting *source poset* (Definition 4.14), which owes its existence to the finiteness result in Theorem 4.11, describes precisely which events occur before others—both macroscopically and infinitesimally. Since wavefront bifurcation is a local phenomenon at an event point, incomparable events can occur simultaneously, or can be viewed as occurring in any desired order. Thus as time progresses, wavefront expansion builds the source poset by adding one event at a time.

The Algorithm

It is one thing to order the set of events after having been given all of them, but it is quite another to predict the "next" event having been given only past events. That the appropriate event to add can be detected locally, and *without knowing future events*, is the content of Theorem 5.2. Its importance is augmented by it being the essential tool in making our algorithm for constructing the source poset, and hence also the source unfolding (Algorithm 6.1) Surprisingly, our geometric analysis of infinitesimal wavefront expansion in Sections 4 and 5 allows us to remove all calculus from Theorem 5.2 and hence Algorithm 6.1: detecting the next event requires only standard tools from linear algebra.

As we mentioned earlier, our original motivation for this paper was its algorithmic applications. Using the theoretical definitions and results in earlier sections, we present pseudocode for our procedure constructing source unfoldings in Algorithm 6.1. That our algorithm provides an *efficient* method to compute source unfoldings is formalized in Theorem 6.5.

There are several arguments in favor of presenting pseudocode. First, it underscores the explicit effective nature of our combinatorial description of the source poset in Theorem 5.2. Second, it emphasizes the simplicity of the algorithm that results from the apparently complicated analysis in Sections 1–5; in particular, the reader interested only in the computational aspects of this paper can start with Section 6 and proceed backwards to read only those earlier parts of the paper addressed in the algorithm. Finally, the pseudocode makes Algorithm 6.1 amenable to actual implementation, which would be of interest but lies outside the scope of this work.[2]

[2]We refer the reader to a recent efficient implementation [35] of the classical $d = 2$ algorithm in [26].

A Note on the Exposition

Proofs of statements that may seem obvious based on intuition drawn from polyhedral surfaces, or even solids of dimension 3, demand surprising precision in the general case. Occasionally, the required adjustments in definitions and lemmas, and even in statements of theorems, were borne out only after considering configurations in dimension 5 or more. The definition of source image is an example, about which we remark in Section 8, in the course of analyzing where various hypotheses (convexity, pseudomanifold, and so on) become essential. Fortunately, once the appropriate notions have been properly identified, the definitions become transparent, and the proofs remain intuitive in low dimension.

1 Geodesics in Polyhedral Boundaries

In this paper a *convex polyhedron* F of dimension d is a finite intersection of closed half-spaces in some Euclidean space $\mathbb{R}^d$, such that F that does not lie in a proper affine subspace of $\mathbb{R}^d$. The polyhedron F need not be bounded, and comes with an induced Euclidean metric. Gluing a finite collection of convex polyhedra by given isometries on pairs of codimension 1 faces yields a (*finite*) *polyhedral cell complex* S. More precisely, S is a regular cell complex endowed with a metric that is piecewise Euclidean, in which every face (closed cell) is isometric to a convex polyhedron.

The case of primary interest is when the polyhedral cell complex S equals the boundary ∂P of a convex polyhedron P of dimension $d+1$ in $\mathbb{R}^{d+1}$.

Convention 1.1 We assume that $S = \partial P$ is a polyhedral boundary in all theorems, proofs, and algorithms from here through Section 6.

We do not require P to be bounded, though the reader interested in polytopes will lose very little of the flavor by restricting to that case. Moreover, with the exception of Lemma 1.3, Proposition 2.10, Corollary 2.11, and Theorem 3.5, the statements of all results from here through Section 6 are worded to hold verbatim for the more abstract class of *convex polyhedral pseudomanifolds*, as we shall see in Section 7.

Denote by μ the metric on S, so $\mu(a, b)$ denotes the distance between points $a, b \in S$. A path $\gamma \subset S$ with endpoints a and b is a *shortest path* if its length equals $\mu(a, b)$. Since we assume S has finitely many *facets* (maximal faces), such length-minimizing paths exist, and are piecewise linear. A path $\eta \subset S$ is a *geodesic* if η is locally a shortest path; i.e., for every $z \in \eta$ that is not an endpoint of η, there exist points $a, b \in \eta \backslash \{z\}$ such that $z \in \gamma \subset \eta$ for some shortest path γ connecting a to b.

Henceforth, as S has dimension d, a face of dimension $d-1$ will be called a *ridge*. For convenience, we say that a point x is *warped* if x lies in the union S_{d-2} of all faces in S of dimension at most $d-2$, and call x *flat* otherwise. Every flat point has a neighborhood isometric to an open subset of $\mathbb{R}^d$.

Proposition 1.2 *If γ is a shortest path in S between its endpoints, then γ has no warped points in its relative interior.*

Proof For any point w lying in the relative interior of γ, the intersection of γ with some neighborhood of w consists of two line segments η and η' that are each straight with one endpoint at w, when viewed as paths in $\mathbb{R}^{d+1}$. This is a consequence of local length-minimization and the fact that each facet of P is isometric to a polytope in $\mathbb{R}^d$. Moreover, if w happens to lie on a ridge while η intersects the relative interior of some facet containing w, then local length-minimization implies that η' is not contained in the facet containing η. Lemma 1.3 shows that w does not lie in S_{d-2}, so the point w is not warped. □

Lemma 1.3 *Let $\eta, \eta' \subset S$ be two paths that* (i) *are straight in $\mathbb{R}^{d+1}$*, (ii) *share a common warped endpoint $w \in S_{d-2}$, and* (iii) *do not both lie in a single facet. There exists a neighborhood $\mathcal{O}$ of w in S such that for every $a \in \eta \cap \mathcal{O}$ and $b \in \eta' \cap \mathcal{O}$, the path η_{ab} from a to w to b along η and η' is not a shortest path in S between a and b.*

Proof Translate P so that w equals the origin $\mathbf{0} \in \mathbb{R}^{d+1}$, and let Q be the unique minimal face of P that contains w. Since η and η' do not lie in a single facet, the 2-plane E spanned by η and η' meets Q at exactly one point, namely $\mathbf{0}$. Since $\dim(Q) \leq d-2$, the span of Q and E has dimension at most d. Choose a line L whose direction is linearly independent from the span of Q and E. Then the 3-plane $H = L + E$ intersects Q only at $\mathbf{0}$. Replacing P by $P \cap H$, we can assume that $\dim(P) = 3$, so that $d = 2$; note that $\mathbf{0}$ is a vertex of $H \cap P$ by construction.

Although the case $d = 2$ was proved in Theorem 4.3.5 of [3] (see also Lemma 4.1 of [33]), we provide a simple argument here, for completeness. Let $\mathcal{O} \subset S$ be the neighborhood of w consisting of all points at some fixed small distance from the vertex w. Then $\mathcal{O}$ can be laid flat on the plane $\mathbb{R}^2$ by slicing along η. One of the two points in this unfolding that glue to $a \in \mathcal{O}$ connects by a straight segment in the unfolding to the unique point corresponding to b. This straight segment shortcuts η_{ab} after gluing back to S. □

An illustration of Lemma 1.3 and its proof is given in Fig. 2.

Corollary 1.4 *Let η be a bounded geodesic in S starting at a point z not on any ridge. Then η intersects each ridge in a discrete set, so η traverses* (*in order*) *the interiors of a well-defined sequence $\mathcal{L}_\eta$ of facets* (*the* facet sequence *of* η).

Proof Since η is locally length minimizing, Proposition 1.2 implies that every intersection of η with a ridge takes place at a flat point. Such points have neighborhoods isometric to open subsets of $\mathbb{R}^d$, and these intersect η in paths isometric to straight segments. It follows that η intersects every ridge transversely. □

For each facet F of $S = \partial P$, let T_F be the affine span of F in $\mathbb{R}^{d+1}$.

Definition 1.5 Suppose two facets F and F' share a ridge $R = F \cap F'$. The *folding map* $\Phi_{F,F'}\colon T_F \to T_{F'}$ is the isometry that identifies the copy of R in T_F with the one in $T_{F'}$ in such a way that the image of F does not intersect the interior of F'.

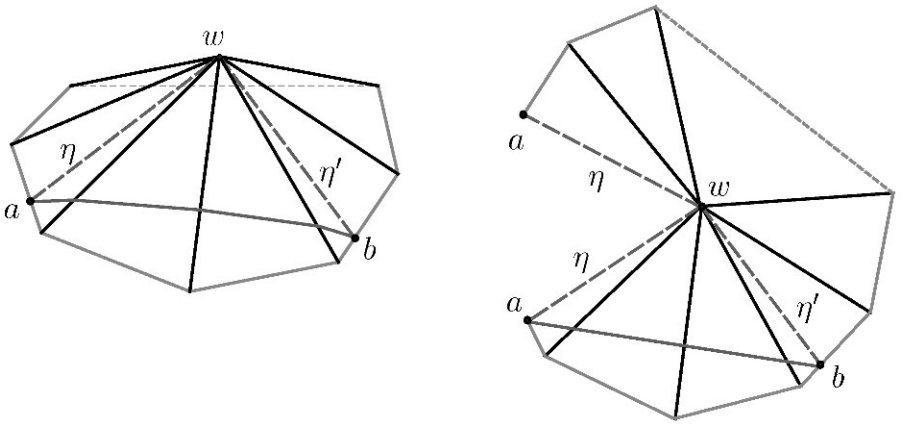


Fig. 2 Neighborhood of a vertex and its foldout after slicing along the segment η. The points a and b are connected by a shortest path.

In other words, the folding map $\Phi_{F,F'}$ is the rotation of T_F with $(d-1)$-dimensional axis $R = F \cap F'$ so that F becomes coplanar with F' and lies on the other side of R from F'. It can be convenient to view $\Phi_{F,F'}$ as rotating all of $\mathbb{R}^{d+1}$ instead of only rotating T_F onto $T_{F'}$. Informally, we say $\Phi_{F,F'}$ *folds* T_F *along* R to lie in the same affine hyperplane as F'.

Definition 1.6 Given an ordered list $\mathcal{L} = (F_1, F_2, \ldots, F_\ell)$ of facets such that F_i shares a (unique) ridge with F_{i+1} whenever $1 \le i < \ell$, we write

$$\Phi_{\mathcal{L}}^{-1} = \Phi_{F_1,F_2}^{-1} \circ \Phi_{F_2,F_3}^{-1} \circ \cdots \circ \Phi_{F_{\ell-1},F_\ell}^{-1}$$

for the *unfolding of* T_{F_ℓ} *onto* T_{F_1}, noting that indeed $\Phi_{\mathcal{L}}^{-1}(T_{F_\ell}) = T_{F_1}$. Setting $\mathcal{L}_i = (F_1, \ldots, F_i)$, the *sequential unfolding* of a subset $\Gamma \subseteq F_1 \cup \cdots \cup F_\ell$ *along* $\mathcal{L}$ is the set

$$(\Gamma \cap F_1) \cup \Phi_{\mathcal{L}_2}^{-1}(\Gamma \cap F_2) \cup \cdots \cup \Phi_{\mathcal{L}_\ell}^{-1}(\Gamma \cap F_\ell) \subset T_{F_1}.$$

By Corollary 1.4, we can sequentially unfold any geodesic. Next, we use this unfolding to show uniqueness of shortest paths traversing given facet sequences.

Lemma 1.7 *Let v and w be flat points in S. Given a sequence $\mathcal{L}$ of facets, there can be at most one shortest path γ connecting v to w such that γ traverses $\mathcal{L}_\gamma = \mathcal{L}$.*

Proof Let γ be a shortest path from v to w traversing $\mathcal{L}$. Inside the union of facets appearing in $\mathcal{L}$, the relative interior of γ has a neighborhood isometric to an open subset of $\mathbb{R}^d$ by Proposition 1.2 and the fact that the set of warped points is closed. Sequential unfolding of γ into T_F for the first facet F in $\mathcal{L}$ thus yields a straight segment in T_F. This identifies γ uniquely as the path in S whose sequential folding along $\mathcal{L}$ is the straight segment in T_F connecting v to $\Phi_{\mathcal{L}}^{-1}(w) \in T_F$. □

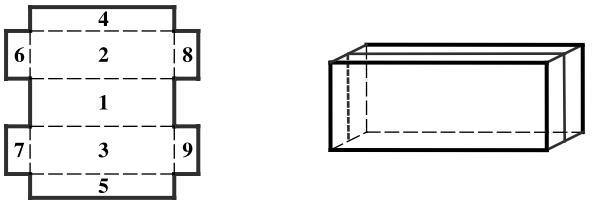


Fig. 3 An unfolding of a $1 \times 1 \times 3$ box.

In the proof of Lemma 1.7, we do not claim that the union of facets in the list $\mathcal{L}$ unfolds sequentially without overlapping, even though some shortest path γ traverses $\mathcal{L}$. However, some neighborhood of γ in this union of facets unfolds without overlapping.

Example 1.8 Consider the unfolding of a $1 \times 1 \times 3$ rectangular box as in Fig. 3. Denote by $F_{\text{bot}}, F_{\text{top}}, F_{\text{front}}, F_{\text{back}}, F_{\text{left}}, F_{\text{right}}$ the bottom, top, front, back, left, and right facets, respectively. Denote by $\mathcal{L}_i$ the list of facets along which the points in the region marked by i have been sequentially unfolded to create the foldout $U \subset T_{F_{\text{bot}}}$ in Fig. 3. Then:

$$\mathcal{L}_1 = (F_{\text{bot}}), \quad \mathcal{L}_2 = (F_{\text{bot}}, F_{\text{back}}), \quad \mathcal{L}_3 = (F_{\text{bot}}, F_{\text{front}}),$$
$$\mathcal{L}_4 = (F_{\text{bot}}, F_{\text{back}}, F_{\text{top}}), \quad \mathcal{L}_5 = (F_{\text{bot}}, F_{\text{front}}, F_{\text{top}}), \quad \mathcal{L}_6 = (F_{\text{bot}}, F_{\text{back}}, F_{\text{left}}),$$
$$\mathcal{L}_7 = (F_{\text{bot}}, F_{\text{front}}, F_{\text{left}}), \quad \mathcal{L}_8 = (F_{\text{bot}}, F_{\text{back}}, F_{\text{right}}), \quad \mathcal{L}_9 = (F_{\text{bot}}, F_{\text{front}}, F_{\text{right}}).$$

2 Cut Loci

Most of this paper concerns the set of shortest paths with one endpoint fixed.

Definition 2.1 Fix a *source point* $v \in S$ lying interior to some facet. A point $x \in S$ is a *cut point*[3] if x has more than one shortest path to v. Denote the set of cut points by K_v, and call its closure the *cut locus* $\overline{K}_v \subset S$.

Here is a consequence of Proposition 1.2.

Corollary 2.2 *No shortest path in S to the source point v has a cut point in its relative interior.*

Proof Suppose c is a cut point in the relative interior of a shortest path from v to w. Replacing the path from v to c with another shortest path from v to c yields a new shortest path from v to w. These two paths to w meet at the flat point $c \in S$ by

[3]Our usage of the term "cut locus" is standard in differential geometry, just as our usage of "ridge" is standard in polyhedral geometry. However, these usages do not agree with terminology in computer science, such as in [33] and [5]: their "ridge points" are what we call "cut points." Furthermore, "cut points" in [5] are what we would call "points on shortest paths to warped points" (when $d = 2$).

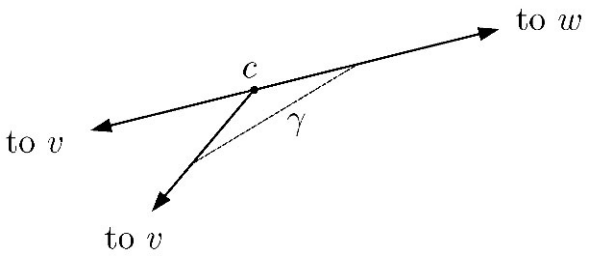


Fig. 4 An intersection that is Y-shaped cannot locally minimize length in $\mathbb{R}^d$ (segment γ is a shortcut).

Proposition 1.2. The resulting Y-shaped intersection at c can be improved upon in a neighborhood of c isometric to an open set in $\mathbb{R}^d$ (Fig. 4), a contradiction. □

Our study of polyhedrality of cut loci will use Voronoi diagrams applied to sets of points from the forthcoming definition, around which the rest of the paper revolves.

Definition 2.3 Suppose that the source point v connects by a shortest path γ to a point x that lies on a facet F or on one of its ridges $R \subset F$, but not on any face of S of dimension $d-2$ or less. If the sequential unfolding of γ into T_F is the segment $[v, x]$, then $v \in T_F$ is called a *source image* for F. Let src_F be the set of source images for F.

Lemma 2.4 *The set* src_F *of source images for any facet* F *of* S *is finite.*

Proof The shortest path in $\mathbb{R}^{d+1}$ between any pair of distinct points x and y in a facet F is the straight segment $[x, y]$. Since this segment is actually contained in S, any shortest path γ in S must contain $[x, y]$ whenever it contains both x and y. Taking x and y to be the first and last points of intersection between γ and the facet F, we find that F can appear at most once in the facet sequence of a shortest path starting at the source point v. Hence there are only finitely many possible facet sequences of shortest paths in S. Now apply Lemma 1.7. □

Example 2.5 Consider a unit cube with a source point in its bottom face, as in Fig. 5. Then the top face has 12 source images, shown in Fig. 5. The four stars "$\star$" are sequential unfoldings of the source point (along three ridges each) that are not source images: each point in the top face is closer to some source image than to any of these stars.

Making Lemma 2.4 quantitative is one of our main open problems; see Section 9.

The next result on the way to Theorem 2.9 generalizes Lemma 3.1 of [27] to arbitrary dimension. Its proof is complicated somewhat by the fact (overlooked in the proof of Lemma 3.1 of [27][4]) that straight segments can lie inside the cut locus, and our lack of a priori knowledge that the cut locus is polyhedral.

[4]Much of [27], but not Lemma 3.1 there, was later incorporated and published in [26].

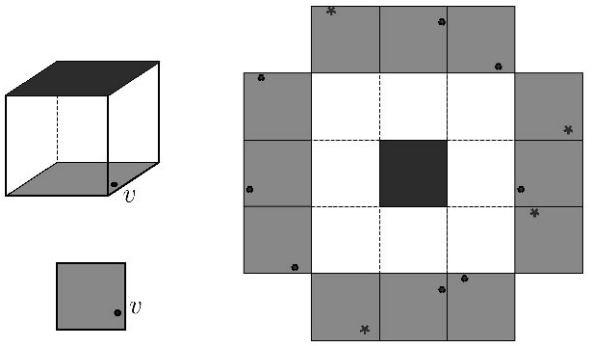


Fig. 5 Source point v on the "bottom" face, 12 source images for the "top" face of a cube and 4 "false" source images (view from the top).

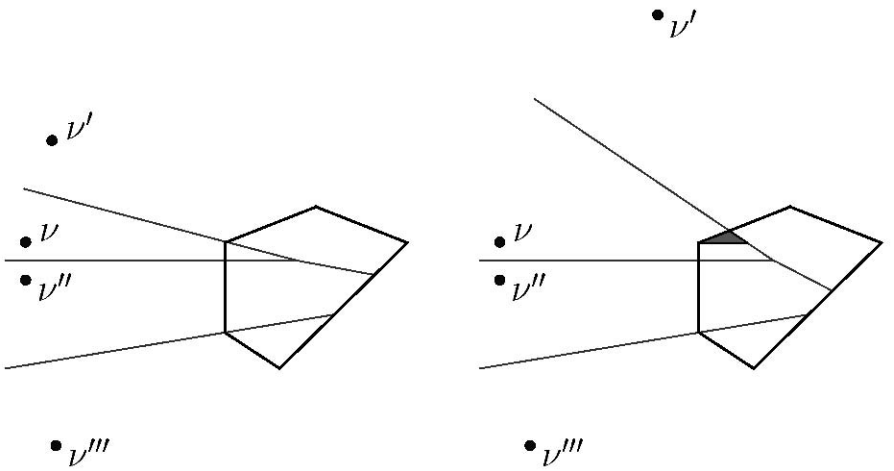


Fig. 6 Generalized Mount's lemma (fails for the shaded region).

Proposition 2.6 (Generalized Mount's lemma) *Let F be a facet of S, and suppose that $v \in \mathrm{src}_F$ is a source image. If $w \in F$, then the straight segment $[v, w] \subset T_F$ has length at least $\mu(v, w)$.*

Example 2.7 In Fig. 6 the left figure is a typical illustration of Proposition 2.6 in dimension $d = 2$: any segment from a source image to $w \in F$ is weakly longer than the one contained in the region with w, and that one sequentially folds to a path in S of length $\mu(v, w)$. In contrast, the right figure will never occur: any point w interior to the shaded region is closest to the source image v, but the straight segment connecting w to v has not been sequentially unfolded along the correct facet sequence.

Proof of Proposition 2.6 Since the two functions $F \to \mathbb{R}$ mapping w to $\mu(v, w)$ and to the length of $[v, w]$ are continuous, we can restrict our attention to those points w lying in any dense subset of F. In particular, the cut locus has dense complement

in F (Corollary 2.2) as does the boundary of F, so we assume throughout that w lies in neither the cut locus nor the boundary of F.

Having fixed $v \in \mathrm{src}_F$, choose a point $x \in F$ as in Definition 2.3, so v connects to x by a shortest path γ that sequentially unfolds to yield the segment $[v, x]$ in T_F. The set $\mathrm{src}_F([x, w])$ of source images sequentially unfolded from shortest paths that end inside the segment $[x, w]$ is finite by Lemma 2.4. Hence we may furthermore assume that w does not lie on any hyperplane H that is equidistant from v and a source image $v' \in \mathrm{src}_F([x, w])$. In other words, we assume w does not lie inside the hyperplane perpendicularly bisecting any segment $[v, v']$.

Claim 2.8 *With these hypotheses, if $v \in \mathrm{src}_F$ but no shortest path unfolds sequentially to the segment $[v, w]$, then w is closer to some point $v' \in \mathrm{src}_F([x, w])$ than to v.*

Assuming this claim for the moment, we may replace v with v' and x with another point x' on $[x, w]$. Repeating this process and again using that the set of source images sequentially unfolded from shortest paths ending in $[x, w]$ is finite, we eventually find that the unique source image $\omega \in \mathrm{src}_F([x, w])$ closest to w is closer to w than v is. Since $[\omega, w]$ has length $\mu(\omega, w)$, it suffices to prove Claim 2.8.

Consider the straight segment $[x, w]$, which is contained in F by convexity. Let Y be the set of points $y \in [x, w]$ having a shortest path γ_y from v that sequentially unfolds to a segment in T_F with endpoint v. Then Y is closed because any limit of shortest paths from v traversing a fixed facet sequence $\mathcal{L}$ is a shortest path that sequentially unfolds along $\mathcal{L}$ to a straight segment from the corresponding source image. Thus, going from x to w, there is a last point $x' \in Y$. This point x' is by assumption not equal to w, so x' must be a cut point (possibly $x = x'$).

There is a facet sequence $\mathcal{L}$ and a neighborhood $\mathcal{O}$ of x' in $[x', w]$ such that every point in $\mathcal{O}$ connects to v by a shortest path traversing $\mathcal{L}$, and such that unfolding the source along $\mathcal{L}$ yields a source image $v' \neq v$ in T_F. This point v' connects to x' by a segment of length $\mu(v, x')$, so the hyperplane H perpendicularly bisecting $[v, v']$ intersects $[x, w]$ at x'. By hypothesis $w \notin H$, and it remains to show that w lies on the side of H closer to v'.

The shortest path from v to x' has a neighborhood in S disjoint from the set of warped points and hence isometric to an open subset of $\mathbb{R}^d$ by Proposition 1.2, because x' is itself not a warped point (we assumed x lies interior to F or to a ridge $R \subset F$). After shrinking $\mathcal{O}$ if necessary, we can therefore ensure that each segment $[v, y]$ for $y \in \mathcal{O}$ is the sequential unfolding of a geodesic η_y in S. The geodesic η_y for $y \in \mathcal{O} \backslash x'$ cannot be a shortest path by definition of x', so $[v, y]$ has length strictly greater than $\mu(v, y)$. We conclude that $\mathcal{O} \backslash x'$, and hence also w, lies strictly closer to v' than to v. This finishes the proof of Claim 2.8 and with it Proposition 2.6.

Before stating the first main result of the paper, we recall the standard notion of *Voronoi diagram* $\mathcal{V}(\Upsilon)$ for a closed discrete set $\Upsilon = \{v, v', \ldots\}$ of points in $\mathbb{R}^d$. This is the subdivision of $\mathbb{R}^d$ whose closed cells are the sets

$$V(\Upsilon, v) = \left\{ \zeta \in \mathbb{R}^d \;\middle|\; \text{every point } v' \in \Upsilon \text{ satisfies } |\zeta - v| \leq |\zeta - v'| \right\}.$$

Thus ζ lies in the interior of $V(\Upsilon, v)$ if ζ is closer to v than to any other point in Υ.

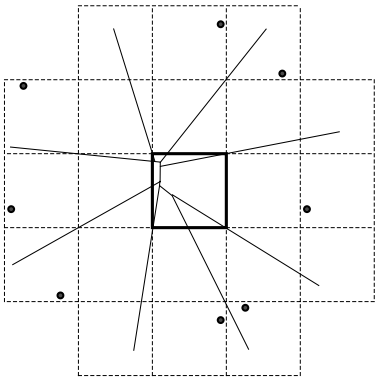


Fig. 7 Cut locus of the "top" face of the cube.

Theorem 2.9 *Fix a facet F of S, and let $V_{d-1} \subseteq T_F$ be the union of the closed cells of dimension $d-1$ in the Voronoi diagram $\mathcal{V}(\mathrm{src}_F)$ for the set of source images in T_F. If F° is the relative interior of F, then the set $F^\circ \cap K_v$ of cut points in F° coincides with the intersection $F^\circ \cap V_{d-1}$. Moreover, if R° is the relative interior of a ridge $R \subset F$, then the set $R^\circ \cap K_v$ of cut points in R° coincides with $R^\circ \cap V_{d-1}$.*

Proof Every shortest path from the source v to a point w in F° or R° unfolds to a straight segment in T_F of length $\mu(v, w)$ ending at a source image for F. Proposition 2.6 therefore says that w lies in the Voronoi cell $V(\Upsilon, v)$ if and only if the segment $[v, w]$ has length exactly $\mu(v, w)$. In particular, v has at least two shortest paths to w if and only if w lies in two such Voronoi cells—that is, $w \in V_{d-1}$. □

To illustrate Theorem 2.9 consider Example 2.5. The Voronoi diagram of source images gives the cut locus in the top face of the cube (see Fig. 7).

Theorem 2.9 characterizes the intersection of the cut locus with faces of dimension d or $d-1$ in S. For faces of smaller dimension, we can make a blanket statement.

Proposition 2.10 *Every warped point lies in the cut locus $\overline{K}_v$; that is, $S_{d-2} \subseteq \overline{K}_v$.*

Proof It is enough to show that every point w in the relative interior of a warped face of dimension $d-2$ is either a cut point or a limit of cut points, because the cut locus $\overline{K}_v$ is closed by definition. Let γ be a shortest path from w to v.

First assume that every neighborhood of w contains a point having no shortest path to v that is a deformation of γ. Suppose that $(y_i)_{i\in\mathbb{N}}$ is a sequence of such points approaching w, with shortest paths $(\gamma_i)_{i\in\mathbb{N}}$ connecting the points y_i to v. Since there are only finitely many facets containing w and finitely many source images for each facet, we may assume (by choosing a subsequence if necessary) that for all i, the sequential unfolding of γ_i connects to the same source image for the same facet. The paths $(\gamma_i)_{i\in\mathbb{N}}$ then converge to a shortest path $\gamma' \neq \gamma$ to w from v, so $w \in K_v$.

Now assume that every point in some neighborhood of w has a shortest path to v that is a deformation of γ. Every point on γ other than w itself is flat in S by Proposition 1.2. Therefore some neighborhood of γ in S is isometric to an open subset of a

product $\mathbb{R}^{d-2} \times C$, where C is a two-dimensional surface that is flat everywhere except at one point $c \in C$ (so C is the boundary of a right circular cone with apex c). The set of points in $\mathbb{R}^{d-2} \times C$ having multiple geodesics to the image of v in $\mathbb{R}^{d-2} \times C$ is a relatively open half-space of dimension $d-1$ whose boundary is $\mathbb{R}^{d-2} \times \{c\}$. Some sequence in this open half-space converges to the image of w. □

Theorem 2.9 and Proposition 2.10 imply the following description of cut loci. For terminology, a subset K of a polyhedral cell complex is called *polyhedral* if its intersection with every facet is a union of convex polyhedra. Note that K need not be a polyhedral cell complex (as defined before Convention 1.1) because it might not come with a cell decomposition. Nonetheless, K can be made into a polyhedral cell complex by suitably subdividing. We call K *pure* of dimension k if it is the closure of a set whose dimension locally near every point is k.

Corollary 2.11 *If v is a source point in S, then*

1 *the cut locus $\overline{K}_v$ is polyhedral and pure of dimension $d-1$, and*
2 *the cut locus $\overline{K}_v$ is the union $K_v \cup S_{d-2}$ of the cut points and warped points.*

Proof Part 2 is a consequence of Theorem 2.9 and Proposition 2.10, the latter taking care of S_{d-2}, and the former showing that points in the cut locus but outside of S_{d-2} are in fact cut points. Since Voronoi diagrams in Euclidean spaces are polyhedral, Theorem 2.9 also implies the polyhedrality in part 1. For the purity, note that if P is any cut point, then the cut set divides a small neighborhood of P into finitely many regions (the regions being determined by the combinatorial types of shortest paths ending therein), with P lying in the closures of at least two of these regions. □

3 Polyhedral Nonoverlapping Unfolding

In this section we again abide by Convention 1.1, so S is the boundary of convex polyhedron P of dimension $d+1$ in $\mathbb{R}^{d+1}$.

Definition 3.1 A polyhedral subset $K \subset S$ of dimension $d-1$ is a *cut set* if K contains the union S_{d-2} of all closed faces of dimension $d-2$, and $S \backslash K$ is open and contractible. A *polyhedral unfolding* of S into $\mathbb{R}^d$ is a choice of cut set K and a map $S \backslash K \to \mathbb{R}^d$ that is an isometry locally on $S \backslash K$. A *nonoverlapping foldout* of S is a surjective piecewise linear map $\varphi\colon \overline{U} \to S$ such that

1. $\overline{U}$ is the closure of its interior U, which is an open topological ball in $\mathbb{R}^d$, and
2. the restriction of φ to U is an isometry onto its image.

Note that K is not required to be a polyhedral subcomplex of S, but only a subset that happens to be a union of polyhedra; thus K can "slice through interiors of facets." The open ball U in item 1 of the definition is usually nonconvex. The polyhedron P is a polytope if and only if $\overline{U}$ is a closed ball—that is, bounded.

When the domain $\overline{U}$ of a nonoverlapping unfolding happens to be polyhedral, so its boundary $\overline{U} \backslash U$ is also polyhedral, the image $K = \varphi(\overline{U} \backslash U)$ is automatically a

cut set in S. Indeed, piecewise linearity of φ implies that K is polyhedral of dimension $d-1$; while the isometry implies that K contains S_{d-2}, and that the open ball $U \cong S\backslash K$ is contractible. Therefore:

Lemma 3.2 *If $\overline{U}$ is polyhedral, then a nonoverlapping foldout $\varphi\colon \overline{U} \to S$ yields an ordinary polyhedral unfolding by taking the inverse of the restriction of φ to U.*

This renders unambiguous the term *polyhedral nonoverlapping unfolding*.

The points in S outside of the $(d-2)$-skeleton S_{d-2} constitute a noncompact flat Riemannian manifold S°. When a point w lies relative interior to a facet F, the tangent space T_w is identified with the tangent hyperplane T_F of F, but when w lies on a ridge, there is no canonical model for T_w.

Most tangent vectors $\zeta \in T_w$ can be exponentiated to get a point $\exp(\zeta) \in S^\circ$ by the usual exponential map from the tangent space T_w to the Riemannian manifold S°. (One can show that the set of tangent vectors that cannot be exponentiated has measure zero in T_w; we shall not use this fact.) In the present case we have a partial compactification S of S°, which allows us to extend this exponential map slightly.

Definition 3.3 Fix a point $w \in S^\circ = S\backslash S_{d-2}$. A tangent vector $\zeta \in T_w$ *can be exponentiated* if the usual exponential of $t\zeta$ exists in S° for all real numbers t satisfying $0 \le t < 1$. In this case, set $\exp(\zeta) = \lim_{t\to 1} \exp(t\zeta)$.

The exponential map $f_\zeta\colon t \to \exp(t\zeta)$ takes the interval $[0, 1]$ to a geodesic $\eta \subset S$, and should be thought of as "geodesic flow" away from w with tangent ζ.

Henceforth fix a source point $v \in S$ not lying on any face of dimension less than d.

Definition 3.4 The *source interior* U_v consists of the tangent vectors $\zeta \in T_v$ at the source point v that can be exponentiated, and such that the exponentials $\exp(t\zeta)$ for $0 \le t \le 1$ do not lie in the cut locus $\overline{K}_v$. The closure of U_v is the *source foldout* $\overline{U}_v$.

Our next main result justifies the terminology for U_v and its closure $\overline{U}_v$.

Theorem 3.5 *Fix a source point v in S. The exponential map* $\exp\colon \overline{U}_v \to S$ *from the source foldout to S is a polyhedral nonoverlapping foldout, and the boundary $\overline{U}_v\backslash U_v$ maps onto the cut locus $\overline{K}_v$. Hence $\overline{K}_v$ is a cut set inducing a polyhedral nonoverlapping unfolding $S\backslash\overline{K}_v \to U_v$ to the source interior.*

Proof It suffices to show the following, in view of parts 1 and 2 from Corollary 2.11:

3. The metric space $S\backslash\overline{K}_v$ is homeomorphic to an open ball.
4. The exponential map $\exp\colon \overline{U}_v \to S$ is piecewise linear and surjective.
5. The exponential map $\exp\colon U_v \to S\backslash\overline{K}_v$ is an isometry.

Every shortest path is the exponential image of some ray in $\overline{U}_v$ by Proposition 1.2, and the set of vectors $\zeta \in \overline{U}_v$ mapping to $S\backslash\overline{K}_v$ is star-shaped by part 2 along with Proposition 1.2 and Corollary 2.2. This implies part 3 and surjectivity in part 4. The space $S^\circ = S\backslash S_{d-2}$ is isometric to a flat Riemannian manifold. Hence the exponential map is a local isometry on any open set of tangent vectors where it is defined. The

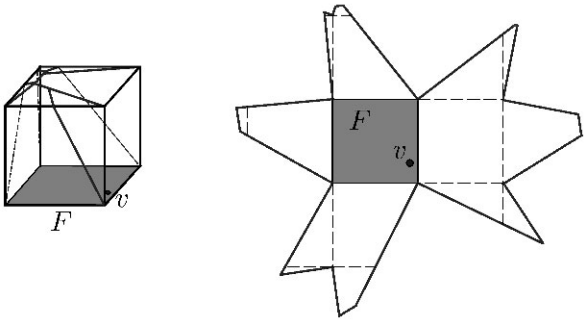


Fig. 8 Cut locus $\overline{K}_v$ and source foldout $\overline{U}_v$ of the cube.

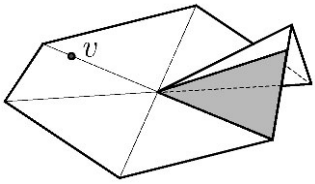


Fig. 9 Shaded region lies outside of $\exp(T_v)$.

definition of $\overline{U}_v$ implies that exp is injective on the interior U_v, so the surjectivity in part 4 shows that exp: $U_v \to S\backslash\overline{K}_v$ is an isomorphism of Riemannian manifolds, proving part 5. Every isometry between two open subsets of affine spaces is linear, so the piecewise linearity in part 4 is a consequence of part 5. □

Example 3.6 Consider a cube P and a source point v located off-center on the bottom face of P, as in Example 2.5 and Fig. 7. The cut locus $\overline{K}_v$ and the corresponding source foldout $\overline{U}_v$ are shown in Fig. 8. See Fig. 1 for the case when v is in the center of the bottom face.

Remark 3.7 Surjectivity of the exponential map does not follow from S° being a Riemannian manifold: convexity plays a crucial role (see Fig. 9 for the case of a nonconvex surface). In fact, surjectivity of exp on a polyhedral manifold is equivalent—in any dimension—to the manifold having positive curvature [34, Lemma 5.1]. Theorem 3.5 extends to the class of *convex polyhedral pseudomanifolds*, but not quite verbatim; see Theorem 7.11 for the few requisite modifications.

4 The Source Poset

In this section we define the source poset (Definition 4.14), and in the next we show how to build it step by step (Theorem 5.2). The reader should consider Definition 4.14 as the main result in this section, although it is the existence and finiteness properties

for minimal jet frames[5] in Theorem 4.11 that endow the source poset with its power to make continuous wavefront expansion combinatorially tractable.

Definition 4.1 Fix a polyhedron V in $\mathbb{R}^d$. Given a list $\bar{\zeta} = (\zeta_1, \ldots, \zeta_r)$ of mutually orthogonal unit vectors in $\mathbb{R}^d$, define for $\varepsilon \in \mathbb{R}$ the unit vector

$$J_{\bar{\zeta}}(\varepsilon) = \frac{\varepsilon\zeta_1 + \cdots + \varepsilon^r\zeta_r}{\sqrt{\varepsilon^2 + \varepsilon^4 + \cdots + \varepsilon^{2r}}}.$$

If $x \in V$ and $x + \varepsilon J_{\bar{\zeta}}(\varepsilon)$ lies in V for all small $\varepsilon > 0$, then the vector-valued function $J_{\bar{\zeta}}$ is a *unit jet* of *order* r at x in V, and $\bar{\zeta}$ is a *partial jet frame* at x along V. If, in addition, $x + \varepsilon J_{\bar{\zeta}}(\varepsilon)$ lies relative interior to V for all small $\varepsilon > 0$, then $\bar{\zeta}$ is a *jet frame*.

The definition will be used later in the case where the convex polyhedron V is a closed Voronoi cell $R \cap V(\mathrm{src}_F, \omega)$ for some ridge R of a facet F, and $\omega \in \mathrm{src}_F$ is a source image. Think of the point $x \in V$ as the closest point in V to ω. It will be important later (but for now may help in understanding the next definition) to note that the relative interior of a polyhedron $V = R \cap V(\mathrm{src}_F, \omega)$ is contained in the relative interior of the ridge R by Definition 2.3 and Theorem 2.9.

We do not assume the polyhedron V has dimension d. However, the order r of a unit jet in V, or equivalently the order of a jet frame along V, is bounded above by the dimension of V. In particular, we allow $\dim(V) = 0$, in which case the only jet frame is empty—that is, a list $\varnothing$ of length zero—and $J_\varnothing \equiv 0$.

The *lexicographic order* on real vectors $\bar{a}$ and $\bar{b}$ of varying lengths is defined by

$$(a_1, \ldots, a_r) < (b_1, \ldots, b_s)$$

if the first nonzero coordinate of $\bar{a} - \bar{b}$ is negative, where by convention we set $a_i = 0$ for $i \geq r + 1$ and $b_j = 0$ for $j \geq s + 1$.

Definition 4.2 Fix a convex polyhedron V in $\mathbb{R}^d$, a point $x \in V$, and an *outer support vector* $\nu \in \mathbb{R}^d$ for V at x, meaning that $\nu \cdot y \leq \nu \cdot x$ for all points $y \in V$. A jet frame $\bar{\zeta}$ at x along V is *minimal* if the *angle sequence* $-(\nu \cdot \zeta_1, \ldots, \nu \cdot \zeta_r)$ is lexicographically smaller than $-(\nu \cdot \zeta'_1, \ldots, \nu \cdot \zeta'_{r'})$ for any jet frame $\bar{\zeta}'$ at x along V.

Again think of $V = R \cap V(\mathrm{src}_F, \omega)$, with $\nu = \omega - x$ being the outer support vector.

In general, that ν is an outer support vector at x means equivalently that x is the closest point in V to $x + \nu$. Minimal jet frames $\bar{\zeta}$ can also be described more geometrically: the angle formed by ν and ζ_1 must be as small as possible, and then the angle formed by ν and ζ_2 must be as small as possible given the angle formed by ν and ζ_1, and so on. It is worth bearing in mind that because ν is an outer support vector, the angle formed by ν and ζ_1 is at least $\pi/2$ (that is, obtuse or right).

[5] The notion of *jet frame* is new; it is motivated by constructions from differential and algebraic geometry, where a *jet* is to a higher-order derivative as a tangent vector is to a first derivative. Our goal is to measure infinitesimal expansion of the wavefront in the directions recorded by jet frames.

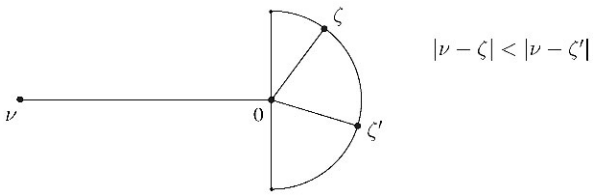


Fig. 10 Dot product vs. length.

Lemma 4.3 *If ζ and ζ' are vectors of equal length in $\mathbb{R}^d$, and $v \in \mathbb{R}^d$ is a vector satisfying $v \cdot \zeta \leq 0$ and $v \cdot \zeta' \leq 0$, then $|v - \zeta| < |v - \zeta'|$ if and only if $v \cdot \zeta > v \cdot \zeta'$.*

Proof Draw v pointing away from the center of the circle containing ζ and ζ', with these vectors on the other side of the diameter perpendicular to v. Then use the law of cosines: the radii ζ and ζ' have equal length, and v has fixed length; only the distances from v to ζ and ζ' change with the angles of ζ and ζ' with v (see Fig. 10). □

Minimal jet frames admit a useful metric characterization as follows.

Proposition 4.4 *Fix two polyhedra V and V' with outer support vectors v and v', of equal length, at points $x \in V$ and $x' \in V'$, respectively. Let $\bar{\zeta}$ and $\bar{\zeta}'$ be partial jet frames at x along V and x' along V', respectively. The angle sequence $-v \cdot \bar{\zeta}$ is smaller than $-v' \cdot \bar{\zeta}'$ in lexicographic order if and only if there exists $\varepsilon_0 > 0$ such that $x + v$ is closer to $x + \varepsilon J_{\bar{\zeta}}(\varepsilon)$ than $x + v'$ is to $x' + \varepsilon J_{\bar{\zeta}'}(\varepsilon)$ for all positive $\varepsilon < \varepsilon_0$.*

Proof Since the dot product of v with each vector $J_{\bar{\zeta}}(\varepsilon)$ or $J_{\bar{\zeta}'}(\varepsilon)$ is negative, and these are unit vectors, it is enough by Lemma 4.3 to show that minimality is equivalent to

$$v \cdot J_{\bar{\zeta}}(\varepsilon) \geq v \cdot J_{\bar{\zeta}'}(\varepsilon) \quad \text{for all nonnegative values of } \varepsilon < \varepsilon_0.$$

If the first nonzero entry of $v \cdot \bar{\zeta} - v \cdot \bar{\zeta}'$ is $c = v \cdot (\zeta_i - \zeta_i')$, then for nonnegative values of ε approaching zero, the difference $v \cdot J_{\bar{\zeta}}(\varepsilon) - v \cdot J_{\bar{\zeta}'}(\varepsilon)$ equals $c\varepsilon^{i-1}$ times a positive function approaching one. The desired result follows easily. □

Corollary 4.5 *Fix an outer support vector v at a point x in a polyhedron V. A jet frame $\bar{\zeta}$ at x along V is minimal if and only if, for every jet frame $\bar{\zeta}'$ at x along V, $x + v$ is weakly closer to $x + \varepsilon J_{\bar{\zeta}}(\varepsilon)$ than to $x + \varepsilon J_{\bar{\zeta}'}(\varepsilon)$ for all small nonnegative ε.*

It is not immediately clear from the definition that minimal jet frames always exist: a priori there could be a continuum of choices for ζ_1, and then a continuum of choices for ζ_2 in such a way that no minimum is attained. Although such continua of choices can indeed occur, we shall see by constructing minimal jet frames explicitly in Theorem 4.11 that a minimum is always attained.

First we need to know more about how (partial) jet frames at x reflect the local geometry of V near x. The *tangent cone* to a polyhedron $V \subseteq \mathbb{R}^d$ at $x \in V$ is the cone

$$T_x V = \mathbb{R}_{\geq 0}\{\zeta \in \mathbb{R}^d \mid x + \zeta \in V\}$$

generated by vectors that land inside V when added to x.

Definition 4.6 Fix a partial jet frame $\bar{\zeta}$ at x along a polyhedron V in $\mathbb{R}^d$. Let $\bar{\zeta}^\perp$ be the linear subspace of $\mathbb{R}^d$ orthogonal to the vectors in $\bar{\zeta}$, and fix a sufficiently small positive real number ε. Then define the *iterated tangent cone*

$$T_x^{\bar{\zeta}} V = T_\xi\big((\xi + \bar{\zeta}^\perp) \cap T_x V\big)$$

as the tangent cone at $\xi = J_{\bar{\zeta}}(\varepsilon)$ to the intersection of $T_x V$ with the affine space $\xi + \bar{\zeta}^\perp$.

Just as the partial jet frames of order 1 generate the tangent cone $T_x V$, we have the following characterization of iterated tangent cones. We omit the easy proof.

Lemma 4.7 *The iterated tangent cone $T_x^{\bar{\zeta}} V$ is generated by all unit vectors ζ_{r+1} in $\mathbb{R}^d$ extending the partial jet frame $\bar{\zeta} = (\zeta_1, \ldots, \zeta_r)$ to a partial jet frame $(\zeta_1, \ldots, \zeta_r, \zeta_{r+1})$ of order $r+1$. In particular, iterated tangent cones do not depend on the small $\varepsilon > 0$.*

Now we set out to construct minimal jet frames inductively.

Lemma 4.8 *Fix a polyhedron V and an outer support vector ν at $x \in V$. If $\zeta \in T_x V$ is a unit vector with $\nu \cdot \zeta$ maximal, then ν is an outer support vector at $\mathbf{0} \in T_x^\zeta V$.*

Proof If $\zeta' \in T_x^\zeta V$ is a unit vector satisfying $\nu \cdot \zeta' > 0$, then $\xi = (\zeta + \varepsilon\zeta')/\sqrt{1+\varepsilon^2}$ for small $\varepsilon > 0$ is a unit vector in $T_x V$ satisfying $\nu \cdot \xi > \nu \cdot \zeta$, contradicting maximality. □

In "generic" cases the functional $\zeta \mapsto \nu \cdot \zeta$ for an outer support vector ν on a cone will take on the maximum value zero uniquely at the origin. In this case, as we now show, there can be only finitely many unit vectors ζ in the cone having $\nu \cdot \zeta$ maximal, and these lie along the *rays*, meaning one-dimensional faces of the cone. Note that genericity forces the cone to be *sharp*, meaning that it contains no linear subspaces.

Proposition 4.9 *Let ν be an outer support vector for a sharp polyhedral cone C, and assume ν is maximized uniquely at the origin $\mathbf{0}$. The minimum angle between ν and a unit vector $\zeta \in C$ occurs when ζ lies on a ray of C.*

Proof Let Z be the set of unit vectors in C. Suppose that L is a two-dimensional subspace inside the span of C, and let $\bar{\nu}$ be the orthogonal projection of ν onto L. View ν and $\bar{\nu}$ as functionals on L via $\zeta \mapsto \nu \cdot \zeta$, and observe that $\nu \cdot \zeta = \bar{\nu} \cdot \zeta$ for all $\zeta \in L$. The circular arc $Z \cap L$ lies inside the unit circle in L, and $\bar{\nu}$ takes nonpositive values on $Z \cap L$ because ν is an outer support vector. Elementary geometry shows that $\bar{\nu}$ is therefore maximized on $Z \cap L$ only at one or both of the endpoints of the arc $Z \cap L$. This argument proves that ν cannot be maximized on Z at a point $\zeta \in Z$ unless ζ lies in the boundary of Z. The result now follows by induction on the dimension of the cone C. □

In "nongeneric" cases, including when the polyhedral cone C has nonzero *lineality*, which is by definition the largest vector space contained in C, the functional ν is maximized along a face of positive dimension. In this case there is always a continuum of choices for unit vectors $\zeta \in C$ having $\nu \cdot \zeta = 0$. However, the sequences of iterated tangent cones to appear in Theorem 4.11 will not in any noticeable way depend on the continuum of choices, because of the next result.

Lemma 4.10 *Fix a polyhedron V, a point $x \in V$, and a face F of V containing x. The iterated tangent cone $T_x^{\bar{\zeta}} V$ is independent of the jet frame $\bar{\zeta}$ for F at x.*

Proof Translate V so $x + \varepsilon J_{\bar{\zeta}}(\varepsilon)$ lies at the origin $\mathbf{0} \in \mathbb{R}^d$. Then F spans a dimension $\dim(F)$ linear subspace $\langle F \rangle \subseteq \mathbb{R}^d$, and the iterated tangent cone is $T_x^{\bar{\zeta}} V = \langle F \rangle^{\perp} \cap T_{\mathbf{0}} V$. Now use the fact that $T_{\mathbf{0}} V = T_{\xi} V$ for all vectors ξ relative interior to F. □

The main theorem in this section says that given an outer support vector ν, there is a finite procedure using elementary linear algebra for producing a single jet frame that is, in a precise sense, tilted as much toward ν as possible.

Theorem 4.11 *Fix a polyhedron V and an outer support vector ν at $x \in V$. Inductively construct a finite set of jet frames for V at x by iterating the following procedure. For each of the finitely many partial jet frames $\bar{\zeta}$ already constructed:*

- *If ν is orthogonal to a nonzero vector in $T_x^{\bar{\zeta}} V$, then add any such vector to $\bar{\zeta}$.*
- *If $\nu \cdot \zeta < 0$ for all nonzero vectors ζ in $T_x^{\bar{\zeta}} V$, then create one new partial jet frame for each of the (finitely many) rays of $T_x^{\bar{\zeta}} V$ minimizing the angle with ν, by appending to $\bar{\zeta}$ the unit vector along that ray.*

At least one of the finitely many jet frames constructed in this way is minimal.

Proof The sequences of vectors constructed by the iterated procedure are jet frames by Lemma 4.8. Given an arbitrary jet frame $\bar{\xi}$ for V at x, it is enough to show that the angle sequence of $\bar{\xi}$ satisfies $\nu \cdot \bar{\zeta} \geq \nu \cdot \bar{\xi}$ in lexicographic order for some constructed jet frame $\bar{\zeta}$. Indeed, then a jet frame whose angle sequence is lexicographically minimal among the constructed ones is minimal. Suppose that the first $i-1$ entries $(\xi_1, \ldots, \xi_{i-1})$ agree with a constructed jet frame, but that $(\xi_1, \ldots, \xi_i)$ do not.

If $\nu \cdot \xi_i < 0$ then $\nu \cdot \xi_i$ is less than $\nu \cdot \zeta_i$ for some constructed jet frame $\bar{\zeta}$ agreeing with $\bar{\xi}$ through the $(i-1)$st entry, by Proposition 4.9.

If, on the other hand, $\nu \cdot \xi_i = 0$, then pick the index j maximal among those satisfying $\xi_j \neq 0$ and also $\nu \cdot \xi_i = \cdots = \nu \cdot \xi_j = 0$. If there is a constructed jet frame $\bar{\zeta}$ that agrees with $\bar{\xi}$ through the jth entry, but has $\nu \cdot \xi_{j+1} < \nu \cdot \zeta_{j+1} = 0$, then we are done already. Therefore we can assume that the constructed jet frame $\bar{\zeta}$ agrees with $\bar{\xi}$ through index $(i-1)$, that $\bar{\zeta}$ has $\nu \cdot \zeta_i = \cdots = \nu \cdot \zeta_j = 0$, and that either $\nu \cdot \zeta_{j+1} < 0$ or else $\bar{\zeta}$ has order j. Replacing the vectors $\xi_i, \ldots, \xi_j$ in $\bar{\xi}$ with $\zeta_i, \ldots, \zeta_j$ yields a new jet frame $\bar{\xi}'$, by Lemma 4.10 applied to the face F of the iterated tangent cone $T_x^{\bar{\zeta}'} V$ orthogonal to ν, where $\bar{\zeta}' = (\zeta_1, \ldots, \zeta_{i-1})$. Downward induction on the number of entries of $\bar{\xi}'$ shared with a constructed jet frame completes the proof. □

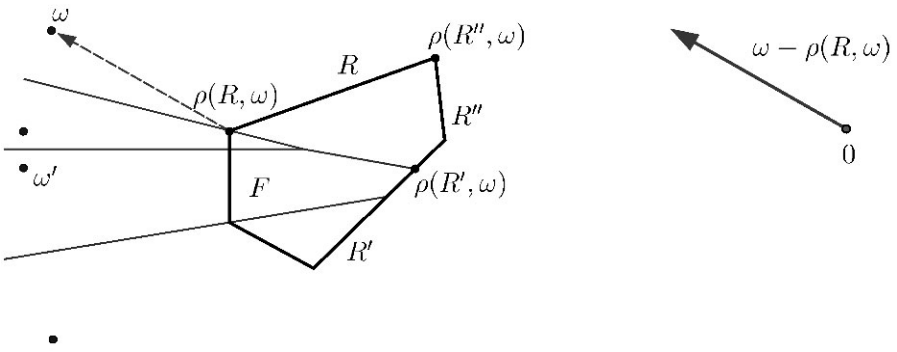


Fig. 11 Illustrations for Definition 4.12.

Our goal is to apply jets to define a poset structure on the set of source images. First, we need some terminology and preliminary concepts. The next definition is made in slightly more generality than required for dealing only with complete sets of source images because we shall need it for Theorem 5.2.

Resume the notation from previous sections regarding the polyhedral complex S. Recall that $T_F \cong \mathbb{R}^d$ is the tangent hyperplane to the facet F. Removing from T_F the affine span T_R of any ridge $R \subset F$ leaves two connected components (open half-spaces). Thus it makes sense to say that a point $v \in T_F \backslash T_R$ lies either on the *same side* or on the *opposite side* of R as does F.

Definition 4.12 Fix a facet F, a ridge $R \subset F$, and a finite set $\Upsilon \subset T_F$.

1. A point $\omega \in \Upsilon$ *can see F through R in $\mathcal{V}(\Upsilon)$* if ω lies on the opposite side of R as F does, and the closed Voronoi cell $V(\Upsilon, \omega)$ contains a point interior to R.
2. A point $\omega \in \Upsilon$ *can see R through F in $\mathcal{V}(\Upsilon)$* if ω lies on the same side of R as F does, and the closed Voronoi cell $V(\Upsilon, \omega)$ contains a point interior to R.
3. In either of the above two cases, the ridge R lies at *radius* $r = r(R, \omega)$ from ω if r equals the smallest distance in T_F from ω to a point of $R \cap V(\Upsilon, \omega)$.
4. The unique *closest point* $\rho(R, \omega)$ to ω in $R \cap V(\Upsilon, \omega)$ has distance r from ω.
5. The *outer support vector* of the pair (R, ω) is $\omega - \rho(R, \omega)$.
6. The *angle sequence* $\angle(R, \omega)$ is the angle sequence $-(\omega - \rho(R, \omega)) \cdot \bar{\zeta}$ for any minimal jet frame $\bar{\zeta}$ at $\rho(R, \omega)$ along $R \cap V(\Upsilon, \omega)$.

Example 4.13 Figure 11 depicts examples of the notions from Definition 4.12. The solid pentagon is the face F, while the set Υ contains four points. The point ω can see F through the ridge R, and can see the ridges R' as well as R'' through F. The three closest points for these are indicated, as is the outer support vector for (R, ω). The point ω' can see the ridge R' through F, but ω' *cannot* see R'' through F, because ω is closer to every point of R''.

In our applications the finite set Υ will always be a subset of source images in src_F, often a proper subset. Now we are ready for the main definition of this section. It may help to recall that each source image $v \in \mathrm{src}_F$ can see F through a

unique ridge R by Theorem 2.9, when $\mathbb{R}^d = T_F$ and the finite set Υ in Definition 4.12 equals src_F.

Definition 4.14 Fix a source point v in S. An *event* is a pair (ν, F) with $\nu \in \mathrm{src}_F$ a source image for the facet F. The event (ν, F) has

1. *radius* $r(\nu, F)$ equal to the radius $r(R, \nu)$ from ν to the ridge R through which ν can see F in the Voronoi subdivision $\mathcal{V}(\mathrm{src}_F)$ of T_F;
2. *event point* $\rho(\nu, F)$ equal to the closest point $\rho(R, \nu)$ in $R \cap V(\mathrm{src}_F, \nu)$ to ν; and
3. *angle sequence* $\angle(\nu, F)$ equal to the angle sequence $\angle(R, \nu)$.

(The trivial event $(v, \mathrm{facet}(v))$ has radius 0, event point v, and empty angle sequence.) The *source poset* $\mathrm{src}(v, S)$ is the set of events, partially ordered with $(\nu, F) \prec (\nu', F')$ if

- $r(\nu, F) < r(\nu', F')$, or if
- $r(\nu, F) = r(\nu', F')$ and $\angle(\nu, F)$ is lexicographically smaller than $\angle(\nu', F')$.

Remark 4.15 Corollary 4.5 says that breaking ties by lexicographically comparing angle sequences at event points is the same as breaking ties by comparing distances from each source image with a minimal jet at its event point. This is the precise sense in which the source poset orders events by comparing infinitesimal expansion of the wavefront along the interiors of ridges containing event points.

5 Constructing Source Images

Aside from its abstract dynamical interpretation, the importance of the source poset here stems from its ability to be computed algorithmically, as we shall see here and in Section 6. Source images are built one by one, using only previously built source images as stepping stones. These stepping stones form an *order ideal* in $\mathrm{src}(v, S)$, meaning a subset $\mathcal{I} \subset \mathrm{src}(v, S)$ closed under going down: $E \in \mathcal{I}$ and $E' \prec E \Rightarrow E' \in \mathcal{I}$.

To make a precise statement in the main result, Theorem 5.2, we need one more dose of terminology, describing constructions in S determined by a choice of order ideal.

Definition 5.1 Fix an order ideal $\mathcal{I}$ in the source poset $\mathrm{src}(v, S)$. For each facet F, let $\Upsilon_F \subset T_F$ be the set of source images $\omega \in \mathrm{src}_F$ with $(\omega, F) \in \mathcal{I}$. The set $\mathcal{E}_\mathcal{I}$ of *potential events* consists of triples (ω, F, R') such that

- ω can see the ridge R' through F in the Voronoi diagram $\mathcal{V}(\Upsilon_F)$, but
- a second facet F' contains R', and the unfolding $\omega' = \Phi_{F,F'}(\omega)$ of ω onto the tangent space $T_{F'}$ results in a pair (ω', F') that does not lie in $\mathcal{I}$.

If (ω', F') is an event in $\mathrm{src}(v, S)\backslash\mathcal{I}$, then we say it is obtained by *processing* (ω, F, R'). A potential event $E \in \mathcal{E}_\mathcal{I}$ is *minimal* if it has minimal radius r among potential events, and lexicographically minimal angle sequence among potential events with radius r.

Tracing back through notation, if $E = (v, F, R')$ is a minimal potential event, then the minimal radius is $r = r(R', v)$, and the minimal angle sequence is $\angle(R', v)$.

Theorem 5.2 *Given a nonempty order ideal $\mathcal{I}$ in the source poset* $\mathrm{src}(v, S)$, *pick a minimal potential event (v, F, R') in $\mathcal{E}_{\mathcal{I}}$. If $v' = \Phi_{F,F'}(v)$ is the unfolding of v to the other facet F' containing R', then $\mathcal{I}' = \mathcal{I} \cup \{(v', F')\}$ is an order ideal in* $\mathrm{src}(v, S)$.

The statement has two parts, really: first, $v' \in T_{F'}$ is indeed a source image; and second, $\mathcal{I}'$ is an order ideal in the poset $\mathrm{src}(v, S)$. To prove the theorem we need a number of preliminaries. We state results requiring an order ideal inside the source poset $\mathrm{src}(v, S)$ using language that assumes an order ideal $\mathcal{I}$ has been fixed.

Recall from Section 1 the notion of facet sequence $\mathcal{L}_\gamma$ for a shortest path γ. If, on the way to a facet F', a shortest path γ from the source point v traverses a facet F, then the corresponding source images in F and F' have a special relationship. Precisely:

Definition 5.3 Let $(v, F) \prec (v', F')$ be events in the source poset. Suppose some shortest path γ has facet sequence $\mathcal{L}_\gamma = (F_1, \ldots, F_{\ell'})$ with a consecutive subsequence

$$\mathcal{L} = (F_\ell, \ldots, F_{\ell'}) \quad \text{in which } F = F_\ell \text{ and } F' = F_{\ell'}.$$

If $v' = \Phi_{\mathcal{L}}(v) = \Phi_{\mathcal{L}_\gamma}(v)$ is the sequential unfolding of the source along γ, and also the sequential unfolding of $v \in T_F$ into $T_{F'}$, then (v, F) *geodesically precedes* (v', F'). We also say that the shortest path γ described above is *geodesically preceded* by (v, F).

Since the Voronoi cells in Theorem 2.9 come up so often, it will be convenient to have easy terminology and notation for them.

Definition 5.4 Given a source image $\omega \in \mathrm{src}_F$, the *cut cell* of ω is $V_\omega = V(\mathrm{src}_F, \omega)$.

Roughly speaking, our next result says that angle sequences increase at successive events along shortest paths, when the event point is pinned at a fixed point x.

Proposition 5.5 *If (v, F) geodesically precedes (v', F') then $(v, F) \prec (v', F')$.*

Proof Because of the way partial order on $\mathrm{src}(v, S)$ is defined, we may as well assume that F and F' share a ridge R', and that $v' = \Phi_{F,F'}(v)$ is obtained by folding along this ridge. In addition, we may as well assume that both event points $\rho(v, F)$ and $\rho(v', F')$ equal the same point $x \in S$, since otherwise $r(v, F) < r(v', F')$. Translate to assume this point x equals the origin $\mathbf{0}$, to simplify notation. Let R be the ridge through which v can see F, and set $V = R \cap V_v$ and $V' = R' \cap V_{v'}$; these are the cut cells through which the source images v and v' see their corresponding facets.

The angle geometry of v' relative to V' in $T_{F'}$ is *exactly* the same as the geometry of v relative to V' in T_F, because v' is obtained by rotation around an axis in $\mathbb{R}^{d+1}$ containing V'. In other words, $v - v'$ is orthogonal to V'. Therefore we need only compare the angles with v of jets along V and V'. All jet frames will be at x.

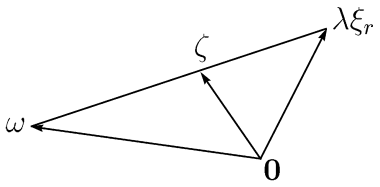


Fig. 12 Geodesic precedence implies a smaller angle sequence.

Suppose the finite sequence $(\xi_1, \xi_2, \ldots)$ is a jet frame along V'. Noting that V and V' have disjoint interiors, choose the index r so that $\bar{\xi} = (\xi_1, \ldots, \xi_{r-1})$ is a partial jet frame along V, but $\bar{\xi}' = (\xi_1, \ldots, \xi_r)$ is not. It is enough to demonstrate that some partial jet frame $(\xi_1, \ldots, \xi_{r-1}, \zeta_r)$ along V has a lexicographically smaller angle sequence than $\bar{\xi}'$. Equivalently, it is enough to produce a unit vector ζ_r in the iterated tangent cone $T_x^{\bar{\xi}} V$ satisfying $\nu \cdot \zeta_r > \nu \cdot \xi_r$.

Since $R' \cap V_{\nu'} = R' \cap V_\nu$ by Theorem 2.9, every line segment from ν to a point in V' passes through V. Therefore, since we have translated to make $x = \mathbf{0}$, every segment connecting ν to $T_x V'$ passes through $T_x V$. This observation will become crucial below; for now, note the resulting inequality $\dim(V) \geq \dim(V')$, which implies that the iterated tangent cone $T_x^{\bar{\xi}} V$ contains nonzero vectors. All such vectors by definition lie in the subspace $\bar{\xi}^\perp$ orthogonal to the space $\langle \bar{\xi} \rangle$ with basis $\xi_1, \ldots, \xi_{r-1}$. The same holds for ξ_r, so we may replace ν with a vector $\omega \in \bar{\xi}^\perp$ by adding a vector in $\langle \bar{\xi} \rangle$, since then

$$\omega \cdot \zeta = \nu \cdot \zeta \quad \text{for all vectors } \zeta \in \bar{\xi}^\perp.$$

Fix a small positive real number ε. The line segment $[\nu, J_{\bar{\xi}'}(\varepsilon)]$ intersects $T_x V$ at a point near $J_{\bar{\xi}}(\varepsilon)$. The image segment in $\bar{\xi}^\perp$ by orthogonal projection modulo $\langle \bar{\xi} \rangle$ is $[\omega, \lambda\xi_r]$, for $\lambda = \varepsilon^r / \sqrt{\varepsilon^2 + \cdots + \varepsilon^{2r}}$. This image segment passes through the cone $T_x^{\bar{\xi}} V$ at some point ζ on its way from ω to $\lambda\xi_r$. Elementary geometry of the triangle with vertices $\mathbf{0}$, ω, and $\lambda\xi_r$ (see Fig. 12) shows that the angle between ω and ζ is smaller than the angle between ω and $\lambda\xi_r$. Taking $\zeta_r = \zeta / |\zeta|$ completes the proof. □

After choosing a minimal potential event E, we must make sure that when all is said and done, none of the other potential events end up below E in the source poset.

Lemma 5.6 *Suppose $(\omega, F, R') \in \mathcal{E}_\mathcal{I}$ is a potential event with angle sequence $\angle$ and radius r. Let F' be the other facet containing R' and let $\omega' = \Phi_{F,F'}(\omega)$ be the unfolding of $\omega \in T_F$ onto $T_{F'}$. If (ω', F') is an actual event, then it either has radius strictly bigger than r, or else its angle sequence $\angle(\omega', F')$ is lexicographically larger than $\angle$.*

Proof Assume that (ω', F') is an event. Quite simply, the result is a consequence of the fact that the cut cell $R' \cap V_{\omega'} = R' \cap V_\omega$ must be contained inside $R' \cap V(\Upsilon_F, \omega)$, which follows because $\Upsilon_F \subseteq \operatorname{src}_F$. □

In comparing a newly processed event (in source poset order) to other as yet unprocessed events, we need to know approximately how those other events will eventually arise. This requires the forthcoming lemma, in which a *flat triangle* inside S is any subset of S isometric to a triangle in the Euclidean plane $\mathbb{R}^2$.

Lemma 5.7 *Fix a point $x \in S$. There is an open neighborhood $\mathcal{O}_x$ of x in S such that, given $y \in \mathcal{O}_x$ and a shortest path γ from the source point v to y, some shortest path γ' from v to x has the following property: the loop formed by traversing γ' and then the segment $[x, y]$ and finally the reverse of γ bounds a flat triangle in S.*

Proof Choose $\mathcal{O}_x$ so small that the only closed faces of the cut locus $\overline{K}_v$ intersecting $\mathcal{O}_x$ are those containing x. Every cut cell containing $y \in \mathcal{O}_x$ also contains x by construction. Convexity of cut cells (Theorem 2.9) implies that the segment $[x, y]$ lies inside every cut cell containing y (there may be more than one if y is itself a cut point). The source image obtained by sequentially unfolding γ therefore connects to every point of $[x, y]$ by a straight segment that sequentially folds to a shortest path. The union of these shortest paths is the flat triangle in question. □

Conveniently, all of the shortest paths to x already yield events in $\mathcal{I}$:

Lemma 5.8 *Suppose some minimal potential event $E \in \mathcal{E}_\mathcal{I}$ has closest point x. Let G be the last facet whose interior is traversed by a shortest path γ from the source point to x. If $\omega \in \mathrm{src}_G$ is the source image sequentially unfolded along γ, then $(\omega, G) \in \mathcal{I}$.*

Proof As γ enters G, it crosses the relative interior of some ridge of G at a point w. The event point $\rho(\omega, G)$ can be no farther than w from ω. On the other hand, $\mu(v, w) < \mu(v, x)$, because γ traverses the interior of G. Therefore (ω, G) has radius less than $r(E) = \mu(v, x)$. □

Proof of Theorem 5.2 Suppose the minimal potential event (v, F, R') has closest point $x = \rho(R', v)$ to $R' \cap V(\Upsilon_F, v)$, of radius r, and a minimal jet frame $\bar{\zeta}$ at x with angle sequence $\angle$.

Let γ be a shortest path from the source that ends at a point in the neighborhood $\mathcal{O}_x$ from Lemma 5.7. By that lemma and Lemma 5.8, γ unfolds to produce a source image whose event either lies in $\mathcal{I}$, or is obtained by processing a potential event in $\mathcal{E}_\mathcal{I}$, or is geodesically preceded by such a processed event. Applying Lemma 5.8 and then Proposition 5.5, we find that all events with event point x that are not in $\mathcal{I}$ have angle sequences lexicographically larger than $\angle$.

For positive ε, set $y(\varepsilon) = x + \varepsilon J_{\bar{\zeta}}(\varepsilon)$. When ε is small enough, $y(\varepsilon)$ lies interior to R', and close to x, in the neighborhood $\mathcal{O}_x$ from Lemma 5.7. By the previous paragraph, every source image containing y in its cut cell is either in $\mathcal{I}$ or has an angle sequence lexicographically larger than $\angle$.

Let us now compare, for all small positive ε, the distance to $y(\varepsilon)$ from v with the distance to $y(\varepsilon)$ from any source image in src_F or $\mathrm{src}_{F'}$. Clearly the distance from a source image ω is minimized when $y(\varepsilon)$ lies in the cut cell V_ω. Moreover, we may restrict our attention to those source images ω whose cut cells V_ω contain $y(\varepsilon)$ for all sufficiently small positive ε. Definition 4.1 says that $\bar{\zeta}$ is a jet frame at x along V_ω.

Therefore, by Proposition 4.4, we conclude using the last sentence of the previous paragraph that $y(\varepsilon)$ is weakly closer to v than to ω for all small positive ε. This argument shows that v' is a source image, so (v', F') is an event. Moreover, it shows:

Claim 5.9 *Any minimal jet frame $\bar{\xi}$ at the event point $x = \rho(R', v)$ along the polyhedron $R' \cap V(\Upsilon_F, v)$ is a minimal jet frame at x along $R' \cap V_{v'}$.*

Every event in $\mathrm{src}(v, S)\backslash\mathcal{I}$ is either obtained by processing a potential event in $\mathcal{E}_{\mathcal{I}}$, or is geodesically preceded by such a processed potential event. Using Claim 5.9, we conclude by Lemma 5.6 and Proposition 5.5 that $\mathcal{I}'$ is an order ideal.

6 Algorithm for Source Unfolding

The primary application of the analysis up to this point is an algorithmic construction of nonoverlapping unfoldings of convex polyhedra, which we present in pseudocode followed by bounds on its running time. In particular, we show that the algorithm is polynomial in the number of source images, when the dimension d is fixed. (Later we state Conjecture 9.2, which posits that the number of source images is polynomial in the number of facets.) Other applications, some of which are further discussed in Section 8, include the discrete geodesic problem (Corollary 6.6) and geodesic Voronoi diagrams (Algorithm 8.1 in Section 8.9).

Roughly, Algorithm 6.1 consists of a single loop that with every iteration constructs one new *event*. Each event is a pair consisting of a facet and a point that we have called a *source image* in the affine span of that facet. The loop is repeated exhaustively until all of the events are computed, so the affine span of every facet has its full complement of source images. The Voronoi diagram for the set of source images in each affine span induces a subdivision of the corresponding facet. For each maximal cell in this subdivision, the algorithm computes a Euclidean motion (composition of rotation and parallel translation) that moves it into the affine span of the facet containing the source point. The union of these moved images of Voronoi cells is the output foldout $\overline{U}_v$ in the tangent space T_v to the source point v.

At each iteration of the loop, the algorithm must choose from a number of *potential events* that it could process into an actual event. Each potential event E consists of an already-computed event (v, F) plus a ridge R in the facet F. Processing the event E applies a rotation to move the source image v into the affine span of the other facet containing R. The potential event that gets chosen must lie as close to the source point as possible; this distance is the radius $r = r(R, v)$ at the beginning of the loop. The loop then calls Routine 6.2 to choose which event to process; although this routine is quite simple in structure, it is the part of the algorithm that most directly encounters the subtlety of working in higher dimensions. The end of the loop consists of updating the sets of source images and potential events; the latter requires Routine 6.3, which we have isolated because it is the only time-consuming part of the algorithm, due to its Voronoi computation.

We emphasize that once a source point v is computed, it is never removed. This claim is part of Theorem 5.2, in which the correctness of Algorithm 6.1—and indeed

the procedure of the algorithm itself—is more or less already implicit, as we shall see in the proof of Theorem 6.4.

We assume that the convex polyhedron P is presented in the input of the algorithm as an intersection of closed half-spaces. Within the algorithm, we omit the descriptions of standard geometric and linear algebraic operations, for which we refer to [18] and [30]. These operations include the determination of lower-dimensional faces (such as ridges) given the facets of P, and the computation of Voronoi diagrams.

Some additional notation will simplify our presentation of the algorithm. Denote by $\mathcal{F}$ and $\mathcal{R}$ the sets of facets and ridges of P, respectively. If a ridge $R \in \mathcal{R}$ lies in a facet $F \in \mathcal{F}$, denote by $\phi(F, R)$ the other facet containing R, so $F \cap \phi(F, R) = R$. Finally, for each facet $F \in \mathcal{F}$, denote by $\widehat{\mathcal{E}}_F$ the set of all triples (v, F, R) such that source point $v \in \Upsilon_F$ lies in the affine span T_F of F, and $R \in \mathcal{R}$ is a ridge contained in F.

Algorithm 6.1 (Computing Source Unfolding)

INPUT convex polyhedron $P \subset \mathbb{R}^{d+1}$ of dimension $d+1$
point v lying in the relative interior of a facet F of P

OUTPUT source foldout of the boundary $S = \partial P$ into $T_v \cong \mathbb{R}^d$ (see Section 3)

DEFINE for each $F \in \mathcal{F}$: a finite set $\Upsilon_F \subset T_F$ of points
for each pair (v, F) satisfying $v \in \Upsilon_F$: an ordered list $\mathcal{L}_{v,F}$ of facets
for each $F \in \mathcal{F}$: a set $\mathcal{E}_F \subset \widehat{\mathcal{E}}_F$ of *potential events*
$\mathcal{E} = \bigcup_{F \in \mathcal{F}} \mathcal{E}_F$, the set of *all potential events*

INITIALIZE for $F \in \mathcal{F}$: if $v \notin F$, then $\Upsilon_F := \varnothing$ and $\mathcal{E}_F = \varnothing$;
otherwise $\Upsilon_F := \{v\}$, $\mathcal{L}_{v,F} := (F)$, $\mathcal{E}_F := \{(v, F, R) \mid R \in \mathcal{R} \text{ and } R \subset F\}$

COMPUTE $\Phi_{F,F'}$ for all $F, F' \in \mathcal{F}$ such that $F \cap F' \in \mathcal{R}$ is a ridge (see Definition 1.5)

WHILE $\mathcal{E} \neq \varnothing$

DO $r := \min\{r(R, v) \mid (v, F, R) \in \mathcal{E}\}$ (see Definition 4.12)
CHOOSE A POTENTIAL EVENT $E = (v, F, R) \in \mathcal{E}$ TO PROCESS
set $F' := \phi(F, R)$, $v' := \Phi_{F,F'}(v)$, $\mathcal{L}_{v',F'} := (\mathcal{L}_{v,F}\, F')$
update $\Upsilon_{F'} \leftarrow \Upsilon_{F'} \cup \{v'\}$
$\mathcal{E}_{F'} \leftarrow \{(\omega, F', R') \in \widehat{\mathcal{E}}_{F'}$ such that $\omega \in \Upsilon_{F'}$, and
POINT $\omega \in \Upsilon_{F'}$ CAN SEE R' THROUGH F', and
$\omega' \notin \Upsilon_G$, where $G = \phi(F', R')$, $\omega' = \Phi_{F',G}(\omega)\}$
$\mathcal{E}_F \leftarrow \mathcal{E}_F \backslash \{E\}$, $\mathcal{E} \leftarrow \bigcup_{G \in \mathcal{F}} \mathcal{E}_G$

END WHILE-DO

COMPUTE for all facets $F \in \mathcal{F}$ and points $v \in \Upsilon_F$:
$\Phi_{\mathcal{L}}$ for $\mathcal{L} = \mathcal{L}_{v,F}$ (see Definition 1.6), and then
$\overline{U}_v(v, F) := \Phi_{\mathcal{L}}^{-1}(F \cap V(\Upsilon_F, v)) \subset T_v$ (see Theorem 2.9)

RETURN the foldout $\overline{U}_v = \bigcup_{(v,F)} \overline{U}_v(v, F)$, the union being over all $F \in \mathcal{F}$ and $v \in \Upsilon_F$

Routine 6.2 (Choose a Potential Event to Process)

INPUT the set $\mathcal{E} = \bigcup_{F \in \mathcal{F}} \mathcal{E}_F$ of potential events, and the radius $r > 0$

OUTPUT an event $E \in \mathcal{E}$ (see Definition 5.1)

COMPUTE the closest potential events $\mathcal{E}_\circ := \{(\omega, F, R) \in \mathcal{E} \mid r(\omega, F) = r\}$
angle sequence $\angle(R, \omega)$ for all $(\omega, F, R) \in \mathcal{E}_\circ$ (see Definition 4.12)
FIND a potential event $E = (\omega, F, R) \in \mathcal{E}_\circ$ with lexicographically
minimal angle sequence $\angle(R, \omega)$ (see Section 4)
RETURN the event $E = (\omega, F, R)$

Routine 6.3 (Point $\omega \in \Upsilon$ Can See R Through F)
INPUT facet $F \in \mathcal{F}$, ridge $R \in \mathcal{R}$, finite set of points $\Upsilon \subset T_F$, and $\omega \in \Upsilon$
OUTPUT boolean variable $\sqsupset \in \{\text{True}, \text{False}\}$ (see Definition 4.12)
COMPUTE Voronoi diagram $\mathcal{V}(\Upsilon)$ (see Section 2)
IF Voronoi cell $V(\Upsilon, \omega) \subset \mathcal{V}(\Upsilon)$ contains a point interior to R
and ω lies on the same side of R as F does in T_F
then $\sqsupset :=$ True;
otherwise $\sqsupset :=$ False
RETURN the variable $\sqsupset$

In the pseudocode we have used the two different symbols "$\leftarrow$" and ":=" to distinguish between those variables that are being updated and those that are being completely redefined at each iteration of the WHILE-DO loop. We hope this clarifies the structure of Algorithm 6.1.

Theorem 6.4 *For every convex polyhedron $P \subset \mathbb{R}^{d+1}$ with boundary $S = \partial P$, and any source point v in a facet of S, Algorithm* 6.1 *computes the source foldout $\overline{U}_v \subseteq T_v$.*

Proof First, we claim by induction that after each iteration of the WHILE-DO loop, the set $\{(v, F) \mid F \in \mathcal{F} \text{ and } v \in \Upsilon_F\}$ is an order ideal in the source poset $\mathrm{src}(v, S)$ from Definition 4.14. The claim is clear at the beginning of the algorithm. By construction, Routine 6.2 picks a minimal potential event E to process. The loop then adds an event by processing E, with the aid of Routine 6.3. Theorem 5.2 implies that what results after processing E is still an order ideal of events, proving our claim. Since the poset $\mathrm{src}(v, S)$ is finite by Lemma 2.4, the algorithm halts after a finite number of loop iterations. Finally, by Theorem 2.9 the Voronoi cells in each facet coincide with the polyhedral subdivision of each facet by the cut locus $\overline{K}_v$, so Theorem 3.5 shows that the foldout in the output is the desired (nonoverlapping) *source foldout* $\overline{U}_v$. □

For purposes of complexity, we assume throughout this paper that the dimension d is fixed. Thus, if the convex polyhedron $P \subset \mathbb{R}^{d+1}$ of dimension d has n facets, so P is presented as an intersection of n closed half-spaces, we can compute all of the vertices and ridges of P in polynomial time [18, 39]. For simplicity, we assume these are precomputed and appended to the input.

The timing of Algorithm 6.1 crucially depends on the number of source images. Let

$$\overline{src}_v := \max_{F \in \mathcal{F}} |\mathrm{src}_F|$$

be the largest number of source images in a tangent plane T_F for a facet F. (This number can change if the source point v is moved. For example, $\overline{src}_v = 4$ if v is in the center of a face, while $\overline{src}_v = 12$ if v is off-center as in Figs. 7 and 8.) Note that

computing Voronoi diagrams for N points in $\mathbb{R}^d$ can be done in $N^{O(d)}$ time [18, p. 381]. See [6, 12], and [17] for details and further references on Voronoi diagrams, and [18] and [30] for other geometric and linear algebraic computations we use.

Theorem 6.5 *When the dimension d is fixed, the cost of Algorithm* 6.1 *is polynomial in the number n of facets and the maximal number $\overline{src}_v$ of source images for a facet.*

Proof From the analysis in the proof of Theorem 6.4, the number of loop iterations is at most $|\mathrm{src}(v, S)| \le |\mathcal{F}|\,\overline{src}_v \le n\,\overline{src}_v$. Within the main body of the algorithm, only standard geometric and linear algebraic operations are used, and these are all polynomial in n. Similarly, Routine 6.2 uses only linear algebraic operations for every potential event $E \in \mathcal{E}$. Note that the cardinality of the set of potential events $\mathcal{E}$ during any iteration of the loop is bounded by $|\mathrm{src}(v, S)| \cdot |\mathcal{F}|^2 \le (n\,\overline{src}_v) \cdot n^2 = n^3\,\overline{src}_v$.

Routine 6.3 constructs Voronoi diagrams $\mathcal{V}(\Upsilon)$ for finite sets $\Upsilon \subset \mathbb{R}^d$. This computation requires $|\Upsilon|^{O(d)} \le (\overline{src}_v)^{O(d)}$ time, which is polynomial for our fixed dimension d. Therefore the total cost of the algorithm is also polynomial in n and $\overline{src}_v$. □

Corollary 6.6 *Let v and w be two points on the boundary S of the convex $(d+1)$-dimensional polyhedron $P \subset \mathbb{R}^{d+1}$, and suppose that v lies interior to a facet. Then the geodesic distance $\mu(v, w)$ on S can be computed in time polynomial in n and $\overline{src}_v$.*

The restriction that v lie interior to a facet is unnecessary, and in fact Algorithm 6.1 can be made to work for arbitrary points v; see Sections 8.8 and 8.9.

Proof Use Algorithm 6.1 to compute the foldout map φ: $\overline{U}_v \to S$. Find $w' \in T_v$ mapping to $w = \varphi(w') \in S$, and compute the distance $|v - w'|$. By the isometry of the exponential map in Theorem 3.5, we conclude that $\mu(v, w) = |v - w'|$. □

Remark 6.7 The complexity of Algorithm 6.1 is exponential in d if the dimension is allowed to grow. For example, the number of vertices of P can be as large as $n^{\Omega(d)}$ [39]. Similarly, the number of cells in Voronoi diagrams of N points in $\mathbb{R}^d$ can be as large as $N^{\Omega(d)}$ [6, 17].

On the other hand, for fixed dimension d Algorithm 6.1 cannot be substantially improved, because the input and the output have costs bounded from below by (a polynomial in) n and $\overline{src}_v$, respectively. This is immediate for the input since P is defined by n hyperplanes. For the output, we claim that the foldout $\overline{U}_v$ in the output of Algorithm 6.1 cannot be presented at a smaller cost because it is a (usually nonconvex) polyhedron that has at least $\overline{src}_v$ *boundary ridges*, meaning faces of dimension $d-1$ in the boundary of $\overline{U}_v$. To see why, let F be a facet with $\overline{src}_v$ source images, and for each $\nu \in \mathrm{src}_F$ consider a shortest path γ_ν whose sequential unfolding into T_F has endpoint ν. If instead we sequentially unfold the paths γ_ν into T_v, we get $|\mathrm{src}_F| = \overline{src}_v$ segments emanating from v. Extend each of these segments to an infinite ray. Some of these infinite rays might pierce the boundary of $\overline{U}_v$ through faces of dimension less than $d-1$, but adjusting their directions slightly ensures that each ray pierces

the boundary of $\overline{U_v}$ through a boundary ridge. These ridges are all distinct because their corresponding rays traverse different facet sequences.

Of course, the efficiency of Algorithm 6.1 does not necessarily imply that it yields an optimal solution to the discrete geodesic problem—or the unfolding problem, for that matter. (The problem of computing *any* nonoverlapping unfolding, not necessarily the source unfolding, is of independent interest in computational geometry [28].) However, although $\overline{src}_v$ is not known to be polynomial in n, we conjecture in Section 9 that it is. See Section 8.10 for more history of the discrete geodesic problem.

Remark 6.8 Following traditions in computational geometry, we have not specified our model of computation. In most computational geometry problems the model is actually irrelevant, since the algorithms are oblivious to it. In our case, however, the situation is more delicate, due to the fact that during each iteration of the loop we make a number of *arithmetic operations* that increase the error. More importantly, we make *comparisons*, which potentially require sharp precision.

Theorem 6.5 and its proof hold as stated for the *complexity over* $\mathbb{R}$ model [9], where there are no errors, and where all arithmetic operations and comparisons have unit cost. While it would be more natural to consider the (usual) *complexity over* $\mathbb{Z}_2$ model [9], arithmetic over $\mathbb{R}$ is unfortunately inherent in the problem: the cut locus, the source unfolding, and geodesic distances can all be irrational.

7 Convex Polyhedral Pseudomanifolds

Recall the notion of polyhedral complex from Section 1. The results in Sections 1–6 hold with relatively little extra work for polyhedral complexes S that are substantially more general than boundaries of polytopes. Since the generality is desirable from the point of view of topology, we complete this extra work here.

Suppose that x is a point in a polyhedral complex S. Denote by

$$S_x(\varepsilon) = \{y \in S \mid \mu(x, y) = \varepsilon\}$$

the *geodesic sphere* in S at radius ε from x. If $\langle x \rangle$ is the smallest face of S containing x, then for sufficiently small positive real numbers ε, the intersection $\langle x \rangle \cap S_x(\varepsilon)$ is an honest (Euclidean) sphere $\langle x \rangle_\varepsilon$ of radius ε around x. The set of points N_x in S near x and equidistant from all points on $\langle x \rangle_\varepsilon$ is the *normal space* at x orthogonal to $\langle x \rangle$ in every face containing x. The *spherical link* of x at radius ε is the set

$$N_x(\varepsilon) = \{y \in N_x \mid \mu(x, y) = \varepsilon\}$$

of points in the normal space at distance ε from x. When ε is sufficiently small, the intersection of $N_x(\varepsilon)$ with any k-dimensional face containing x is a sector inside a sphere of dimension $k - 1 - \dim\langle x \rangle$. The metric μ on S induces a subspace metric on the spherical link $N_x(\varepsilon)$. Always assume ε is sufficiently small when $N_x(\varepsilon)$ is written.

Definition 7.1 Let S be a connected finite polyhedral cell complex of dimension d whose facets all have dimension d. Given a point x inside the union S_{d-2} of all

faces in S of dimension at most $d-2$, we say that S is *positively curved* at x if the spherical link $N_x(\varepsilon)$ is connected and has diameter less than $\pi\varepsilon$. The space S is a *convex*[6] *polyhedral complex* if S is positively curved at every point $x \in S_{d-2}$.

This definition of positive curvature is derived from the one appearing in [34]. It includes as special cases all boundaries of convex polyhedra; this is essentially the content of Proposition 1.2.

Spherical links give local information about geodesics, as noticed by Stone (but see also Section 4.2.2 of [10]).

Lemma 7.2 [34, Lemma 2.2] *Suppose S is a convex polyhedral complex. Then $\tilde{\gamma}$ is a shortest path of length $\alpha\varepsilon$ in the spherical link $N_x(\varepsilon)$ of a point $x \in S$ if and only if the union of all segments connecting points of $\tilde{\gamma}$ to x is isometric (with distances given by the metric on S) to a sector of angle α inside a disk in $\mathbb{R}^2$ of radius ε.*

Although Stone only uses simplicial complexes, we omit the straightforward generalization to polyhedral complexes. Stone's lemma forces shortest paths to avoid low-dimensional faces in the presence of positive curvature.

Proposition 7.3 *Proposition* 1.2 *holds for convex polyhedral complexes S.*

Proof Using notation from Lemma 7.2, suppose that $\alpha < \pi$, and let γ be the segment connecting the endpoints of $\tilde{\gamma}$ through the sector of angle α. Then γ misses x. □

The rest of Section 1 goes through without change for convex polyhedral complexes after we fix, once and for all, a *tangent hyperplane* $T_F \cong \mathbb{R}^d$ for each facet F. The choice of a tangent hyperplane is unique up to isometry. For convenience, we identify F with an isometric copy in T_F, so that (for instance) we may speak as if F is contained inside T_F. This makes Definition 1.5, in particular, work verbatim here.

The main difficulty to overcome in the remainder of Sections 1–6 is the finiteness in Lemma 2.4. In the context of convex polyhedral complexes, this finiteness is fundamental. It comes down to the fact that shortest paths never wind arbitrarily many times around a single face inside of a fixed small neighborhood of a point. The statement of the upcoming Proposition 7.4 would be false if we allowed infinitely many facets, though it could still be made to hold in that case if the sizes of the facets and their dihedral angles were forced to be uniformly bounded away from zero.

Proposition 7.4 *Fix a real number $r \geq 0$ and a convex polyhedral complex S. There is a fixed positive integer $N = N(r, S)$ such that the facet sequence $\mathcal{L}_\gamma$ of each shortest path γ of length r in S has size at most N.*

[6]Using "convex" instead of "positively curved' allows usage of the term "nonconvex polyhedral complex" without ambiguity: "nonpositively curved" is already established in the context of CAT(0) spaces to mean (for polyhedral manifolds, at least) that no point has positive sectional curvature in any direction. In contrast, "nonconvex" means that some point has a negative sectional curvature.

Proof Pick a real number $\varepsilon > 0$ small enough so that the following holds. First, the sphere $S_x(\varepsilon)$ of radius ε centered at each vertex x only intersects faces containing x. Then, for every point x on an edge but outside the union of the radius ε balls around vertices, the sphere $S_x(\varepsilon/2)$ only intersects faces containing x. Iterating, for every point x on a face of dimension i but outside the union of all the previously constructed neighborhoods of smaller-dimensional faces, the sphere $S_x(\varepsilon/2^i)$ only intersects faces containing x. The existence of such a number ε follows from the fact that every facet of S is convex, and that S has finitely many facets (Definition 7.1).

It suffices to prove the lemma with $r = \varepsilon/2^d$. Let y be the midpoint of γ. The closed ball $B_y(\varepsilon/2^{d+1})$ of radius $\varepsilon/2^{d+1}$ centered at y intersects some collection of faces, and among these there is a face of minimal dimension k. Fix a point x_k lying in the intersection of this face with the ball $B_y(\varepsilon/2^{d+1})$. The ball $B_{x_k}(\varepsilon/2^k)$ contains γ by the triangle inequality. However, $B_{x_k}(\varepsilon/2^k)$ might also contain a point x_j on a face of dimension $j < k$. If so, then choose j to be minimal. Iterating this procedure (at most d times) eventually results in a point x on a face of dimension i such that $B_x(\varepsilon/2^i)$ contains γ and only intersects faces containing x.

The metric geometry of S inside the ball $B_x(\varepsilon/2^i)$ is the same as in $B_{x'}(\varepsilon/2^i)$ for every point x' on the smallest face containing x, as long as $B_{x'}(\varepsilon/2^i)$ only intersects faces containing x'. Since S has finitely many faces by Definition 7.1, we reduce to proving the lemma for shortest paths γ after replacing S by the ball $B = B_x(\varepsilon/2^i)$. In fact, we uniformly bound the number of facets traversed by *any* shortest path in B. For simplicity, inflate the metric by a constant factor so that B has radius 2. By a *face of* B we mean the intersection of B with a face of S.

Note that B is isometric to a neighborhood of the apex on the boundary of a right circular cone when the dimension is $d = 2$. In this case shortest paths in B can pass at most once through each ray emanating from x. We conclude that the lemma holds in full (not just for B) when $d = 2$. Using induction on d, we assume that the lemma holds in full for convex polyhedral complexes of dimension at most $d - 1$.

First suppose that x is not a vertex of S, so the smallest face $\langle x \rangle$ containing x has positive dimension. Then B is isometric to a neighborhood of x in the product $\langle x \rangle \times N_x$ of the face $\langle x \rangle$ with the normal space N_x. Projecting γ onto N_x yields a shortest path $\bar{\gamma}$ whose facet sequence in the convex polyhedral complex N_x has the same size as $\mathcal{L}_\gamma$. Induction on d completes the proof in this case.

Now assume that x is a vertex of S. If one of the endpoints of γ is x itself, then γ is contained in some face of B. Hence we may assume from now on that x does not lie on γ. Consider the radial projection from $B \backslash \{x\}$ to the unit sphere $S_x(1)$ centered at x in B. If the image of γ is a point, then again γ lies in a single face; hence we may assume that radial projection induces a bijection from γ to its image curve $\tilde{\gamma}$. Since the geometry of B is scale invariant, every path γ' in $B \backslash \{x\}$ mapping bijectively to $\tilde{\gamma}$ under radial projection has a well-defined facet sequence equal to $\mathcal{L}_\gamma$.

Choose another small real number ε as in the first paragraph of the proof, but with B in place of S. Assume in addition that $\varepsilon < 1/2\pi$. Subdivide $\tilde{\gamma}$ into at least $2^d/\varepsilon$ equal arcs, and use Lemma 7.2 to connect the endpoints of each arc by straight segments in (the cone over $\tilde{\gamma}$ in) B. Lemma 7.2 implies that $\tilde{\gamma}$ has length at most π, because γ is a shortest path. Therefore each of the at least $2^{d+1}\pi$ chords of $\tilde{\gamma}$ has length at most 2^d. The argument in the second paragraph of the proof now produces a new center x' for each chord, and we are assured that $x' \neq x$ because the ε-ball

around x does not contain any of the chords. Hence the smallest face $\langle x' \rangle$ containing x' has positive dimension, and we are done by induction on d as before. □

We shall see in Corollary 7.7 that Proposition 7.4 implies finiteness of the set of source images. However, first we need to introduce the class of polyhedral complexes for which the notion of source image—and hence the rest of Sections 1–6—makes sense.

Definition 7.5 A convex polyhedral complex S of dimension d is a *convex polyhedral pseudomanifold* if S satisfies two additional *pseudomanifold conditions*: (i) each facet is a bounded polytope of dimension d, and (ii) each ridge lies in at most two facets.

Remark 7.6 The "A.D. Aleksandrov spaces with curvature bounded below by 0" of [10] include convex polyhedral pseudomanifolds; see Example 2.9(6) there. Some of our results here, such as surjectivity of exponential maps and nonbranching of geodesics, are general—and essentially local—properties of spaces with curvature bounded below by zero. However, our focus is on decidedly global issues pertaining to the combinatorial and polyhedral nature of convex polyhedral pseudomanifolds, rather than on a local analogy with Riemannian geometry. That being said, many of our results here can be extended to convex "polyhedral" pseudomanifolds with facets of constant positive curvature instead of curvature zero. We leave this extension to the reader.

A flat point in an arbitrary convex polyhedral complex need not have a neighborhood isometric to an open subset of $\mathbb{R}^d$, because more than two facets could meet there. In a convex polyhedral pseudomanifold, on the other hand, every flat point not lying on the topological boundary has a neighborhood isometric to an open subset of $\mathbb{R}^d$. This condition is necessary for even the most basic of our results to hold, including Corollary 2.2 (whose proof works verbatim for convex polyhedral pseudomanifolds), and the definition of source image (which would require modification without it; see Section 8.3).

We would have preferred to avoid the boundedness condition on facets, but the finiteness of the set of source images in Lemma 2.4 can fail without it; see Section 8.6.

Corollary 7.7 *Lemma* 2.4 *holds for convex polyhedral pseudomanifolds* S.

Proof Since every facet is bounded, the lengths of all shortest paths in S are uniformly bounded. Proposition 7.4 therefore implies that there are only finitely many possible facet sequences among all shortest paths in S from the source. □

Corollary 7.7 yields the following consequences, with the same proofs.

Theorem 7.8 *Proposition* 2.6 *on the generalization of Mount's lemma and Theorem* 2.9 *on Voronoi diagrams hold verbatim for convex polyhedral pseudomanifolds* S.

The rest of Section 2 requires slight modification due to the fact that a convex polyhedral pseudomanifold S can have a nonempty topological boundary ∂S.

Proposition 7.9 *Fix a source point v in a convex polyhedral pseudomanifold S. Every warped point lies either in the topological boundary of S or in the cut locus $\overline{K}_v$.*

Proof The same as Proposition 2.10, assuming w is not in the boundary of S. □

In view of Proposition 7.9, the statement of Corollary 2.11 fails for convex polyhedral pseudomanifolds. Instead we get the following, with essentially the same proof.

Corollary 7.10 *If v is a source point in a convex polyhedral pseudomanifold S, then*

1. *$\overline{K}_v \cup \partial S$ is polyhedral and pure of dimension $d-1$, and*
2. *$\overline{K}_v \cup \partial S$ is the union $K_v \cup S_{d-2} \cup \partial S$ of the cut, warped, and boundary points.*

The considerations in Section 3 go through with one small modification: the noncompact flat Riemannian manifold S° is the complement in S of not just the $(d-2)$-skeleton S_{d-2}, but also the topological boundary ∂S of S. The notion of what it means that a tangent vector at $w \in S$ *can be exponentiated* (Definition 3.3) remains unchanged, as long as w lies neither in S_{d-2} nor the boundary of S. Similarly, the notion of *source interior* (Definition 3.4) remains unchanged except that the exponentials $\exp(t\zeta)$ for $0 \le t \le 1$ must lie in neither the cut locus $\overline{K}_v$ nor the boundary ∂S.

Theorem 7.11 *Fix a source point v in the convex polyhedral pseudomanifold S. The exponential map* exp: $\overline{U}_v \to S$ *on the source foldout is a polyhedral nonoverlapping foldout, and the boundary $\overline{U}_v \backslash U_v$ maps onto $\overline{K}_v \cup \partial S$. Hence $\overline{K}_v \cup \partial S$ is a cut set inducing a polyhedral nonoverlapping unfolding $S \backslash (\overline{K}_v \cup \partial S) \to U_v$ to the source interior.*

Proof Using Corollary 7.10 in place of Corollary 2.11, the proof is the same as that of Theorem 3.5, except that every occurrence of $S \backslash \overline{K}_v$ must be replaced by $S \backslash (\overline{K}_v \cup \partial S)$, and the open subspace S° must be defined as $S \backslash (S_{d-2} \cup \partial S)$ instead of $S \backslash S_{d-2}$. □

Corollary 7.12 *Every convex polyhedral pseudomanifold of dimension d is, as a metric space, obtained from a closed, star-shaped, polyhedral ball in $\mathbb{R}^d$ by identifying pairs of isometric boundary components.*

Section 4 concerns local geometry in the context of convex polyhedra, and therefore requires no modification for convex pseudomanifolds, given that all of the earlier results in the paper hold in this more general context.

In Section 5 the only passage that does not seem to work verbatim for convex polyhedral pseudomanifolds is the proof of Proposition 5.5. That proof is presented using language as if F and F' were embedded in the same Euclidean space $\mathbb{R}^{d+1}$, as they are in the case $S = \partial P$. This embedding can be arranged in the general case here by choosing identifications of T_F and $T_{F'}$ as subspaces of $\mathbb{R}^{d+1}$ in such a way that the copies of F and F' intersect as they do in S.

Finally, the algorithm in Section 6 works just as well for convex polyhedral pseudomanifolds, as long as these spaces are presented in a manner that includes the structure of each facet as a polytope and the adjacency relations among facets. For example, folding maps along ridges shared by adjacent facets can be represented as linear transformations after assigning a vector space basis to each tangent hyperplane.

For the record, let us summarize the previous three paragraphs.

Theorem 7.13 *The results in Sections* 4–6 *hold verbatim for convex polyhedral pseudomanifolds S in place of boundaries of convex polyhedra.*

8 Limitations, Generalizations, and History

The main results in this paper are more or less sharp, in the sense that further extension would make certain aspects of them false. In this section we make this sharpness precise, and also point out some alternative generalizations of our results that might hold with requisite modifications. Along the way, we provide more history.

8.1 Polyhedral versus Riemannian

The study of geodesics on convex surfaces, where $d = 2$, goes back to ancient times and has been revived by Newton and the Bernoulli brothers in modern times. The study of explicit constructions of geodesics on two-dimensional polyhedral surfaces was initiated in [23], and is perhaps much older.

The idea of studying the exponential map on polyhedral surfaces goes back to Aleksandrov [3, Section 9.5], who introduced it locally when $d = 2$. He referred to images of lines in the tangent space T_F to a facet F as *quasi-geodesic* lines on the surface, and proved some results on them specific to the dimension $d = 2$. Among his other results was the $d = 2$ case of Proposition 1.2.

A detailed analysis of the cut locus of two-dimensional convex polyhedral surfaces was presented in [37]. This paper, seemingly overlooked in the West, gives a complete description of certain convex regions called "peels" in [5], which can be used to construct source unfoldings. The approach in [37] is inherently two-dimensional and nonalgorithmic.

The study of exponential maps on Riemannian manifolds is classical [21]. Wolter [38] proved properties of cut loci in the Riemannian context that are quite similar to our results describing the cut locus as the closure of the set of cut points. In fact, we could deduce part 2 of our Corollary 2.11 from Lemma 2 of [38]—in the manifold case, at least—using Proposition 1.2 (which has no analogue in Riemannian geometry). The method would be to "smooth out" the warped locus to make a sequence of complete Riemannian manifolds converging (as metric spaces) to the polyhedral complex S, such that the complement of an ever decreasing neighborhood of the warped locus in S is isometric to the corresponding subset in the approximating manifold. Every shortest path to v in S is eventually contained in the bulk complement of the smoothed neighborhood.

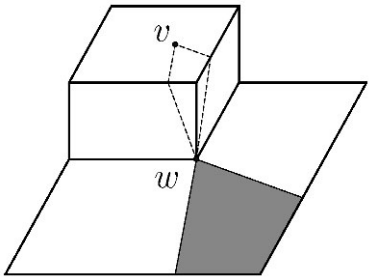


Fig. 13 Points in the shaded region have exactly two shortest paths to v; all of these paths go through the warped point w.

This method does not extend to the polyhedral case where S is allowed to be nonconvex, because Proposition 1.2 fails: shortest paths (between flat points) can pass through warped points (Fig. 13). Moreover, the polyhedrality in the first part of Corollary 2.11 fails systematically when S is allowed to be nonconvex (see [26]).

8.2 Low-Dimensional Flat Faces

We assumed in Definition 7.1 that faces of dimension $d-2$ or less in convex polyhedral complexes must be nontrivially curved. Allowing convex polyhedral complexes where low-dimensional faces can be flat would break the notion of a facet sequence in Corollary 1.4, and would cause the set of warped points to differ from the union of all closed faces of dimension $d-2$, in general. The resulting definitions of folding map and sequential unfolding would be cumbersome if not completely opaque. Nonetheless, the resulting definitions would be possible, because shortest paths would still enter facets (and, in fact, all faces whose interiors are flat) at well-defined angles. The notion of an exponential map would remain unchanged.

Definition 2.3 and Theorem 2.9 should hold verbatim for the modified notion of convex polyhedral pseudomanifold in which low-dimensional flat faces are allowed, because the generalized Mount Lemma (Proposition 2.6) should remain true. Note that Mount's lemma relies mainly on Proposition 1.2 and Corollary 2.2. The latter might be more difficult to verify in the presence of low-dimensional flat faces, because it needs every flat point to have a neighborhood isometric to an open subset of $\mathbb{R}^d$. Thus one might have to assume S is a *manifold*, and not just a pseudomanifold.

Observe that Fig. 6 depends on not having low-dimensional flat faces: it uses the fact that the vertex bordering the shaded region must lie in the cut locus.

8.3 Why the Pseudomanifold Conditions?

Theorem 2.9 fails for convex polyhedral complexes that are not pseudomanifolds, even when there are no flat faces of small dimension. Indeed, with the notion of cut point set forth in Definition 2.1, entire facets could consist of cut points. To see why, suppose there is a cut point interior to a ridge lying on the boundary of three or more facets, and note that the argument using Fig. 4 in the proof of Corollary 2.2

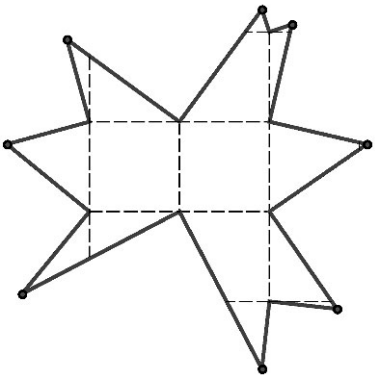


Fig. 14 Aleksandrov unfolding of the cube (the source point v is on the front face, while the left and back faces have no cuts).

fails. For a more concrete construction in dimension $d = 2$, find a convex polyhedral pseudomanifold with a source point so that some edge in the cut locus connects two vertices (for example, take a unit cube with a source point in the center of a facet; see Fig. 1), and then attach a triangle along that edge of the cut locus. The attached triangle ("dorsal fin") consists of cut points.

The proof of Theorem 2.9 fails for nonpseudomanifolds S when we use the thinness of the cut set in the proof of Proposition 2.6. The appropriate definition of cut point x for convex polyhedral complexes more general than pseudomanifolds should say that two shortest paths from x to the source leave x in different directions—that is, they pierce the geodesic sphere $S_x(\varepsilon)$ at different points. However, Corollary 2.2 would still fail for shortest paths entering the "dorsal fin" constructed above.

8.4 Aleksandrov Unfoldings

The dimension $d = 2$ foldouts called "star unfoldings" in [5, 13], and [1] were conceived of by Aleksandrov in Section 6.1 of [3]. Thus we propose here to use the term "Aleksandrov unfolding" instead of "star unfolding," since in any case these foldouts need not be star-shaped polygons. We remark that a footnote in the same section in [3] indicates that Aleksandrov did not realize the nonoverlapping property, which was only established four decades later [5].

Aleksandrov unfoldings are defined for three-dimensional polytopes P similarly to source unfoldings. The idea is again to fix a source point v, but then slice the boundary S of P open along each shortest path connecting v in S to a vertex. An example of the Aleksandrov unfolding of the cube is given in Fig. 14 (see also Fig. 5). Note that when the source point is in the center of the face, the resulting Aleksandrov unfolding agrees with the source unfolding in Fig. 1.

There is a formal connection between source and Aleksandrov unfoldings. Starting from the source unfolding, cut the star-shaped polygon $\overline{U_v}$ into sectors—these are "peels" as in Section 8.1—by slicing along the shortest paths to images of vertices. Rearranging the peels so that the various copies of v lie on the exterior cycle yields a nonoverlapping foldout [5] containing an isometric copy of the bulk of the cut locus.

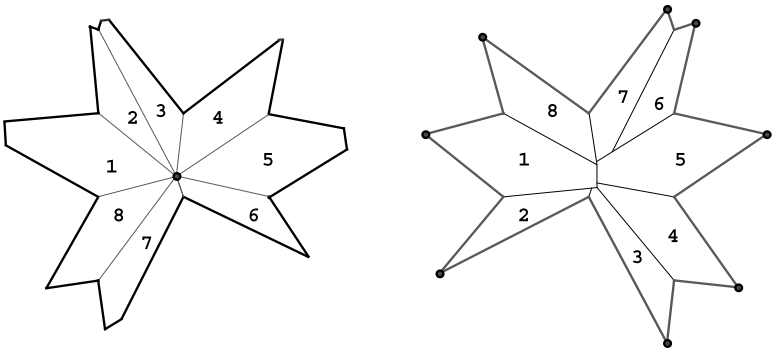


Fig. 15 Source and Aleksandrov unfoldings of the cube, where the corresponding peels in both unfoldings are numbered from 1 to 8.

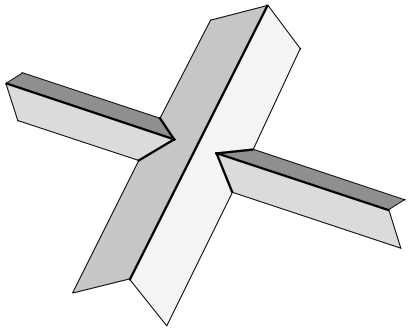


Fig. 16 Shortest paths to warped points making edges look disconnected.

This rearrangement is illustrated in Fig. 15, which continues Example 3.6 (see [1] for further references).

No obvious higher-dimensional analogue of the Aleksandrov unfolding exists, because although the union of all shortest paths connecting the source point to warped points is polyhedral, this complex is not a cut set as per Definition 3.1. Indeed, thinking in terms of source foldouts again, the union of all rays passing from the origin through the images of warped points does not form the $(d-1)$-skeleton of a fan of polyhedral cones. Even when $d = 3$, edges of S closer to the source point can make edges farther away look disconnected, as seen from the source point. An example of how this phenomenon looks from v is illustrated in Fig. 16, where the picture is meant to look like the roof of a building as seen from above.

Circumventing the above failure of Aleksandrov unfoldings in high dimension would necessarily involve dealing with the fact that the set S_{d-2} of warped points generically intersects the cut locus $\overline{K}_v$ in a polyhedral set of dimension $d-3$. This "warped cut locus" usually contains points interior to maximal faces of $\overline{K}_v$, making it impossible for these interiors of maximal cut faces to have neighborhoods in S isometric to open sets in $\mathbb{R}^d$, even locally. Thus the picture in Fig. 15, where most of the cut locus can lie intact in $\mathbb{R}^2$, is impossible in dimension $d \geq 3$. The only

remedy would be to make further slices across the interiors of the maximal faces of the cut locus $\overline{K}_v$ before attempting to lay it flat in $\mathbb{R}^d$. Making these extra slices in a canonical way, to generalize Aleksandrov unfoldings to arbitrary dimension, remains an open problem.

8.5 Definition of Source Image

Some subtle geometry dictated our choice of definition of "source image" (Definition 2.3). With no extra information available, we might alternatively have tried defining src_F as the (finite) set of endpoints of sequentially unfolded shortest paths

- ending at a point interior to F; or
- ending anywhere on F, including at a warped point.

Both look reasonable enough; but the first fails to detect faces of dimension $d-1$ in the cut locus that lie entirely within ridges of S, while the second causes problems with verifying the generalized Mount Lemma (Proposition 2.6) as well as Proposition 5.5 and Lemma 5.8. It is not that the generalized Mount Lemma would be false with these "bonus" source images included, but the already delicate proof would fail. In addition, having these extra source images would add unnecessary bulk to the source poset.

8.6 Finiteness of Source Images

As we saw in Lemma 2.4 for boundaries of polyhedra, or Proposition 7.4 and Corollary 7.7 for convex polyhedral pseudomanifolds, the number of source images is finite. The argument we gave in Lemma 2.4 relies on the embedding of S as a polyhedral complex inside $\mathbb{R}^{d+1}$ in such a way that each face is part of an affine subspace (i.e. not bent or folded). This embedding can be substituted by the more general condition that the polyhedral metric on each facet is induced by the metric on S (so pairs of points on a single facet are the same distance apart in S as in the metric space consisting of the isolated facet). With this extra hypothesis, we would get finiteness of the set of source images even for convex polyhedral pseudomanifolds whose facets were allowed to be unbounded. However, allowing unbounded facets in arbitrary convex polyhedral pseudomanifolds can result in facets with infinitely many source images.

For example, consider an infinite strip in the plane, subdivided into three substrips (one wide and two narrow, to make the picture clearer). Fix a distance $\ell > 0$, and glue each point on one (infinite) boundary edge of the strip to the point ℓ units away from its closest neighbor on the opposite (infinite) edge of the strip. What results is the cylinder S in Fig. 17. This cylinder would be a convex polyhedral manifold if its facets were bounded. The source foldout $\overline{U}_v$ determined by a source point v in the middle of the wide substrip is depicted beneath S, shrunken vertically by a factor of about 2. The cut locus $\overline{K}_v$, which is a straight line along the spine of S, divides each substrip into infinitely many regions, so each substrip has infinitely many source images.

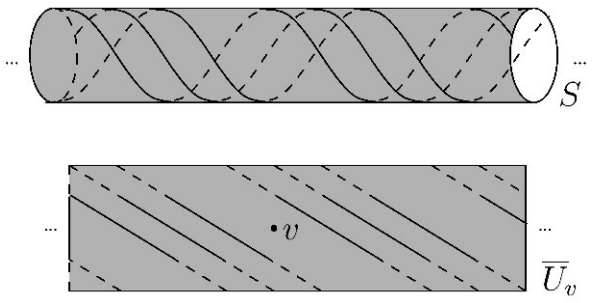


Fig. 17 A source foldout with infinitely many source images.

8.7 Generic Source Points

For generic choices of source point v, the source poset will be a chain—that is, a total order on events. The reason is that moving v infinitesimally changes differently the angles from different source images to ridges containing the same event point, precisely because these source images are sequentially unfolded along shortest paths leaving v in different directions. Note, however, that the *distances* from these various source images to the same event point always remain equal.

8.8 Warped Source Points

We assumed that the source point $v \in S$ lies in the relative interior of some facet; however, nothing really changes when v lies in the relative interior of some ridge. This can be seen by viewing the exponential map as living on the interior flat points $S^\circ \subseteq S$, as in Section 3.

Moreover, simple modifications can generalize the exponential map to the case where v is warped. However, exponentiation on the complement of the cut locus cannot produce a nonoverlapping foldout in $\mathbb{R}^d$ if $v \in S_{d-2}$, because the resulting cut locus would not be a cut set. Indeed, the cut locus would fail to contain all of S_{d-2}, so its complement could not possibly be isometric to an open subset of $\mathbb{R}^d$. On the other hand, exponentiation would instead produce a foldout of S onto the tangent cone to S at v. The main point is that the source point still connects to a dense set of points in S via shortest paths not passing through warped points, by Proposition 1.2.

8.9 Multiple Source Points

Let $\Upsilon = \{v_1, \ldots, v_k\} \subset S$ be a finite set of points on the boundary $S = \partial P$ of a convex polyhedron P. Define the *geodesic Voronoi diagram* $\mathcal{V}_S(\Upsilon)$ to be the subdivision of S whose closed cells are the sets

$$V_S(\Upsilon, v_i) = \big\{w \in S \mid \mu(v_i, w) \le \mu(v_j, w) \text{ for all } 1 \le j \le k\big\}.$$

Just like the (usual) Voronoi diagrams, computing geodesic Voronoi diagrams is an important problem in computational geometry, with both theoretical and practical applications [2, 22, 29] (see [24] and [18] for additional references).

Below we modify Algorithm 6.1 to compute the geodesic Voronoi diagrams in S when multiple source points are input. The modified algorithm outputs subdivisions of the facets of S that indicate which source point is closest. More importantly, it also computes which combinatorial type of geodesic gives a shortest path. In the code below, the WHILE-DO loop and the routines remain completely unchanged. The only differences are in the initial and final stages of the pseudocode.

Algorithm 8.1 (Computing the Geodesic Voronoi Diagram)

INPUT convex polyhedron $P \subset \mathbb{R}^{d+1}$ of dimension $d+1$, and
flat points $v_1, \ldots, v_k$ in the boundary $S = \partial P$

OUTPUT geodesic Voronoi diagram $\mathcal{V}_S(\Upsilon)$ in S

INITIALIZE $\Upsilon_F := \{v_i \mid v_i \in F\}$, $\mathcal{L}_{v_i,F} := (F)$ for $v_i \in \Upsilon_F$, and
$\mathcal{E}_F := \{(v_i, F, R) \in \widehat{\mathcal{E}}_F \mid$ POINT $v_i \in \Upsilon_F$ CAN SEE R THROUGH $F\}$
[. . .]

COMPUTE for each i: $\mathrm{src}_i := \{(v, F) \mid F \in \mathcal{F}, v \in \Upsilon_F$, and $\mathcal{L}_{v,F}$ begins with $F_i\}$, and
for each i: the subset $V_S(\Upsilon, v_i) := \bigcup_{(v,F)\in \mathrm{src}_i} V(\Upsilon_F, v) \cap F$ of S

RETURN geodesic Voronoi diagram $\mathcal{V}_S(\Upsilon) = (V_S(\Upsilon, v_1), \ldots, V_S(\Upsilon, v_k))$

That some of the source points $v_1, \ldots, v_k$ might lie in the same facet necessitates the call to Routine 6.3 in the initialization of $\mathcal{E}_F$. As we did before Theorem 6.5, define $\overline{src}$ to be the maximal number of source images for a single facet.

Theorem 8.2 *Let $P \subset \mathbb{R}^{d+1}$ be a convex polyhedron and let $S = \partial P$, with source points $v_1, \ldots, v_k$ in $S \backslash S_{d-2}$. For fixed dimension d, Algorithm* 8.1 *computes the geodesic Voronoi diagram $\mathcal{V}_S(\Upsilon)$ in time polynomial in k, the number n of facets, and $\overline{src}$.*

The proof is a straightforward extension of the proof of Theorems 6.4 and 6.5; it is omitted. Using observations in Section 8.8, it is possible to modify Algorithm 8.1 to work for a set of arbitrary (that is, possibly warped) source points.

8.10 The Discrete Geodesic Problem

One of our motivating applications for this paper was to the *discrete geodesic problem* of computing geodesic distances and the shortest paths between points v and w in S. The reduction of this problem to computing source unfoldings is easy: construct the source foldout $\overline{U}_v$ in the tangent cone at v, and compute the Euclidian distance between the images.

We should mention here that for $d = 2$ essentially two methods are used in the literature to resolve the discrete geodesic problem: the construction of nonoverlapping unfoldings as above (see [1, 13], and [33]), and the so-called "continuous Dijkstra" method, generalizing Dijkstra's classical algorithm [15] for finding shortest paths in graphs. The second method originated in [26] and is applicable to nonconvex surfaces (see also [20] and [32], where the appendix to the latter paper contains a critique of the former). Interestingly, this method constructs an explicit geodesic wavefront, and then selects and performs "events" one at a time. However, the time-ordering of

events is based on the $d = 2$ fact that the wavefront intersects the union of ridges (edges, in this case) in a finite set of points. Our approach is a combination of these two algorithmic methods, which have previously been separated in the literature. We refer the reader to [24] for more references and results on the complexity of discrete geodesic problems. In general, computing geodesic distances on arbitrary polyhedral complexes remains a challenging problem of both theoretical and practical interest.

9 Open Problems and Complexity Issues

The source poset succeeds at time-ordering the events during wavefront expansion, but it fails to describe accurately how the wavefront bifurcates during expansion, because every event of radius less than r occurs before the first event of radius r in the source poset. On the other hand, the notion of "geodesic precedence" from Definition 5.3 implies a combinatorial structure recording bifurcation exactly.

Definition 9.1 Given a source point v on a convex polyhedral pseudomanifold S, the *vistal tree* $\mathcal{T}(v, S)$ is the set of events, partially ordered by geodesic precedence.

The definition of geodesic precedence immediately implies that $\mathcal{T}(v, S)$ is indeed a rooted tree. It records the facet adjacency graph of the polyhedral decomposition of the source foldout $\overline{U}_v$ into cut cells of dimension d. Equivalently, this data describes the "vista" seen by an observer located at the source point—that is, how the visual field of the observer is locally subdivided by pieces of warped faces. Proposition 5.5 says precisely that the identity map on the set of events induces a poset map from the vistal tree to the source poset. In particular, when the source point is generic as in Section 8.7, the source poset is a linear extension of the vistal tree.

There are numerous interesting questions to ask about the vistal tree, owing to its geometric bearing on the nature of wavefront expansion on convex polyhedra. For example, its size, which is controlled by the extent of branching at each node, is important for reasons of computational complexity (Theorem 6.5).

Conjecture 9.2 *The cardinality* $|\mathrm{src}(v, S)|$ *of the set of source images for a polyhedral boundary* S *is polynomial in the number of facets when the dimension* d *is fixed.*

Hence we conjecture that there is a fixed polynomial f_d, independent of both S and v, such that $|\mathrm{src}(v, S)| < f_d(n)$ for all boundaries $S = \partial P$ of convex polyhedra P of dimension $d + 1$ with n facets, and all source points $v \in S$. Note that the cardinality in question is at most factorial in the number of facets: $|\mathrm{src}(v, S)| < n \cdot (n-1)! = n!$. Indeed, each source image yields a facet sequence, and each of these has length at most n, starts at with facet F containing v, and does not repeat any facet.

To demonstrate the strength of Conjecture 9.2, the following weaker (but perhaps more natural) claim is an immediate consequence.

Conjecture 9.3 *The number of shortest paths joining any pair of points in a polyhedral boundary is polynomial in the number of facets when the dimension is fixed.*

Conjecture 9.3 says that only polynomially many cut cells can meet at a single point, whereas Conjecture 9.2 says there are only polynomially many cut cells in total.

In contrast, we also believe a stronger statement than Conjecture 9.2 holds for boundaries of convex polyhedra. Given a shortest path γ, both of whose endpoints lie interior to facets, call the facet sequence $\mathcal{L}_\gamma$ traversed by γ, the *combinatorial type* of γ (this is called the *edge sequence* in [27] for the $d = 2$ case).

Conjecture 9.4 *The cardinality of the set of combinatorial types of shortest paths in the boundary S of a convex polyhedron is polynomial in the number of facets of S, when the dimension is fixed.*

That is, we do not require one endpoint to be fixed at the source point. The statement is stronger than Conjecture 9.2 because source images are in bijection with combinatorial types of shortest paths in S with endpoint v. In all three of the previous conjectures, the degree of the polynomial will increase with d, even perhaps linearly. When $d = 2$ all three conjectures have been proved (see [1] and [13]).

The intuition for Conjecture 9.2 is that, as seen from the source point in a convex polyhedral boundary S, the faces of dimension $d - 2$ more or less subdivide the horizon into regions. (The horizon is simply the boundary of the source foldout $\overline{U}_v$, as seen from v.) The phrase "more or less" must be made precise, of course; and our inability to delete it altogether is a result of exactly the same phenomenon in Fig. 16 that breaks the notion of Aleksandrov unfoldings in higher dimension.

The reason we believe Conjecture 9.4 is that we believe Conjecture 9.2, and there should not be too many combinatorial types of vistal trees. More precisely, moving the source point a little bit should not alter the combinatorics of the vistal tree, and there should not be more than polynomially many possible vistal trees. In fact, we believe a stronger, more geometric statement. It requires a new notion.

Definition 9.5 Two source points are *equivistal* if their vistal trees are isomorphic, and corresponding nodes represent the same facet sequences.

Again, the facet sequence corresponding to a node of the vistal tree is the list of facets traversed by any shortest path whose sequential unfolding yields the corresponding source image. Hence two source points are equivistal when their views of the horizon look combinatorially the same.

Conjecture 9.6 *The equivalence relation induced by equivistality constitutes a convex polyhedral subdivision of the boundary S of any convex polyhedron. Moreover, the number of open regions in this subdivision is polynomial in the number of facets of S.*

Independent from the conjecture's validity, the *vistal subdivision* it speaks of—whether convex polyhedral or not—is *completely canonical*: it relies only on the metric structure of S. In addition, lower-dimensional strata of the vistal subdivision should reflect combinatorial transitions between neighboring isomorphism classes of vistal trees. Thus Conjecture 9.6 gets at the heart of a number of issues surrounding the interaction of the metric and combinatorial structures of convex polyhedra.

Remark 9.7 An important motivation behind the above ideas lies in the computation of the geodesic diameter of the boundary of a convex polytope. This is a classical problem in computational geometry, not unlike computing diameters of finite graphs (for the $d = 2$ case see [5] and [1]). One possibility, for example, would be to compute the vistal subdivision in Conjecture 9.6, and use this data to list the combinatorial types of shortest paths. Each combinatorial type could then be checked to determine how long its corresponding shortest paths can be. Conjectures 9.4 and 9.6 give hope that the geodesic diameter problem can be solved in polynomial time.

We remark here that the polynomial complexity conjectures fail for nonconvex polyhedral manifolds of dimension $d \geq 2$. Note that this does not contradict the fact that when $d = 2$ there exists a polynomial time algorithm to solve the discrete geodesic problem (see Section 8.10 above). Indeed, the number of source images gives only a lower bound for *our* algorithm, while the problem is resolved by a different kind of algorithm. On the other hand, we show below that for $d \geq 3$ the discrete geodesic problem is NP-hard. The following result further underscores the difference between the convex and nonconvex case.

Proposition 9.8 *On* (*nonconvex*) *polyhedral manifolds*, *the number of distinct combinatorial types of shortest paths can be exponential in the number of facets*. *In addition*, *finding a shortest path on a* (*nonconvex*) *polyhedral manifold is* NP-*hard*.

We present two proofs of the first part: one that is more explicit and works for all $d \geq 2$, and the other that is easy to modify to prove the second part. For the proof of the second part we construct a three-dimensional polyhedral manifold, which is essentially due to Canny and Reif [11]. See Remark 9.9 for comments on how to doctor these manifolds to make them compact and without boundary.

Proof To obtain a polyhedral domain with exponentially many shortest paths between two points x and y, we consider a dimension $d = 2$ example. Simply take a pyramid shape polyhedral surface as shown in Fig. 18 and observe that there exist 2^k shortest paths between top point v and bottom vertex w, where k is the number of terraces in the pyramid. The omitted details are straightforward.

Now consider a dimension $d = 3$ example of a different type. Polyhedrally subdivide $\mathbb{R}^3$ by taking the product of a line $\ell = \mathbb{R}$ with the subdivision of $\mathbb{R}^2$ in Fig. 19. Observe that there are only finitely many cells. Now add $4n$ hyperplanes $H_0, \ldots, H_{4n-1}$ orthogonal to ℓ, and equally spaced along ℓ. This still leaves finitely many convex cells. Between hyperplanes H_{4k} and H_{4k+1}, for all $k = 0 \cdots n-1$, remove all cells *except* the prisms whose bases are the top and bottom triangles in Fig. 19. Similarly, between hyperplanes H_{4k+2} and H_{4k+3}, remove all cells *except* the prisms whose bases are the left and right triangles in Fig. 19.

Now choose x and y to be points on ℓ, with x being on one side of all the hyperplanes, and y being on the other side. Any shortest path connecting x to y must pass alternately through vertical and horizontal pairs of triangular prisms, and there is no preference for which of the two prisms in each pair the shortest path chooses. Thus the number of shortest paths is at least 4^n, while the number of cells is linear in n.

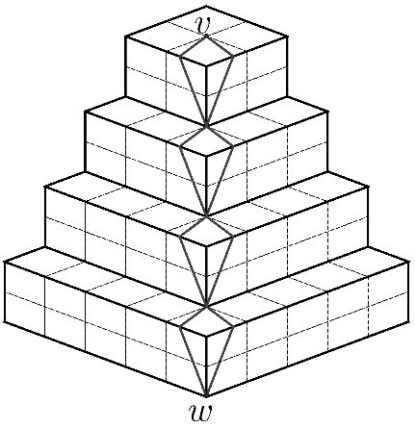


Fig. 18 Nonconvex polyhedral surface in $\mathbb{R}^3$ and shortest paths between points v and w.

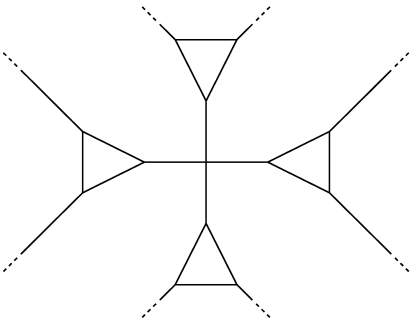


Fig. 19 Polyhedral subdivision of a planar slice in $\mathbb{R}^3$.

For the second part, a construction in [11] presents a polyhedral domain B where the shortest path solution is NP-hard. This domain B is obtained by removing a set of parallel equilateral triangles from $\mathbb{R}^3$. To produce a manifold one has to thicken the triangles into nearly flat triangular prisms. We omit the details. □

Remark 9.9 The polyhedral manifold S in the above proof is noncompact and has nonempty boundary; but with a little extra work, we could accomplish the same effect using a compact polyhedral manifold without boundary. The idea is to draw a large cube C around S in $\mathbb{R}^3$, and place copies C_{top} and C_{bot} of C as the top and bottom facets of a hollow hypercube inside $\mathbb{R}^4$. The remaining six facets of the hollow hypercube are to remain solid. The result is compact, but still has nonempty boundary in C_{top} and C_{bot}. This we fix by building tall three-dimensional prisms in $\mathbb{R}^4$ on the boundary faces, orthogonal to C_{top} and C_{bot}, pointing away from the hypercube. Then we can cap off the prisms with copies of the cells originally excised from $C \subset \mathbb{R}^3$ to get a nonconvex polyhedral 3-sphere in $\mathbb{R}^4$.

Remark 9.10 The reader should not be surprised by the fact that computing the geodesic distance is NP-hard for nonconvex manifolds. On the contrary, in most situations the problem of computing the shortest distance is intractable, and in general is not in NP. We refer to [25] for further hardness results in the geometric context. In a different, more traditional, context, finding the shortest distance in a Cayley graph between two elements in a permutation group (presented by a list of generators in S_N) is known to be NP-hard even for abelian groups [16]. Furthermore, for directed Cayley graphs the problem is PSPACE-complete [19].

Our final conjecture concerns the process of unfolding boundaries of convex polyhedra: if someone provides a polyhedral nonoverlapping foldout made of hinged wood, is it always possible to glue its corresponding edges together? Because wood is rigid, we need not only a nonoverlapping property on the foldout as it lies flat on the ground, but also a nonintersecting property as we continuously fold it up to be glued.

Viewing this process in reverse, can we continuously unfold the polyhedral boundary so that all dihedral angles monotonically increase, until the whole polyhedral boundary lies flat on a hyperplane? This idea was inspired by recent works [8, 14] and was suggested by Connelly.[7] While the monotone increase of the dihedral angles may seem an unnecessary condition justified only by the aesthetics of the blooming, it is in fact crucial in the references above.

As we have phrased things above, we asked for continuous unfolding of an arbitrary nonoverlapping foldout. However, in fact, we only want to ask that there *exist* a foldout that can be continuously glued without self-intersection. Let us be more precise.

Definition 9.11 Let S be the boundary of a convex polyhedron of dimension $d+1$ in $\mathbb{R}^{d+1}$. A *continuous blooming* of S is a choice of nonoverlapping foldout $\overline{U} \to S$, and a homotopy $\{\phi_t \colon \overline{U} \to \mathbb{R}^{d+1} \mid 0 \le t \le 1\}$ such that

1. ϕ_0 is the foldout map $\overline{U} \to S$;
2. ϕ_1 is the identity map on $\overline{U}$;
3. ϕ_t is an isometry from the interior U of $\overline{U}$ to its image, and ϕ_t is linear on each component of the complement of the cut set in each facet, for $0 < t < 1$; and
4. the dihedral angles between corresponding facets of $\phi_t(\overline{U})$ increase as t increases.

An example of a continuous blooming is given in Fig. 20.

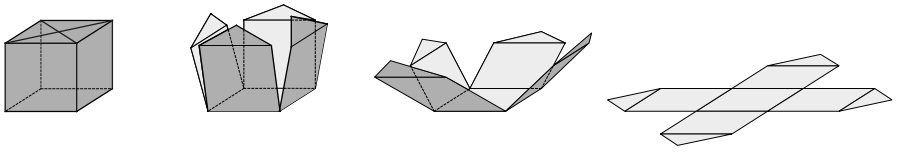


Fig. 20 An example of a continuous blooming of the surface of the cube.

[7] Private communication.

Conjecture 9.12 *Every convex polyhedral boundary has a continuous blooming.*

Even though we ask only for existence, we believe that in fact the source unfolding can be continuously bloomed. As far as we know, this is open even for $d = 2$. Interestingly, we know of no nonoverlapping unfolding that cannot be continuously bloomed, and remain in disagreement on their potential existence.

Acknowledgements We are grateful to Boris Aronov, Yuri Burago, Bob Connelly, Erik Demaine, Maksym Fedorchuk, Robin Forman, Tracy Hall, Bob MacPherson, Jon McCammond, David Mount, János Pach, Micha Sharir, Alexey Tarasov, and Santosh Vempala for helpful conversations. We are indebted to Robin Forman, Joe Malkevitch, and Frank Wolter for references, to Frank Sottile for finding a subtle error in an earlier version, and especially to Günter Rote and (nonanonymous) referees Joe Mitchell and Joseph O'Rourke for a careful reading and numerous helpful comments.

References

1. Agarwal, P.K., Aronov, B., O'Rourke, J., Schevon, C.A.: Star unfolding of a polytope with applications. SIAM J. Comput. **26**(6), 1689–1713 (1997)
2. Aggarwal, A., Guibas, L.J., Saxe, J., Shor, P.W.: A linear-time algorithm for computing the Voronoi diagram of a convex polygon. Discrete Comput. Geom. **4**(6), 591–604 (1989)
3. Aleksandrov, A.D.: Vnutrennyaya geometriya vypuklykh poverkhnostey. Gostekhizdat, Moscow (1948) (in Russian). English translation: Selected Works. Intrinsic Geometry of Convex Surfaces, vol. 2. Chapman & Hall/CRC, Boca Raton (2005)
4. Aleksandrov, A.D.: Vypuklye mnogogranniki. Gostekhizdat, Moscow (1950) (in Russian). English translation: Convex Polyhedra, Springer, Berlin (2005)
5. Aronov, B., O'Rourke, J.: Nonoverlap of the star unfolding. Discrete Comput. Geom. **8**(3), 219–250 (1992)
6. Aurenhammer, F.: Voronoi diagrams—a survey of a fundamental geometric data structure. ACM Comput. Surv. **23**, 345–405 (1991)
7. Bern, M., Demaine, E.D., Eppstein, D., Kuo, E., Mantler, A., Snoeyink, J.: Ununfoldable polyhedra with convex faces. Comput. Geom. **24**, 51–62 (2003)
8. Bezdek, K., Connelly, R.: Pushing disks apart—the Kneser–Poulsen conjecture in the plane. J. Reine Angew. Math. **553**, 221–236 (2002)
9. Blum, L., Cucker, F., Shub, M., Smale, S.: Complexity and Real Computation. Springer, New York (1998)
10. Burago, Yu., Gromov, M., Perelman, A.D.: Alexandrov spaces with curvature bounded below. Russ. Math. Surv. **47**(2), 1–58 (1992)
11. Canny, J.F., Reif, J.H.: New lower bound techniques for robot motion planning problems. Proc. 28th IEEE FOCS, pp. 49–60 (1987)
12. Chazelle, B.: An optimal convex hull algorithm and new results on cuttings. In: Proc. 32nd IEEE FOCS, pp. 29–38 (1991)
13. Chen, J., Han, Y.: Shortest paths on a polyhedron. I. Computing shortest paths. Int. J. Comput. Geom. Appl. **6**(2), 127–144 (1996)
14. Connelly, R., Demaine, E.D., Rote, G.: Straightening polygonal arcs and convexifying polygonal cycles. Discrete Comput. Geom. **30**(2), 205–239 (2003)
15. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numer. Math. **1**, 269–271 (1959)
16. Even, S., Goldreich, O.: The minimum-length generator sequence problem is NP-hard. J. Algorithms **2**(3), 311–313 (1981)
17. Fortune, S.: Voronoi diagrams and Delaunay triangulations. In: Hwang, F., Du, D.Z. (eds.) Computing in Euclidean Geometry, pp. 225–265. World Scientific, Singapore (1995)
18. Goodman, J.E., O'Rourke, J. (eds.): Handbook of Discrete and Computational Geometry. CRC Press, Boca Raton (1997)
19. Jerrum, M.R.: The complexity of finding minimum-length generator sequences. Theor. Comput. Sci. **36**(2–3), 265–289 (1985)
20. Kapoor, S.: An efficient computation of geodesic shortest paths. In: Proc. of the 31st ACM STOC, pp. 770–779 (1999)

21. Kobayashi, S.: On conjugate and cut loci. In: Global Differential Geometry, pp. 140–169. MAA, Washington (1989)
22. Kunze, R., Wolter, F.E., Rausch, T.: Geodesic Voronoi diagrams on parametric surfaces. In: Proc. Comput. Graphics Int., Hasselt-Diepenbeek, Belgium, pp. 230–237 (1997)
23. Lyusternik, L.A.: Geodesic Lines. The Shortest Paths on Surfaces. Gostekhizdat, Moscow (1940) (in Russian)
24. Mitchell, J.S.B.: Geometric shortest paths and network optimization. In: Handbook of Computational Geometry, pp. 633–701. North-Holland, Amsterdam (2000)
25. Mitchell, J.S.B., Sharir, M.: New results on shortest paths in three dimensions. In: Proc. 20th ACM Sympos. Comput. Geom., New York, pp. 124–133 (2004)
26. Mitchell, J.S.B., Mount, D.M., Papadimitriou, C.H.: The discrete geodesic problem. SIAM J. Comput. **16**(4), 647–668 (1987)
27. Mount, D.M.: On finding shortest paths on convex polyhedra. Technical Report 1495, Dept. of Computer Science, University of Maryland, Baltimore, MD (1985)
28. O'Rourke, J.: Folding and unfolding in computational geometry. In: Discrete and Computational Geometry, Tokyo, 1998, pp. 258–266. Springer, Berlin (2000)
29. Papadopoulou, E., Lee, D.T.: A new approach for the geodesic Voronoi diagram of points in a simple polygon and other restricted polygonal domains. Algorithmica **20**(4), 319–352 (1998)
30. Preparata, F.P., Shamos, M.I.: Computational Geometry. An Introduction. Texts and Monographs in Computer Science. Springer, New York (1985)
31. Schlickenrieder, W.: Nets of polyhedra. Diplomarbeit, TU Berlin, Berlin (1997)
32. Schreiber, Y., Sharir, M.: An efficient algorithm for shortest paths on a convex polytope in three dimensions, Preliminary version (see also the extended abstract in Proc. 22nd ACM Sympos. Comput. Geom., Sedona, AZ, 2006.)
33. Sharir, M., Schorr, A.: On shortest paths in polyhedral spaces. SIAM J. Comput. **15**(1), 193–215 (1986)
34. Stone, D.A.: Geodesics in piecewise linear manifolds. Trans. Am. Math. Soc. **215**, 1–44 (1976)
35. Surazhsky, V., Surazhsky, T., Kirsanov, D., Gortler, S.J., Hoppe, H.: Fast exact and approximate geodesics on meshes. ACM Trans. Graph. **24**(3), 553–560 (2005)
36. Tarasov, A.S.: Polyhedra that do not admit natural unfoldings. Russ. Math. Surv. **54**(3), 656–657 (1999)
37. Volkov, J.A., Podgornova, E.G.: The cut locus of a polyhedral surface of positive curvature. Ukrain. Geom. Sb. **11**, 15–25 (1971) (in Russian)
38. Wolter, F.E.: Cut loci in bordered and unbordered Riemannian manifolds. Ph.D. thesis, TU Berlin, FB Mathematik, Berlin, Germany (1985)
39. Ziegler, G.M.: Lectures on Polytopes. Springer, New York (1995)

Empty Simplices of Polytopes and Graded Betti Numbers

Uwe Nagel

Abstract The conjecture of Kalai, Kleinschmidt, and Lee on the number of empty simplices of a simplicial polytope is established by relating it to the first graded Betti numbers of the polytope and applying a result of Migliore and the author. This approach allows us to derive explicit optimal bounds on the number of empty simplices of any given dimension. As a key result, we prove optimal bounds for the graded Betti numbers of any standard graded K-algebra in terms of its Hilbert function.

1 Introduction

Let $P \subset \mathbb{R}^d$ be a simplicial d-polytope, i.e., the d-dimensional convex hull of finitely many points in $\mathbb{R}^d$ such that all its faces are simplices. The simplest combinatorial invariant of P is its f-vector $\underline{f} = (f_{-1}, f_0, \ldots, f_{d-1})$ where $f_{-1} := 1$ and f_i is the number of i-dimensional faces of P if $i \geq 0$. In [14] McMullen conjectured a characterization of the possible f-vectors. In order to state his conjecture we use an equivalent set of invariants, the h-vector $\underline{h} := (h_0, \ldots, h_s)$. It is defined as the sequence of coefficients of the polynomial

$$\sum_{j=0}^{s} h_j z^j := \sum_{j=0}^{d} f_{j-1} \cdot z^j (1-z)^{d-j}.$$

The f-vector can be recovered from the h-vector because

U. Nagel (✉)
Department of Mathematics, University of Kentucky, 715 Patterson Office Tower, Lexington, KY 40506-0027, USA
e-mail: uwenagel@ms.uky.edu

$$f_{j-1} = \sum_{i=0}^{j} \binom{d-i}{j-i} h_i .$$

Using h-vectors we can state McMullen's conjecture which has become a proven statement by combining the results of Billera and Lee [2] and Stanley [20] (see also [15]).

Theorem 1.1 (g-Theorem) *A sequence $\underline{h} = (h_0, \ldots, h_s)$ of positive integers is the h-vector of a simplicial d-polytope if and only if $s = d$ and $\underline{h}$ is an SI-sequence, i.e., $\underline{h}$ satisfies*:

(i) (*Dehn–Sommerville equations*) $h_i = h_{d-i}$ *for* $i = 0, \ldots, d$.
(ii) $\underline{g} := (h_0, h_1 - h_0, \ldots, h_{\lfloor d/2 \rfloor} - h_{\lfloor d/2 \rfloor - 1})$ *is an O-sequence.*

Being an O-sequence is a purely numerical condition (see Sect. 3). Note that O-sequences are precisely the Hilbert functions of Artinian standard graded K-algebras.

In order to prove sufficiency of these conditions, in [2] Billera and Lee construct, for each SI-sequence $\underline{h} := (h_0, \ldots, h_d)$, a certain simplicial d-polytope $P_{\mathrm{BL}}(\underline{h})$ whose h-vector is the given SI-sequence $\underline{h}$. The Billera–Lee polytopes are rather particular which has led to expectations that they have some extremal properties. In order to state one such instance recall (see [11]) that an *empty simplex* of the polytope P is a smallest subset S of the vertex set of P such that S is not a face of P, but each proper subset of S is a face of P. Sometimes, empty simplices are called *missing faces*. They are just minimal non-faces of the vertex set of P. Empty simplices play an important role in the classification of polytopes (see, e.g., [9] and Remark 4.19). In [10] Kalai states as Conjecture 2:

Conjecture 1.2 (Kalai, Kleinschmidt, Lee) *For all simplicial d-polytopes with prescribed h-vector $\underline{h}$, the number of j-dimensional empty simplices is maximized by the Billera–Lee polytope $P_{\mathrm{BL}}(\underline{h})$.*

Kalai has pointed out in Theorem 19.5.35 of [11] that this conjecture is a consequence of results by Migliore and the author [16], but his argument needs some adjustment. The starting point of this note is to point out in detail the connection to the results in [16] that leads to a proof of the conjecture in Theorem 2.3.

The construction of the Billera–Lee polytopes is rather involved. In general, the number of empty j-simplices of a given Billera–Lee polytope $P_{\mathrm{BL}}(\underline{h})$ has not been known. Hence, the proof of Conjecture 1.2 leaves open the problem of giving an explicit bound in terms of the h-vector. The bulk of this paper is devoted to solving this problem. The key is given by our proof of Conjecture 1.2. It identifies the number of missing j-simplices of the polytope P with a certain graded Betti number of its Stanley–Reisner ring $K[P]$. Since the h-vector of P is determined by the Hilbert function of $K[P]$, we are led to consider the problem of finding sharp upper bounds for the graded Betti numbers of the Stanley–Reisner ring $K[P]$ in terms of its Hilbert

function. We solve this problem in Sect. 3 in greater generality, namely for Gorenstein algebras with the Weak Lefschetz property (Theorem 3.17). Its proof requires explicit bounds for all graded Betti numbers of any standard graded K-algebra A in terms of its Hilbert function. These are established in Theorem 3.12. They are optimal. Because of the importance of graded Betti numbers, it seems fair to expect that Theorem 3.12 will find applications in other contexts as well.

In Sect. 4 we apply the results of Sect. 3 to derive explicit optimal bounds for the number of missing j-simplices of a simplicial polytope in terms of its g-vector (see Corollary 4.6). Note that the g-vector is easily obtained from the h-vector (Definition 4.2). We conclude with some applications. Let $N(k)$ be the number of empty j-simplices of a simplicial polytope P such that $j \leq k$. In Theorem 4.15 we establish an upper bound for $N(k)$ that depends on k, the dimension d, and the number of vertices f_0 of the polytope. For each given triple (k, d, f_0), this bound is attained when P is a certain Billera–Lee polytope. For example, the bound for the number of empty edges reads as

$$N(1) \leq \begin{cases} f_0(f_0-3)/2 & \text{if } d=2, \\ \binom{f_0-d}{2} & \text{if } d \geq 3. \end{cases}$$

It is always sharp if the polytope is stacked (see Example 4.18(ii)). McMullen's famous Upper Bound Theorem [13] states that the cyclic polytope $C(f_0, d)$ has the maximal f-vector among all simplicial d-polytopes with f_0 vertices. Our Theorem 4.15 shows that it also has the maximal *total* number of empty simplices among these polytopes (Example 4.18(i)).

As a consequence of Theorem 4.15, for a simplicial d-polytope we obtain a bound for the number $N(k)$ that depends only on k and $f_0 - d$ (Corollary 4.16). Following Kalai [10], such a bound is the key to a central result of Perles [18] in the theory of arbitrary polytopes with "few vertices" (see Remark 4.19). Finally, we show that very little information on the g-vector is sufficient to bound the number of empty j-simplices of a simplicial d-polytope if d is large enough (Corollary 4.22). This result slightly corrects and improves Theorem 3.8 of [10].

2 The Conjecture of Kalai, Kleinschmidt, and Lee

The goal of this section is to prove Conjecture 1.2. To this end we need some more notation. Let P be a simplicial d-polytope. Denote its vertex set by $\{v_1, \ldots, v_{f_0}\}$ and let $R := K[x_1, \ldots, x_{f_0}]$ be the polynomial ring in f_0 variables over an arbitrary field K. Then the *Stanley–Reisner ring* of P is $K[P] := R/I_P$ where the *Stanley–Reisner ideal* is generated by all square-free monomials $x_{i_1}x_{i_2}\cdots x_{i_t}$ such that $\{v_{i_1}, v_{i_2}, \ldots, v_{i_t}\}$ is not a face of P. It is well-known (see Corollary 5.6.5 of [3]) that $K[P]$ is a Gorenstein ring of dimension $d = \dim P$. Since $h_1 = f_0 - d$, its minimal graded free resolution is of the form

$$0 \to \bigoplus_{j\in\mathbb{Z}} R(-j)^{\beta^K_{h_1,j}(P)} \to \cdots \to \bigoplus_{j\in\mathbb{Z}} R(-j)^{\beta^K_{1,j}(P)} \to R \to R/I_P \to 0.$$

The non-negative integers $\beta^K_{i,j}(P) = \dim_K [\mathrm{Tor}^R_i(K[P], K)]_j$, $i, j \in \mathbb{Z}$, are called the *graded Betti numbers* of P.

The following result is shown in [16]:

Theorem 2.1 *Let K be a field of characteristic zero and let P be a simplicial d-polytope with h-vector $\underline{h}$. Then we have for all integers i, j,*

$$\beta_{i,j}^K(P) \leq \beta_{i,j}^K\big(P_{\mathrm{BL}}(\underline{h})\big).$$

Proof The claim is a consequence of Theorem 9.6 of [16], because its proof shows (see page 57) that the extremal polytope that is not specified in part (b) of this theorem is indeed the Billera–Lee polytope $P_{\mathrm{BL}}(\underline{h})$. □

Remark 2.2 The assumption on the characteristic of the field K is needed to ensure that the Stanley–Reisner ring $K[P]$ has the so-called Weak Lefschetz property (see Sect. 3). This property also plays a crucial role in Stanley's necessity part of the g-Theorem in [20].

The conjecture of Kalai, Kleinschmidt, and Lee now follows easily.

Theorem 2.3 *For all simplicial polytopes with prescribed h-vector $\underline{h}$, the number of j-dimensional empty simplices is maximized by the Billera–Lee polytope $P_{\mathrm{BL}}(\underline{h})$.*

Proof It follows from its definition that $\beta_{1,j}^K(P)$ is the number of minimal generators of degree j of the Stanley–Reisner ideal I_P. A j-dimensional empty simplex S of P corresponds to a monomial m_S of degree $j+1$ in I_P. Since each proper subset of S is a face of P, the monomial m_S is not a proper multiple of any monomial in I_P, i.e., m_S is a minimal generator of I_P. Therefore, the conjecture of Kalai, Kleinschmidt, and Lee is a consequence of Theorem 2.1 applied with $i = 1$. □

The combinatorial interpretation of the first Betti numbers allows us to drop the assumption on the characteristic in Theorem 2.1 for certain Betti numbers.

Corollary 2.4 *Let P be a simplicial d-polytope with h-vector $\underline{h}$. Then we have for all integers j,*

$$\beta_{1,j}^K(P) \leq \beta_{1,j}^K\big(P_{\mathrm{BL}}(\underline{h})\big), \qquad \beta_{h_1-1,j}^K(P) \leq \beta_{h_1-1,j}^K\big(P_{\mathrm{BL}}(\underline{h})\big),$$

and

$$\beta_{h_1,j}^K(P) = \begin{cases} 0 & \text{if } j \neq h_1 + d, \\ 1 & \text{if } j = h_1 + d. \end{cases}$$

Proof Denote by $n_j(P)$ the number of empty j-simplices of P. We have seen that, for every field K,

$$n_{j-1}(P) = \beta_{1,j}^K(P).$$

Let K be a field of characteristic zero. Then Theorem 2.3 provides

$$n_{j-1}(P) \leq n_{j-1}\big(P_{\mathrm{BL}}(\underline{h})\big).$$

Let now K be an arbitrary field. Then, applying the above equality again, the claim for the first Betti numbers follows.

Since $K[P]$ is a Gorenstein ring, its minimal free resolution is self-dual. In particular, for all integers i, j, we have

$$\beta_{i,j}^K(P) = \beta_{h_1-i,h_1+d-j}^K(P).$$

This implies the remaining assertions. □

Remark 2.5 Note that the conjecture of Kalai, Kleinschmidt, and Lee has been shown by giving a combinatorial interpretation of the first graded Betti numbers of a simplicial polytope. By duality, it follows that the second last non-trivial graded Betti numbers have a combinatorial interpretation, too. However, it is not possible to find combinatorial interpretations of all graded Betti numbers because, in general, the Betti numbers depend on the characteristic of the ground field (see Example 3.3 of [23]).

3 Upper Bounds for Betti Numbers

The key to proving the conjecture of Kalai, Kleinschmidt, and Lee has been to identify the number of missing i-simplices as a certain first graded Betti number. The results in [16] show that in order to compute an upper bound for this number in terms of the h-vector of the polytope, we need to know an upper bound for the Betti numbers of Cohen–Macaulay algebras. The goal of this section is to establish such bounds. Since the general case does not take more work than the special case of a Cohen–Macaulay algebra, we derive upper bounds for the graded Betti numbers of any arbitrary standard graded K-algebra in terms of its Hilbert function.

The applications in Sect. 4 rely on the results about the first graded Betti numbers of certain Gorenstein algebras. However, we cannot restrict ourselves to considering first Betti numbers in this section. In order to apply Theorem 8.13 in [16], we also need optimal bounds for the last non-trivial Betti numbers of a Cohen–Macaulay algebra. This forces us to discuss simultaneously *all* graded Betti numbers here.

Throughout this section we denote by R the polynomial ring $K[x_1, \ldots, x_n]$ over an arbitrary field K with its standard grading where every variable has degree one. $A \neq 0$ will be a standard graded K-algebra R/I where $I \subset R$ is a proper homogeneous ideal. For a finitely generated graded R-module $M = \bigoplus_{j\in\mathbb{Z}}[M]_j$, we denote its graded Betti numbers by

$$\beta_{ij}^R(M) := \dim_K \left[\operatorname{Tor}_i^R(M, K)\right]_j.$$

Since the graded Betti numbers of M do not change under field extensions of K, we may and will assume that the field K is infinite.

The Hilbert function of M is the numerical function $h_M\colon \mathbb{Z} \to \mathbb{Z}, h_M(j) := \dim_K [M]_j$. The Hilbert functions of graded K-algebras have been completely classified by Macaulay. In order to state his result we need some notation.

Notation 3.1 (i) We always use the following convention for binomial coefficients: If $a \in \mathbb{R}$ and $j \in \mathbb{Z}$, then

$$\binom{a}{j} := \begin{cases} a(a-1)\cdots(a-j+1)/j! & \text{if } j > 0, \\ 1 & \text{if } j = 0, \\ 0 & \text{if } j < 0. \end{cases}$$

(ii) Let b, d be positive integers. Then there are uniquely determined integers $m_d > m_{d-1} > m_s \geq s \geq 1$ such that

$$b = \binom{m_d}{d} + \binom{m_{d-1}}{d-1} + \cdots + \binom{m_s}{s}.$$

This is called the *d-binomial expansion* of b. For any integer j we set

$$b^{\langle d,j\rangle} := \binom{m_d+j}{d+j} + \binom{m_{d-1}+j}{d-1+j} + \cdots + \binom{m_s+j}{s+j}.$$

Of particular importance are the cases where $j = 1$ or $j = -1$. To simplify notation, we further define

$$b^{\langle d\rangle} := b^{\langle d,1\rangle} = \binom{m_d+1}{d+1} + \binom{m_{d-1}+1}{d} + \cdots + \binom{m_s+1}{s+1}$$

and

$$b_{[d]} := b^{\langle d,-1\rangle} = \binom{m_d-1}{d-1} + \binom{m_{d-1}-1}{d-2} + \cdots + \binom{m_s-1}{s-1}.$$

(iii) If $b = 0$, then we put $b^{\langle d\rangle} = b_{[d]} = b^{\langle d,j\rangle} := 0$ for all $j, d \in \mathbb{Z}$.

Recall that a sequence of non-negative integers $(h_j)_{j\geq 0}$ is called an *O-sequence* if $h_0 = 1$ and $h_{j+1} \leq h_j^{\langle j\rangle}$ for all $j \geq 1$. Now we can state Macaulay's characterization of Hilbert functions [12] (see also [19]).

Theorem 3.2 (Macaulay) *For a numerical function $h\colon \mathbb{Z} \to \mathbb{Z}$, the following conditions are equivalent*:

(a) *h is the Hilbert function of a standard graded K-algebra.*
(b) *$h(j) = 0$ if $j < 0$ and $\{h(j)\}_{j\geq 0}$ is an O-sequence.*

For later use we record some formulas for sums involving binomial coefficients.

Lemma 3.3 *For any positive real numbers a, b and every integer $j \geq 0$, there are the following identities*:

(i) $\displaystyle\sum_{k=0}^{j} (-1)^k \binom{a+k-1}{k}\binom{b}{j-k} = \binom{b-a}{j};$

(ii) $\displaystyle\sum_{k=0}^{j}\binom{a+k-1}{k}\binom{b+j-k-1}{j-k}=\binom{a+b+j-1}{j}$;

(iii) $\displaystyle\sum_{k=0}^{j}(-1)^k\binom{a+k}{m}\binom{b}{j-k}=\sum_{k=0}^{m}\binom{a-k-1}{m-k}\binom{b-k-1}{j}$ *if* $0\le m\le a$ *are integers.*

Proof (i) and (ii) are probably standard. In any case they follow immediately by comparing coefficients of power series using the identities $(1+x)^{b-a}=(1+x)^{-a}\cdot(1+x)^b$ and $(1-x)^{-a-b}=(1-x)^{-a}\cdot(1-x)^{-b}$.

To see part (iii), we first use (ii) and finally (i); we get

$$\begin{aligned}\sum_{k=0}^{j}(-1)^k\binom{a+k}{m}\binom{b}{j-k}&=\sum_{k=0}^{j}(-1)^k\binom{b}{j-k}\left\{\sum_{i=0}^{m}\binom{k+i}{i}\binom{a-1-i}{m-i}\right\}\\&=\sum_{i=0}^{m}\binom{a-1-i}{m-i}\left\{\sum_{k=0}^{j}(-1)^k\binom{k+i}{k}\binom{b}{j-k}\right\}\\&=\sum_{k=0}^{m}\binom{a-1-i}{m-i}\binom{b-i-1}{j},\end{aligned}$$

as claimed. □

After these preliminaries we are ready to derive bounds for Betti numbers. We begin with the special case of modules having a d-linear resolution. Recall that the graded module M is said to have a *d-linear resolution* if it has a graded minimal free resolution of the form

$$\cdots\to R^{\beta_i}(-d-i)\to\cdots\to R^{\beta_1}(-d-1)\to R^{\beta_0}(-d)\to M\to 0.$$

Here $\beta_i^R(M)=\sum_{j\in\mathbb{Z}}\beta_{i,j}^R(M):=\beta_i$ is the ith total Betti number of M.

Proposition 3.4 *Let $M\neq 0$ be a graded R-module with a d-linear resolution. Then, for every $i\ge 0$, its i*th *total graded Betti number is*

$$\beta_i^R(M)=\sum_{j=0}^{i}(-1)^j\cdot h_M(d+j)\cdot\binom{n}{i-j}.$$

Proof We argue by induction on i. The claim is clear if $i=0$. Let $i>0$. Using the additivity of vector space dimensions along exact sequences and the induction hypothesis we get

$$\begin{aligned}\beta_i^R(M)&=(-1)^i h_M(d+i)+\sum_{j=0}^{i-1}(-1)^{i-1-j}\cdot\beta_j^R(M)\binom{n-1+i-j}{i-j}\\&=(-1)^i h_M(d+i)\end{aligned}$$

$$
\begin{aligned}
&+\sum_{j=0}^{i-1}(-1)^{i-1-j}\cdot\binom{n-1+i-j}{i-j}\left\{\sum_{k=0}^{j}(-1)^j\cdot h_M(d+k)\binom{n}{j-k}\right\}\\
&=(-1)^i h_M(d+i)\\
&\quad+\sum_{k=0}^{i-1}(-1)^k\cdot h_M(d+k)\left\{\sum_{j=k}^{i-1}(-1)^{i-1-j}\binom{n-1-i-j}{i-j}\binom{n}{j-k}\right\}\\
&=(-1)^i h_M(d+i)\\
&\quad+\sum_{k=0}^{i-1}(-1)^k\cdot h_M(d+k)\left\{\sum_{j=1}^{i-k}(-1)^{j-1}\binom{n+j-1}{j}\binom{n}{i-k-j}\right\}\\
&=(-1)^i h_M(d+i)+\sum_{k=0}^{i-1}(-1)^k\cdot h_M(d+k)\binom{n}{i-k}
\end{aligned}
$$

according to Lemma 3.3(i). Now the claim follows. □

It is amusing and useful to apply this result to a case where we know the graded Betti numbers.

Example 3.5 Consider the ideal $I=(x_1,\ldots,x_n)^d$, where $d>0$. Its minimal free resolution is given by an Eagon–Northcott complex. It has a d-linear resolution and its Betti numbers are (see, e.g., the proof of Corollary 8.14 of [16])

$$\beta_i^R(I)=\binom{d+i-1}{i}\binom{n+d-1}{d+i}.$$

Since the Hilbert function of I is, for all $j\geq 0$,

$$h_I(d+j)=h_R(d+j)=\binom{n+d+j-1}{d+j},$$

a comparison with Proposition 3.4 yields

$$\binom{d+i-1}{i}\binom{n+d-1}{d+i}=\sum_{j=0}^{i}(-1)^j\cdot\binom{n+d+j-1}{d+j}\binom{n}{i-j}. \tag{3.1}$$

Now we compute the graded Betti numbers of lex-segment ideals. Recall that an ideal $I\subset R$ is called a *lex-segment* ideal if, for every d, the ideal $I_{\langle d\rangle}$ is generated by the first $\dim_k[I]_d$ monomials in the lexicographic order of the monomials in R. Here $I_{\langle d\rangle}$ is the ideal that is generated by all the polynomials of degree d in I. For every graded K-algebra $A=R/I$ there is a unique lex-segment ideal $I^{\mathrm{lex}}\subset R$ such that A and R/I^{lex} have the same Hilbert function. For further information on lex-segment ideals we refer to [3].

Lemma 3.6 *Let $I \subset R$ be a proper lex-segment ideal whose generators all have degree d. Consider the d-binomial expansion of $b := h_{R/I}(d)$:*

$$b = \binom{m_d}{d} + \binom{m_{d-1}}{d-1} + \cdots + \binom{m_s}{s}.$$

Then the Betti numbers of $A := R/I$ are for all $i \geq 0$,

$$\begin{aligned}\beta_{i+1}^R(A) &= \beta_{i+1,i+d}^R(A) \\ &= \binom{n+d-1}{d+i}\binom{d+i-1}{d-1} - \sum_{k=s}^{d}\sum_{j=0}^{m_k-k}\binom{m_k-j-1}{k-1}\binom{n-1-j}{i}.\end{aligned}$$

(*Note that according to Notation* 3.1, *the sum on the right-hand side is zero if $b = 0$.*)

Proof Gotzmann's Persistence Theorem [4] implies that the Hilbert function of A is, for $j \geq 0$, $h_A(d+j) = b^{\langle d,j\rangle}$ and that I has a d-linear resolution. Hence Proposition 3.4 in conjunction with formula (3.1) and Lemma 3.3(iii) provides

$$\begin{aligned}\beta_{i+1}^R(A) = \beta_i^R(I) &= \sum_{j=0}^{i}(-1)^j \cdot h_I(d+j)\binom{n}{i-j} \\ &= \sum_{j=0}^{i}(-1)^j\left[\binom{n+d+j-1}{d+j} - b^{\langle d,j\rangle}\right]\binom{n}{i-j} \\ &= \binom{n+d-1}{d+i}\binom{d+i-1}{i} - \sum_{j=0}^{i}(-1)^j\cdot\left[\sum_{k=s}^{d}\binom{m_k+j}{k+j}\right]\binom{n}{i-j} \\ &= \binom{n+d-1}{d+i}\binom{d+i-1}{i} - \sum_{k=s}^{d}\left[\sum_{j=0}^{i}(-1)^j\cdot\binom{m_k+j}{m_k-k}\binom{n}{i-j}\right] \\ &= \binom{n+d-1}{d+i}\binom{d+i-1}{i} - \sum_{k=s}^{d}\sum_{j=0}^{m_k-k}\binom{m_k-j-1}{k-1}\binom{n-1-j}{i},\end{aligned}$$

as claimed. □

The above formulas simplify in the extremal cases.

Corollary 3.7 *Adopt the notation and assumptions of Lemma* 3.6. *Then*

(a) $\beta_1^R(A) = \binom{n+d-1}{d} - b$;

(b) $\beta_n^R(A) = \binom{n+d-2}{d-1} - b_{[d]}$.

Proof Part (a) being clear, we restrict ourselves to showing (b). Since $\binom{n-1-j}{n-1} = 0$ for $j > 0$, Lemma 3.6 immediately gives $\beta_n^R(A) = \binom{n+d-2}{d-1} - \sum_{k=s}^{d} \binom{m_k-1}{k-1} = \binom{n+d-2}{d-1} - b_{[d]}$. □

Now, we can compute the non-trivial graded Betti numbers of an arbitrary lex-segment ideal. The basic idea is to reduce the computation to the special case treated in Lemma 3.6 by exploiting algebraic properties of lex-segment ideals.

Proposition 3.8 *Let $I \subset R$ be an arbitrary proper lex-segment ideal and let $d \geq 2$ be an integer. Set $A := R/I$ and consider the d-binomial expansion*

$$h_A(d) =: \binom{m_d}{d} + \binom{m_{d-1}}{d-1} + \cdots + \binom{m_s}{s}$$

and the $(d-1)$-binomial expansion

$$h_A(d-1) =: \binom{n_{d-1}}{d-1} + \binom{n_{d-2}}{d-2} + \cdots + \binom{n_t}{t}.$$

Then we have for all $i \geq 0$

$$\beta_{i+1,i+d}^R(A) = \beta_{i+1,i+d}(h_A, n),$$

where

$$\beta_{i+1,i+d}(h_A, n) := \sum_{k=t}^{d-1} \sum_{j=0}^{n_k-k} \binom{n_k - j}{k} \binom{n-1-j}{i} - \sum_{k=s}^{d} \sum_{j=0}^{m_k-k} \binom{m_k - 1 - j}{k-1} \binom{n-1-j}{i}.$$

Proof As noted above, since I is a lex-segment ideal, for every $j \in \mathbb{Z}$, the ideal $I_{\langle j \rangle}$ has a j-linear resolution, i.e., the ideal I is componentwise linear. Hence, Proposition 1.3 of [7] gives for all $i \geq 0$,

$$\beta_{i+1,i+d}^R(A) = \beta_{i+1}^R(R/I_{\langle d \rangle}) - \beta_{i+1}^R(R/\mathfrak{m}I_{\langle d-1 \rangle}), \tag{3.2}$$

where $\mathfrak{m} = (x_1, \ldots, x_n)$ is the homogeneous maximal ideal of R.

Since $I_{\langle d-1 \rangle}$ is generated in degree $d-1$, the ideals $I_{\langle d-1 \rangle}$ and $\mathfrak{m}I_{\langle d-1 \rangle}$ have the same Hilbert function in all degrees $j \geq d$. Thus, using the assumption $d \geq 2$, Gotzmann's Persistence Theorem [4] provides

$$h_{R/\mathfrak{m}I_{\langle d-1 \rangle}}(d-1+j) = h_{R/I_{\langle d-1 \rangle}}(d-1+j) = h_A(d-1)^{\langle d-1, j \rangle} \quad \text{for all } j \geq 1.$$

It is easy to see that $\mathfrak{m}I_{\langle d-1 \rangle}$ has a d-linear resolution because $I_{\langle d-1 \rangle}$ has a $(d-1)$-linear resolution. Hence, as in the proof of Lemma 3.6, Proposition 3.4 provides

$$\beta_{i+1}^R(R/\mathfrak{m}I_{\langle d-1 \rangle}) = \binom{n+d-1}{d+i} \binom{d+i-1}{d-1} - \sum_{k=t}^{d-1} \sum_{j=0}^{n_k-k} \binom{n_k - j}{k} \binom{n-1-j}{i}.$$

Plugging this and the result of Lemma 3.6 into the formula (3.2), we get our claim. □

Again, the formula simplifies in the extremal cases. We use the result in the following section.

Corollary 3.9 *Adopt the notation and assumptions of Proposition* 3.8. *Then*:

(a) $\beta^R_{1,d}(A) = \beta_{1,d}(h_A, n) = h_A(d-1)^{\langle d-1\rangle} - h_A(d)$.
(b) $\beta^R_{n,n-1+d}(A) = \beta_{n,n-1+d}(h_A, n) = h_A(d-1) - (h_A(d))_{[d]}$.

Proof This follows from the formula given in Proposition 3.8. □

In Proposition 3.8 we left out the case $d \leq 1$ which is easy to deal with. We need:

Definition 3.10 Let h be the Hilbert function of a graded K-algebra such that $h(1) \leq n$. Then we define, for all integers $i \geq 0$ and d, the numbers $\beta_{i+1,i+d}(h, n)$ as in Proposition 3.8 if $d \geq 2$ and otherwise:

$$\beta_{i+1,i+d}(h, n) := \begin{cases} \binom{n-h(1)}{i+1} & \text{if } d = 1, \\ 0 & \text{if } d \leq 0. \end{cases}$$

Moreover, if $i \leq 0$ we set

$$\beta_{i,j}(h, n) := \begin{cases} 1 & \text{if } (i, j) = (0, 0), \\ 0 & \text{otherwise.} \end{cases}$$

Lemma 3.11 *Let $A = R/I \neq 0$ be any graded K-algebra. Then we have for all integers i, d with $d \leq 1$,*

$$\beta^R_{i+1,i+d}(A) = \beta_{i+1,i+d}(h_A, n).$$

Proof Since A has as an R-module just one generator in degree zero, this is clear if $d \leq 0$. Furthermore, $I_{\langle 1\rangle}$ is generated by a regular sequence of length $n - h_A(1)$. Its minimal free resolution is given by the Koszul complex. Hence, the claim follows for $d = 1$ because $\beta^R_{i+1,i+1}(A) = \beta^R_{i,i+1}(I_{\langle 1\rangle})$. □

Combined with results of Bigatti, Hullet, and Pardue, we get the main result of this section: bounds for the graded Betti numbers of a K-algebra as an R-module in terms of its Hilbert function and the dimension of R.

Theorem 3.12 *Let $A = R/I \neq 0$ be a graded K-algebra. Then its graded Betti numbers are bounded by*

$$\beta^R_{i+1,i+j}(A) \leq \beta_{i+1,i+j}(h_A, n) \qquad (i, j \in \mathbb{Z}).$$

Furthermore, equality is attained for all integers i, j if I is a lex-segment ideal.

Proof Let $I^{\text{lex}} \subset R$ be the lex-segment ideal such that A and R/I^{lex} have the same Hilbert function. Then we have for all integers i, j that

$$\beta_{i+1,i+j}^R(A) \leq \beta_{i+1,i+j}^R\big(R/I^{\text{lex}}\big)$$

according to Bigatti [1] and Hulett [8] if $\operatorname{char} K = 0$ and to Pardue [17] if K has positive characteristic. Since Proposition 3.8 and Lemma 3.11 yield

$$\beta_{i+1,i+j}^R\big(R/I^{\text{lex}}\big) = \beta_{i+1,i+j}(h_A, n) \quad (i, j \in \mathbb{Z}),$$

our claims follow. □

Remark 3.13 Note that Theorem 3.12 gives in particular that $\beta_{i+1,i+d}^R(A) = 0$ if $i \geq n$, in accordance with Hilbert's Syzygy Theorem.

We conclude this section by discussing the graded Betti numbers of Cohen–Macaulay algebras with the so-called Weak Lefschetz property. The special case of a Gorenstein algebra is crucial for the applications to polytopes in the following section.

Let $A = R/I$ be a graded Cohen–Macaulay K-algebra of Krull dimension d and let $l_1, \ldots, l_d \in [R]_1$ be sufficiently general linear forms. Then $\overline{A} := A/(l_1, \ldots, l_d)A$ is called the *Artinian reduction* of A. Its Hilbert function and graded Betti numbers as a module over $\overline{R} := R/(l_1, \ldots, l_d)R$ do not depend on the choice of the forms $l_1, \ldots, l_d$. The Hilbert function of $\overline{A}$ takes positive values in only finitely many degrees. The sequence of these positive integers $\underline{h} = (h_0, h_1, \ldots, h_r)$ is called the *h-vector* of A. We set $\beta_{i+1,i+d}(\underline{h}, n-d) := \beta_{i+1,i+d}(h_{\overline{A}}, n-d)$. Using this notation we get:

Corollary 3.14 *Let $A = R/I$ be a Cohen–Macaulay graded K-algebra of dimension d with h-vector $\underline{h}$. Then its graded Betti numbers satisfy*

$$\beta_{i+1,i+j}^R(A) \leq \beta_{i+1,i+j}(\underline{h}, n-d) \quad (i, j \in \mathbb{Z}).$$

Proof If $l \in [R]_1$ is a not a zero-divisor of A, then the graded Betti numbers of A as an R-module agree with the graded Betti numbers of A/lA as an R/lR module (see, e.g., Corollary 8.5 of [16]). Hence, by passing to the Artinian reduction of A, Theorem 3.12 provides the claim. □

Remark 3.15 Note that, for any O-sequence $\underline{h} = (1, h_1, \ldots, h_r)$ with $h_r > 0$, Definition 3.10 provides $\beta_{i+1,i+j}(\underline{h}, m) = 0$ for all $i, m \geq 0$ if $j \leq 1$ or $j \geq r+2$.

Recall that an Artinian graded K-algebra A has the so-called Weak Lefschetz property if there is an element $l \in A$ of degree one such that, for each $j \in \mathbb{Z}$, the multiplication $\times l\colon [A]_{j-1} \to [A]_j$ has maximal rank. The Cohen–Macaulay K-algebra A is said to have the *Weak Lefschetz property* if its Artinian reduction has the Weak Lefschetz property.

Remark 3.16 The Hilbert functions of Cohen–Macaulay algebras with the Weak Lefschetz property have been completely classified in Proposition 3.5 of [6]. Moreover, Theorem 3.20 in [6] gives optimal upper bounds on their graded Betti numbers in terms of the Betti numbers of certain lex-segment ideals. Thus, combining this result with Theorem 3.12, one gets upper bounds for the Betti numbers of these algebras in terms of their Hilbert functions. In general, these bounds are strictly smaller than the bounds of Corollary 3.14 for Cohen–Macaulay algebras that do not necessarily have the Weak Lefschetz property.

The h-vectors of graded Gorenstein algebras with the Weak Lefschetz property are precisely the SI-sequences (see Theorem 6.3 of [16] or Theorem 1.2 of [5]). For their Betti numbers we obtain:

Theorem 3.17 *Let $\underline{h} = (1, h_1, \ldots, h_u, \ldots, h_r)$ be an SI-sequence where $h_{u-1} < h_u = \cdots = h_{r-u} > h_{r-u+1}$. Put $\underline{g} = (1, h_1 - 1, h_2 - h_1, \ldots, h_u - h_{u-1})$. If $A = R/I$ is a Gorenstein graded K-algebra of dimension d with the Weak Lefschetz property and h-vector $\underline{h}$, then its graded Betti numbers satisfy*

$$\beta^R_{i+1,i+j}(A) \leq \begin{cases} \beta_{i+1,i+j}(\underline{g}, m) & \text{if } j \leq r-u, \\ \beta_{i+1,i+j}(\underline{g}, m) + \beta_{g_1-i,r+h_1-i-j}(\underline{g}, m) & \text{if } r-u+1 \leq j \leq u+1, \\ \beta_{g_1-i,r+h_1-i-j}(\underline{g}, m) & \text{if } j \geq u+2, \end{cases}$$

where $m := n - d - 1 = \dim R - d - 1$.

Proof This follows immediately by combining Theorem 8.13 of [16] and Theorem 3.12. □

4 Explicit Bounds for the Number of Missing Simplices

We now return to the consideration of simplicial polytopes. To this end we specialize the results of Sect. 3 and then discuss some applications.

We begin by simplifying our notation somewhat. Let P be a simplicial d-polytope with f-vector $\underline{f}$. It is well known that the h-vector of the Stanley–Reisner ring $K[P]$ agrees with the h-vector of P as defined in the Introduction. Furthermore, in Sect. 2 we defined the graded Betti numbers of $K[P] = R/I_P$ by resolving $K[P]$ as an R-module where R is a polynomial ring of dimension f_0 over K, i.e.,

$$\beta^K_{i,j}(P) = \beta^R_{i,j}\big(K[P]\big).$$

Note that the Stanley–Reisner ideal I_P does not contain any linear forms. The graded Betti numbers of P agree with the graded Betti numbers of the Artinian reduction of $K[P]$ as a module over a polynomial ring of dimension $f_0 - d = h_1$. Thus, we can simplify the statements of the bounds of $\beta^K_{i,j}(P)$ by setting:

Notation 4.1 Using the notation introduced above Corollary 3.14 we define for every O-sequence $\underline{h}$,

$$\beta_{i+1,i+j}(\underline{h}) := \beta_{i+1,i+j}(\underline{h}, h_1).$$

Notice that $\beta_{i+1,i+j}(\underline{h}) = 0$ if $i \geq 0$ and $j \leq 1$.

In this section we primarily use the g-vector of a polytope which is defined as follows:

Definition 4.2 Let P be a simplicial polytope with h-vector $\underline{h} := (h_0, \ldots, h_d)$. Then the g-Theorem (Theorem 1.1) shows that there is a unique integer u such that $h_{u-1} < h_u = \cdots = h_{d-u} > h_{d-u+1}$. The vector $\underline{g} = (g_0, \ldots, g_u) := (1, h_1 - 1, h_2 - h_1, \ldots, h_u - h_{u-1})$ is called the *g-vector* of P. All its entries are positive.

Some observations are in order.

Remark 4.3 (i) By its definition, the g-vector of the polytope P is uniquely determined by the h-vector of P. The g-Theorem shows that the h-vector of P (thus also its f-vector) can be recovered from its g-vector, provided the dimension of P is given.

(ii) The g-Theorem also gives an estimate of the length of the g-vector because it implies $2u \leq d = \dim P$.

Now we can state our explicit bounds for the Betti numbers of a polytope.

Theorem 4.4 *Let K be a field of characteristic zero and let $\underline{g} = (g_0, \ldots, g_u)$ be an O-sequence with $g_u > 0$. Then we have*:

(a) *If P is a simplicial d-polytope with g-vector $\underline{g}$, then*

$$\beta^K_{i+1,i+j}(P) \leq \begin{cases} \beta_{i+1,i+j}(\underline{g}) & \text{if } j \leq d-u, \\ \beta_{i+1,i+j}(\underline{g}) + \beta_{g_1-i,d+h_1-i-j}(\underline{g}) & \text{if } d-u+1 \leq j \leq u+1, \\ \beta_{g_1-i,d+g_1+1-i-j}(\underline{g}) & \text{if } j \geq u+2. \end{cases}$$

(b) *In* (a) *equality is attained for all integers i, j if P is the d-dimensional Billera–Lee polytope with g-vector $\underline{g}$.*

Proof It is well known that the Stanley–Reisner ring of every simplicial polytope is a Gorenstein algebra. Furthermore, according to Stanley [20] (see also [15]), it has the Weak Lefschetz property. Hence part (a) is a consequence of Theorem 3.17. Part (b) follows from Theorem 9.6 of [16] and Theorem 3.12, as pointed out in the proof of Theorem 2.1. □

We have seen in Sect. 2 that the number of empty j-simplices of the simplicial polytope P is equal to the Betti number $\beta^K_{1,j+1}(P)$. Thus, we want to make the preceding bounds more explicit if $i = 0$. At first, we treat a trivial case.

Remark 4.5 Notice that the g-vector has length one, i.e., $u = 0$ if and only if the polytope P is a simplex. In this case, its Stanley–Reisner ideal is a principal ideal generated by a monomial of degree $d = \dim P$.

In the following result we stress when the Betti numbers vanish. Because of Remark 4.5, it is harmless to assume that $u \geq 1$. We use Notation 3.1.

Corollary 4.6 *Let $\underline{g} = (g_0, \ldots, g_u)$ be an O-sequence with $g_u > 0$ and $u \geq 1$. Set $g_{u+1} := 0$. Then we have*:

(a) *If P is a simplicial d-polytope with g-vector $\underline{g}$, then there are the following bounds*:
(i) *If $d \geq 2u+1$, then*

$$\beta_{1,j}^K(P) \leq \begin{cases} g_{j-1}^{\langle j-1\rangle} - g_j & \text{if } 2 \leq j \leq u+1, \\ g_{d+1-j} - (g_{d+2-j})_{[d+2-j]} & \text{if } d-u+1 \leq j \leq d, \\ 0 & \text{otherwise.} \end{cases}$$

(ii) *If $d = 2u$, then*

$$\beta_{1,j}^K(P) \leq \begin{cases} g_{j-1}^{\langle j-1\rangle} - g_j & \text{if } 2 \leq j \leq u, \\ g_u^{\langle u\rangle} + g_u & \text{if } j = u+1, \\ g_{d+1-j} - (g_{d+2-j})_{[d+2-j]} & \text{if } u+2 \leq j \leq d, \\ 0 & \text{otherwise.} \end{cases}$$

(b) *In* (a) *equality is attained for all integers j if P is the d-dimensional Billera–Lee polytope with g-vector $\underline{g}$.*

Proof Since the first Betti numbers of any polytope do not depend on the characteristic of the field, the claims follow from Theorem 4.4 by taking into account Corollary 2.4, Corollary 3.9, and the fact that $\beta_{i+1,i+j}(\underline{g}) = 0$ if $i \geq 0$ and either $j \leq 1$ or $j \geq u+2$ due to Remark 3.15. □

To illustrate the last result, we consider an easy case.

Example 4.7 Let P be a simplicial d-polytope with $g_1 = 1$. Then its Stanley–Reisner ideal I_P is a Gorenstein ideal of height two, thus a complete intersection. Indeed, since the g-vector of P is an O-sequence, it must be $\underline{g} = (g_0, \ldots, g_u) = (1, \ldots, 1)$. Hence Corollary 4.6 provides that I_P has exactly two minimal generators, one of degree $u+1$ and one of degree $d-u+1$. Equivalently, P has exactly two empty simplices, one of dimension u and one of dimension $d-u$.

As an immediate consequence of Corollary 4.6 we partially recover Proposition 3.6 of [10].

Corollary 4.8 *Every simplicial d-polytope has no empty faces of dimension j if $u+1 \leq j \leq d-u-1$.*

Remark 4.9 Kalai's Conjecture 8 in [10] states that the following converse of Corollary 4.8 should be true: If there is an integer k such that $d \geq 2k$ and the simplicial d-polytope has no empty simplices of dimension j whenever $k \leq j \leq d-k$, then $u < k$. Kalai has proved this if $k = 2$ in [9]. Our results provide the following weaker version of Kalai's conjecture:

If there is an integer k such that $d \geq 2k$ and *every* simplicial d-polytope with g-vector $(g_0, \ldots, g_u)$ has no empty simplices of dimension j whenever $k \leq j \leq d-k$, then $u < k$.

Indeed, this follows by the sharpness of the bounds in Corollary 4.6.

Now we want to make some existence results of Kalai and Perles effective. As preparation, we state:

Corollary 4.10 *Let P be a simplicial d-polytope with g-vector $\underline{g} = (g_0, \ldots, g_u)$ where $u \geq 1$. Set $g_{u+1} = 0$. Then the number $N(k)$ of empty simplices of P whose dimension is at most k is bounded above as follows*:

$$N(k) \leq \begin{cases} g_1 + \sum_{j=1}^{k} \{g_j^{\langle j \rangle} - g_j\} - g_{k+1} & \text{if } 1 \leq k \leq \min\{u, d-u-1\}, \\ N(u) & \text{if } u < k < d-u, \\ g_1 + g_{d-k}^{\langle d-k \rangle} + \sum_{j=1}^{d-k-1} \{g_j^{\langle j \rangle} - g_j\} & \\ \quad + \sum_{j=d-k+1}^{u} \{g_j^{\langle j \rangle} - (g_j)_{[j]}\} & \text{if } d-u \leq k < d. \end{cases}$$

Furthermore, for each k, the bound is attained if P is the Billera–Lee d-polytope with g-vector $\underline{g}$.

Proof By Corollary 4.8, this is clear if $u < k < d-u$. In any case, we know that $N(k) = \sum_{j=2}^{k+1} \beta_{1,j}^K(P)$. Thus, using Corollary 4.6 carefully, elementary calculations provide the claim. We omit the details. □

The last result immediately gives:

Corollary 4.11 *If P is a simplicial polytope with g-vector $\underline{g} = (g_0, \ldots, g_u)$, where $u \geq 1$, then its total number of empty simplices is at most*

$$\binom{g_1+2}{2} - 1 + \sum_{j=2}^{u} \{g_j^{\langle j \rangle} - (g_j)_{[j]}\}.$$

Furthermore, this bound is attained if P is any Billera–Lee polytope with g-vector $\underline{g}$.

Proof Use Corollary 4.10 with $k = d-1$ and recall that $g_1^{\langle 1 \rangle} = \binom{g_1+1}{2}$. □

Remark 4.12 It is somewhat surprising that the bound in Corollary 4.11 does *not* depend on the dimension of the polytope. In contrast, the other bounds (see, e.g., Corollary 4.10) do depend on the dimension d of the polytope.

In view of Corollary 4.10, the following elementary facts will be useful.

Lemma 4.13 *Let k be a positive integer. If $a \geq b$ are non-negative integers, then*

(a) $a^{\langle k \rangle} - a_{[k]} \geq b^{\langle k \rangle} - b_{[k]}$;
(b) $a^{\langle k \rangle} - a \geq b^{\langle k \rangle} - b$;
(c) $a_{[k]} \geq b_{[k]}$.

Proof We show only (a). The proofs of the other claims are similar and easier.

To see (a), we begin by noting, for integers $m \geq j > 0$, the identity

$$\binom{m+1}{j+1} - \binom{m-1}{j-1} = \binom{m}{j+1} + \binom{m-1}{j}. \tag{4.1}$$

Now we use induction on $k \geq 1$. Since $a^{\langle 1 \rangle} - a_{[1]} = \binom{a+1}{2} - 1$, the claim is clear if $k = 1$. Let $k \geq 2$. Consider the k-binomial expansions

$$a =: \binom{m_k}{k} + \binom{m_{k-1}}{k-1} + \cdots + \binom{m_s}{s}$$

and

$$b =: \binom{n_k}{k} + \binom{n_{k-1}}{k-1} + \cdots + \binom{n_t}{t}.$$

Since $a \geq b$, we get $m_k \geq n_k$. We distinguish two cases.

Case 1. Let $m_k = n_k$. Then the claim follows by applying the induction hypothesis to

$$a - \binom{m_k}{k} \geq b - \binom{m_k}{k}.$$

Case 2. Let $m_k > n_k$. Using $n_i \leq n_k - k + i$ and formula (4.1), we get

$$\begin{aligned} b^{\langle k \rangle} - b_{[k]} &= \sum_{i=t}^{k} \left\{ \binom{n_i}{i+1} + \binom{n_i - 1}{i} \right\} \\ &\leq \sum_{i=1}^{k} \left\{ \binom{n_k - k + i}{i+1} + \binom{n_k - k - 1 + i}{i} \right\} \\ &= \binom{n_k + 1}{k+1} + \binom{n_k}{k} - (n_k - k + 2) \\ &< \binom{m_k}{k+1} + \binom{m_k - 1}{k} \end{aligned}$$

because $n_k < m_k$. The claim follows since formula (4.1) gives $\binom{m_k}{k+1} + \binom{m_k - 1}{k} \leq a^{\langle k \rangle} - a_{[k]}$. □

Remark 4.14 In general, it is not true that $a > b$ implies $a^{\langle k \rangle} - a_{[k]} > b^{\langle k \rangle} - b_{[k]}$. For example, if $k \geq 2$ and $a - 1 = b = \binom{m}{k} > 0$, then $a^{\langle k \rangle} - a_{[k]} = b^{\langle k \rangle} - b_{[k]}$.

We are ready to establish optimal bounds that depend only on the dimension and the number of vertices.

Theorem 4.15 *Let P be a simplicial d-polytope with $d+g_1+1$ vertices which is not a simplex. Then there is the following bound on the number $N(k)$ of empty simplices of P whose dimension is $\leq k$:*

$$N(k) \leq \begin{cases} \binom{g_1+k}{g_1-1} & \text{if } 1 \leq k < d/2; \\ \binom{g_1+\lfloor d/2 \rfloor}{g_1-1} + \binom{g_1+\lfloor d/2 \rfloor -1}{g_1-1} & \text{if } d/2 \leq k < d. \end{cases}$$

Furthermore, for each k, the bound is attained if P is the Billera–Lee d-polytope with g-vector $(g_0, \ldots, g_u)$ where $g_j = \binom{g_1+j-1}{j}$, $0 \leq j \leq u$, and $u = \min\{k, \lfloor d/2 \rfloor\}$.

Proof Let $\underline{g} = (g_0, \ldots, g_u)$ be the g-vector of P. Since P is not a simplex, we have $u \geq 1$. We have to distinguish two cases.

Case 1. Let $k < d/2$. If $k > u$, then we formally set $g_{u+1} = \cdots = g_{\lfloor d/2 \rfloor} = 0$. Since $k < d/2 \leq d-u$, Corollary 4.10 provides

$$N(k) \leq g_1 + \sum_{j=1}^{k} \left\{ g_j^{\langle j \rangle} - g_j \right\} - g_{k+1}.$$

According to Lemma 4.13, the sum on the right-hand side becomes maximal if $g_2, \ldots, g_k$ are as large as possible and $g_{k+1} = 0$. The latter means $u = k$. Macaulay's Theorem 3.2 implies $g_j \leq \binom{g_1+j-1}{j}$. Now an easy computation provides the bound in this case. It is sharp because $(g_0, \ldots, g_k)$, where $g_j = \binom{g_1+j-1}{j}$, is a g-vector of a simplicial d-polytope by the g-Theorem, thus Corollary 4.10 applies.

Case 2. Let $d/2 \leq k < d$. First, we also assume that $k \geq d-u$. Then Corollary 4.10 gives

$$N(k) \leq g_1 + g_{d-k}^{\langle d-k \rangle} + \sum_{j=1}^{d-k-1} \left\{ g_j^{\langle j \rangle} - g_j \right\} + \sum_{j=d-k+1}^{u} \left\{ g_j^{\langle j \rangle} - (g_j)_{\lfloor j \rfloor} \right\}.$$

Again, Lemma 4.13 shows that, for fixed u, the bound is maximized if $g_j = \binom{g_1+j-1}{j}$, $0 \leq j \leq u$. This provides

$$N(k) \leq \binom{g_1+u}{g_1-1} + \binom{g_1+u-1}{g_1-1}.$$

Since $u \leq d/2$, our bound follows in this case.

Second, assume $k < d-u$. Then $u \leq d/2 \leq k < d-u$ yields $u < d/2$. Thus Corollary 4.10 provides $N(k) = N(u)$, but $N(u) \leq \binom{g_1+u}{g_1-1}$ by Case 1. This concludes the proof of the bound in Case 2. Its sharpness is shown as in Case 1. □

As an immediate consequence we obtain:

Corollary 4.16 *Every simplicial polytope, which is not a simplex, has at most* $\binom{g_1+k}{g_1-1} + \binom{g_1+k-1}{g_1-1}$ *empty simplices of dimension* $\leq k$.

Remark 4.17 Kalai [10, Theorem 2.7] has first given an estimate as in Corollary 4.16. His bound is

$$N(k) \leq (g_1+1)^{k+1} \cdot (k+1)!.$$

Comparing with our bound, we see that Kalai's bound is asymptotically not optimal for $g_1 \gg 0$.

Notice that the bound on $N(k)$ in Theorem 4.15 does not depend on k if $k \geq d/2$. This becomes plausible by considering cyclic polytopes.

Example 4.18 (i) Recall that a *cyclic polytope* $C(f_0, d)$ is a d-dimensional simplicial polytope which is the convex hull of f_0 distinct points on the moment curve

$$\left\{\left(t, t^2, \ldots, t^d\right) \mid t \in \mathbb{R}\right\}.$$

Its combinatorial type depends only on f_0 and d.

According to McMullen's Upper Bound Theorem [13], the cyclic polytope $C(f_0, d)$ has the maximal f-vector among all simplicial d-polytopes with f_0 vertices. Theorem 4.15 shows that it also has the maximal total number of empty simplices among these polytopes. Indeed, this follows by comparing with the main result in [22] (see also Corollary 9.10 of [16]) which provides that $C(f_0, d)$ has $\binom{g_1+\lfloor d/2 \rfloor}{g_1-1} + \binom{g_1+\lfloor d/2 \rfloor - 1}{g_1-1}$ empty simplices. Moreover, the empty simplices of $C(f_0, d)$ have either dimension $d/2$ if d is even or dimensions $(d-1)/2$ and $(d+1)/2$ if d is odd. This explains why the bound on $N(k)$ in Theorem 4.15 does not change if $k \geq d/2$.

(ii) If P is a simplicial d-polytope with $f_0 \geq d+2$ vertices, then Theorem 4.15 gives for its number of empty edges

$$N(1) \leq \begin{cases} f_0(f_0-3)/2 & \text{if } d = 2, \\ \binom{f_0-d}{2} & \text{if } d \geq 3. \end{cases}$$

If $d = 2$, the bound is always attained because $f_0(f_0-3)/2$ is the number of "missing diagonals" of a convex f_0-gon. The results in [24] (see also Remark 9.9 of [16]) provide that the bound is sharp for stacked d-polytopes for all $d \geq 2$.

Remark 4.19 Recall that the *k-skeleton* of an arbitrary d-polytope P is the set of all faces of P whose dimension is at most k. Perles [18] has shown:

The number of combinatorial types of k-skeleta of d-polytopes with $d + g_1 + 1$ vertices is bounded by a function in k and g_1.

In [10] Kalai gave a new proof of this result that relies on the concept of missing faces. Indeed, in the simplicial case one concludes by using a bound on $N(k)$ because the k-skeleton of a simplicial polytope is determined by its set of empty simplices of dimension $\leq k$.

In [10] Kalai sketches an argument showing that the number of empty simplices can be bounded with very little information on the g-vector. Below, we slightly correct Theorem 3.8 of [10] and give explicit bounds. We use Notation 3.1.

Theorem 4.20 *Fix integers $j \geq k \geq 1$ and $b \geq 0$. Let P be a simplicial d-polytope P with $g_k \leq b$ where we define $g_i = 0$ if $i > u$. If $d \geq j + k$, then the number of empty j-simplices of P is bounded by*

$$\begin{cases} b^{\langle k, j-k+1\rangle} & \text{if } j < d/2, \\ b^{\langle k, j-k+1\rangle} + b^{\langle k, j-k\rangle} & \text{if } j = d/2, \\ b^{\langle k, d-j-k\rangle} & \text{if } j > d/2. \end{cases}$$

Proof We have to bound $\beta^K_{1,j+1}(K[P])$. By Corollary 4.8, P has no empty j-simplices if $u + 1 \leq j \leq d - u - 1$. Thus, we may assume that $1 \leq j \leq u$ or $d - u \leq j \leq d - 1$.

Case 1. Assume $1 \leq j \leq u \leq d/2$. Then Corollary 4.6 provides if $j < d/2$,

$$\beta^K_{1,j+1}\big(K[P]\big) \leq g_j^{\langle j\rangle} - g_{j+1}.$$

Using Lemma 4.13, we see that the bound is maximized if $g_{j+1} = 0$ and g_j is as large as possible. Since the g-vector is an O-sequence, we get $g_j \leq g_k^{\langle k, j-k\rangle} \leq b^{\langle k, j-k\rangle}$. Our claimed bound follows.

If $j = d/2$, then we get $j = u = d/2$. Hence Corollary 4.6 gives

$$\beta^K_{1,j+1}\big(K[P]\big) \leq g_j^{\langle j\rangle} + g_j.$$

Now the bound is shown as above.

Case 2. Assume $d/2 \leq d - u \leq j \leq d - 1$. By the above considerations, we may also assume that $j \neq d/2$. Thus, Corollary 4.6 provides

$$\beta^K_{1,j+1}\big(K[P]\big) \leq g_{d-j} - (g_{d+1-j})_{[d+1-j]}.$$

Using our assumption $d - j \geq k$, we conclude as above. □

Remark 4.21 (i) Theorem 3.8 of [10] the existence of bounds as in the above result is claimed without assuming $d \geq j + k$. However, this is impossible, as Case 2 in the above proof shows. Indeed, if $d - j < k$ and $d > j > d/2$, then knowledge of g_k does not give any information on g_{d-j}. In particular, g_{d-j} can be arbitrarily large preventing the existence of a bound on $\beta^K_{1,j+1}(K[P])$ in terms of g_k, j, k in this case.

For a somewhat specific example, fix $k = j = 2$ and $d = 3$. Then the Billera–Lee 3-polytope with g-vector $(1, g_1)$ has g_1 empty 2-simplices.

(ii) Note that the bounds in Theorem 4.20 are sharp if $g_k = b$. This follows from the proof.

If we only know that d is large enough compared with j and k, then we have the following weaker bound.

Corollary 4.22 *Fix integers $j \geq k \geq 1$, $b \geq 0$, and $d \geq j + k$. Then the number of empty j-simplices of every simplicial d-polytope with $g_k \leq b$ is at most $b^{\langle k, j-k+1\rangle} + b^{\langle k, j-k\rangle}$.*

Proof By Theorem 4.20, it remains to consider the case where $j > d/2$. However, then $d - j < j$, thus $b^{\langle k, d-j-k\rangle} \leq b^{\langle k, j-k\rangle}$, and we conclude again by using Theorem 4.20. □

Remark 4.23 Notice that the bound in Corollary 4.22 is independent of the number of vertices of the polytope and its dimension, provided the latter is large enough.

In essence, all the bounds on the number of empty simplices are bounds on certain first graded Betti numbers of the Stanley–Reisner ring of a simplicial polytope. As such, using Theorem 3.17, they can be extended to bounds for the first graded Betti numbers of any graded Gorenstein algebra with the Weak Lefschetz property. We leave this and analogous considerations for higher Betti numbers to the interested reader.

We conclude this note by pointing out some directions for future research:

Remark 4.24 (i) It is an open problem whether the upper bounds on the number of empty simplices of simplicial polytopes obtained in this paper extend to the case of empty pyramids of arbitrary polytopes. Recall that an empty pyramid of a polytope P is a subcomplex of the face complex of P that consists of all the proper faces of a pyramid over a face of P.

More generally, it would be very interesting to investigate whether prescribing the (toric) g-vector (see Sect. 3.14 of [21]) bounds the number of empty simplices (or possibly even empty pyramids) of non-simplicial polytopes.

(ii) It is natural to wonder also about good lower bounds on the number of empty simplices for polytopes with a given f-vector. This problem seems difficult. For simplicial d-polytopes with f_0 vertices, Krull's Principal Ideal Theorem implies that the total number of empty simplices is at least $f_0 - d$. Equality is true if $f_0 - d \leq 2$, but in most other cases this bound seems far from being optimal.

Acknowledgements The author thanks Gil Kalai, Carl Lee, and Juan Migliore for motivating discussions, encouragement, and helpful comments. He also thanks the referees whose suggestions helped to improve the presentation.

References

1. Bigatti, A.: Upper bounds for the Betti numbers of a given Hilbert function. Commun. Algebra **21**, 2317–2334 (1993)
2. Billera, L.J., Lee, C.W.: A proof of the sufficiency of McMullen's conditions for f-vectors of simplicial convex polytopes. J. Comb. Theory Ser. A **31**, 237–255 (1981)
3. Bruns, W., Herzog, J.: Cohen–Macaulay Rings, rev. edn. Cambridge Studies in Advanced Mathematics, vol. 39. Cambridge University Press, Cambridge (1998)
4. Gotzmann, G.: Eine Bedingung für die Flachheit und das Hilbertpolynom eines graduierten Ringes. Math. Z. **158**, 61–70 (1978)
5. Harima, T.: Characterization of Hilbert functions of Gorenstein–Artin algebras with the Weak Stanley property. Proc. Am. Math. Soc. **123**, 3631–3638 (1995)

6. Harima, T., Migliore, J., Nagel, U., Watanabe, J.: The Weak and Strong Lefschetz properties for Artinian K-algebras. J. Algebra **262**, 99–126 (2003)
7. Herzog, J., Hibi, T.: Componentwise linear ideals. Nagoya Math. J. **153**, 141–153 (1999)
8. Hulett, H.: Maximum Betti numbers of homogeneous ideals with a given Hilbert function. Commun. Algebra **21**, 2335–2350 (1993)
9. Kalai, G.: Rigidity and the lower bound theorem. Invent. Math. **88**, 125–151 (1987)
10. Kalai, G.: Some aspects of the combinatorial theory of convex polytopes. In: Polytopes: Abstract, Convex and Computational, Scarborough, ON, 1993. NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci., vol. 440, pp. 205–229. Kluwer, Dordrecht (1994)
11. Kalai, G.: Polytope skeletons and paths. In: Goodman, O'Rourke (eds.) Handbook of Discrete and Computational Geometry. CRC Ser. Discrete Math. Appl., pp. 331–353. CRC, Boca Raton (1997)
12. Macaulay, F.S.: Some properties of enumeration in the theory of modular systems. Proc. Lond. Math. Soc. **26**, 531–555 (1927)
13. McMullen, P.: The maximum number of faces of a convex polytope. Mathematika **17**, 179–184 (1970)
14. McMullen, P.: The number of faces of simplicial polytopes. Israel J. Math. **9**, 559–570 (1971)
15. McMullen, P.: On simple polytopes. Invent. Math. **113**, 419–444 (1993)
16. Migliore, J., Nagel, U.: Reduced arithmetically Gorenstein schemes and simplicial polytopes with maximal Betti numbers. Adv. Math. **180**, 1–63 (2003)
17. Pardue, K.: Deformation classes of graded modules and maximal Betti numbers. Ill. J. Math. **40**, 564–585 (1996)
18. Perles, M.: Truncation of atomic lattices, unpublished manuscript (around 1970)
19. Stanley, R.: Hilbert functions of graded algebras. Adv. Math. **28**, 57–82 (1978)
20. Stanley, R.: The number of faces of a simplicial convex polytope. Adv. Math. **35**, 236–238 (1980)
21. Stanley, R.: Enumerative Combinatorics, vol. 1. Cambridge Studies in Advanced Mathematics, vol. 49. Cambridge University Press, Cambridge (1997)
22. Terai, N., Hibi, T.: Computation of Betti numbers of monomial ideals associated with cyclic polytopes. Discrete Comput. Geom. **15**, 287–295 (1996)
23. Terai, N., Hibi, T.: Some results on Betti numbers of Stanley–Reisner rings. Discrete Math. **157**, 311–320 (1996)
24. Terai, N., Hibi, T.: Computation of Betti numbers of monomial ideals associated with stacked polytopes. Manuscripta Math. **92**, 447–453 (1997)

Rigidity and the Lower Bound Theorem for Doubly Cohen–Macaulay Complexes

Eran Nevo

Abstract We prove that for $d \geq 3$, the 1-skeleton of any $(d-1)$-dimensional doubly Cohen–Macaulay (abbreviated 2-CM) complex is generically d-rigid. This implies that Barnette's lower bound inequalities for boundary complexes of simplicial polytopes (Barnette, D. Isr. J. Math. 10:121–125, 1971; Barnette, D. Pac. J. Math. 46:349–354, 1973) hold for every 2-CM complex of dimension ≥ 2 (see Kalai, G. Invent. Math. 88:125–151, 1987). Moreover, the initial part (g_0, g_1, g_2) of the g-vector of a 2-CM complex (of dimension ≥ 3) is an M-sequence. It was conjectured by Björner and Swartz (J. Comb. Theory Ser. A 113:1305–1320, 2006) that the entire g-vector of a 2-CM complex is an M-sequence.

1 Introduction

The g-theorem gives a complete characterization of the f-vectors of boundary complexes of simplicial polytopes. It was conjectured by McMullen in 1970 and proved by Billera and Lee [5] (sufficiency) and by Stanley [13] (necessity) in 1980. A major open problem in f-vector theory is the g-conjecture, which asserts that this characterization holds for all homology spheres. The open part of this conjecture is to show that the g-vector of every homology sphere is an M-sequence, i.e. it is the f-vector of some order ideal of monomials. Based on the fact that homology spheres are doubly Cohen–Macaulay (abbreviated 2-CM) and that the g-vector of some other classes of 2-CM complexes is known to be an M-sequence (e.g. [14]), Björner and Swartz [14] recently suspected that

Conjecture 1.1 ([14], *a weakening of Problem* 4.2.) *The g-vector of any 2-CM complex is an M-sequence.*

E. Nevo (✉)
Institute of Mathematics, The Hebrew University, Jerusalem, Israel
e-mail: eranevo@math.huji.ac.il

We prove a first step in this direction, namely:

Theorem 1.2 *Let K be a $(d-1)$-dimensional 2-CM simplicial complex (over some field) where $d \geq 4$. Then $(g_0(K), g_1(K), g_2(K))$ is an M-sequence.*

This theorem follows from the following theorem, combined with an interpretation of rigidity in terms of the face ring (Stanley–Reisner ring), due (implicitly) to Lee [10].

Theorem 1.3 *Let K be a $(d-1)$-dimensional 2-CM simplicial complex (over some field) where $d \geq 3$. Then K has a generically d-rigid 1-skeleton.*

Kalai [8] showed that if a simplicial complex K of dimension ≥ 2 satisfies the following conditions then it satisfies Barnette's lower bound inequalities:

(a) K has a generically $(\dim(K)+1)$-rigid 1-skeleton.
(b) For each face F of K of codimension > 2, its link $lk_K(F)$ has a generically $(\dim(lk_K(F))+1)$-rigid 1-skeleton.
(c) For each face F of K of codimension 2, its link $lk_K(F)$ (which is a graph) has at least as many edges as vertices.

Kalai used this observation to prove that Barnette's inequalities hold for a large class of simplicial complexes.

Observe that the link of a vertex in a 2-CM simplicial complex is 2-CM, and that a 2-CM graph is 2-connected. Combining it with Theorem 1.3 and the above result of Kalai we conclude:

Corollary 1.4 *Let K be a $(d-1)$-dimensional 2-CM simplicial complex where $d \geq 3$. For all $0 \leq i \leq d-1$ $f_i(K) \geq f_i(n,d)$ where $f_i(n,d)$ is the number of i-faces in a (equivalently every) stacked d-polytope on n vertices. (Explicitly, $f_{d-1}(n,d) = (d-1)n - (d+1)(d-2)$ and $f_i(n,d) = \binom{d}{i}n - \binom{d+1}{i+1}i$ for $1 \leq i \leq d-2$.)*

Theorem 1.3 is proved by decomposing K into a union of minimal $(d-1)$-cycle complexes (Fogelsanger's notion [6]). Each of these pieces has a generically d-rigid 1-skeleton ([6]), and the decomposition is such that gluing the pieces together results in a complex with a generically d-rigid 1-skeleton. The decomposition is detailed in Theorem 3.4.

This paper is organized as follows: In Sect. 2 we give the necessary background from rigidity theory, explain the connection between rigidity and the face ring, and reduce the results mentioned in the Introduction to Theorem 3.4. In Sect. 3 we give the necessary background on 2-CM complexes, prove Theorem 3.4 and discuss related problems and results.

2 Rigidity

The presentation of rigidity here is based mainly on the one in Kalai [8].

Let $G = (V, E)$ be a graph. A map $f : V \to \mathbb{R}^d$ is called a $d-$*embedding*. It is *rigid* if any small enough perturbation of it which preserves the lengths of the edges is induced by an isometry of $\mathbb{R}^d$. Formally, f is called *rigid* if there exists an $\varepsilon > 0$ such that if $g : V \to \mathbb{R}^d$ satisfies $d(f(v), g(v)) < \varepsilon$ for every $v \in V$ and $d(g(u), g(w)) = d(f(u), f(w))$ for every $\{u, w\} \in E$, then $d(g(u), g(w)) = d(f(u), f(w))$ for every $u, w \in V$ (where $d(a, b)$ denotes the Euclidean distance between the points a and b).

G is called *generically d-rigid* if the set of its rigid d-embeddings is open and dense in the topological vector space of all of its d-embeddings.

Let $V = [n]$, and let $\mathrm{Rig}(G, f)$ be the $dn \times |E|$ matrix which is defined as follows: for its column corresponding to $\{v < u\} \in E$ put the vector $f(v) - f(u)$ (resp. $f(u) - f(v)$) at the entries of the d rows corresponding to v (resp. u) and zero otherwise. G is generically d-rigid iff $\mathrm{Im}(\mathrm{Rig}(G, f)) = \mathrm{Im}(\mathrm{Rig}(K_V, f)$ for a generic f, where K_V is the complete graph on V. $\mathrm{Rig}(G, f)$ is called the *rigidity matrix* of G (its rank is independent of the generic f that we choose).

Let G be the 1-skeleton of a $(d-1)$-dimensional simplicial complex K. We define d generic degree-one elements in the polynomial ring $A = \mathbb{R}[x_1, \ldots, x_n]$ as follows: $\Theta_i = \sum_{v \in [n]} f(v)_i x_v$ where $f(v)_i$ is the projection of $f(v)$ on the i-th coordinate, $1 \le i \le d$. Then the sequence $\Theta = (\Theta_1, \ldots, \Theta_d)$ is a linear system of parameters for the face ring $\mathbb{R}[K] = A/I_K$ (I_K is the ideal in A generated by the monomials whose support is not an element of K). Let $H(K) = \mathbb{R}[K]/(\Theta) = H(K)_0 \oplus H(K)_1 \oplus \cdots$ where (Θ) is the ideal in A generated by the elements of Θ and the grading is induced by the degree grading in A. Consider the multiplication map $\omega : H(K)_1 \longrightarrow H(K)_2$, $m \to \omega m$ where $\omega = \sum_{v \in [n]} x_v$. Lee [10] proved that

$$\dim_{\mathbb{R}} \mathrm{Ker}(\mathrm{Rig}(G, f)) = \dim_{\mathbb{R}} H(K)_2 - \dim_{\mathbb{R}} \omega(H(K)_1). \tag{1}$$

Assume that G is generically d-rigid. Then $\dim_{\mathbb{R}} \mathrm{Ker}(\mathrm{Rig}(G, f)) = f_1(K) - \mathrm{rank}(\mathrm{Rig}(K_V, f)) = g_2(K) = \dim_{\mathbb{R}} H(K)_2 - \dim_{\mathbb{R}} H(K)_1$. Combining with (1), the map ω is injective, and hence $\dim_{\mathbb{R}}(H(K)/(\omega))_i = g_i(K)$ for $i = 2$; clearly this holds for $i = 0, 1$ as well. Hence $(g_o(K), g_1(K), g_2(K))$ is an M-sequence. We conclude that Theorem 1.3 implies Theorem 1.2, via the following algebraic result:

Theorem 2.1 *Let K be a $(d-1)$-dimensional 2-CM simplicial complex (over some field) where $d \ge 3$. Then the multiplication map $\omega : H(K)_1 \longrightarrow H(K)_2$ is injective.*

In order to prove Theorem 1.3, we need the concept of minimal cycle complexes, introduced by Fogelsanger [6]. We summarize his theory below.

Fix a field k (or more generally, any Abelian group) and consider the formal chain complex on a ground set $[n]$, $C = (\bigoplus\{kT : T \subseteq [n]\}, \partial)$, where $\partial(1T) = \sum_{t \in T} \mathrm{sign}(t, T)T \setminus \{t\}$ and $\mathrm{sign}(t, T) = (-1)^{|\{s \in T : s < t\}|}$. Define *subchain*, *minimal d-cycle* and *minimal d-cycle complex* as follows: $c' = \sum\{b_T T : T \subseteq [n], |T| = d + 1\}$ is a *subchain* of a d-chain $c = \sum\{a_T T : T \subseteq [n], |T| = d + 1\}$ iff for every such T, $b_T = a_T$ or $b_T = 0$. A d-chain c is a *d-cycle* if $\partial(c) = 0$, and is a *minimal d-cycle* if its only subchains which are cycles are c and 0. A simplicial complex K which is spanned by the support of a *minimal d-cycle* is called a *minimal d-cycle complex* (over k), i.e. $K = \{S : \exists T \ S \subseteq T, a_T \neq 0\}$ for some minimal d-cycle c as above. For example, triangulations of connected manifolds without boundary are minimal cycle complexes—fix $k = \mathbb{Z}_2$ and let the cycle be the sum of all facets.

The following is the main result in Fogelsanger's thesis.

Theorem 2.2 (*Fogelsanger* [6]) *For $d \geq 3$, every minimal $(d-1)$-cycle complex has a generically d-rigid 1-skeleton.*

We will need the following gluing lemma, due of Asimov and Roth, who introduced the concept of generic rigidity of graphs [1].

Theorem 2.3 (*Asimov and Roth* [2]) *Let G_1 and G_2 be generically d-rigid graphs. If $G_1 \cap G_2$ contains at least d vertices, then $G_1 \cup G_2$ is generically d-rigid.*

Now we are ready to conclude Theorem 1.3 from the decomposition theorem, Theorem 3.4.

Proof of Theorem 1.3 Consider a decomposition sequence of K as guaranteed by Theorem 3.4, $K = \bigcup_{i=1}^{m} S_i$. By Theorem 2.2 each S_i has a generically d-rigid 1-skeleton. By Theorem 2.3 for all $2 \leq i \leq m$ $\bigcup_{j=1}^{i} S_j$ has a generically d-rigid 1-skeleton, in particular K has a generically d-rigid 1-skeleton ($i = m$). □

Remark One can verify that Theorems 2.2 and 2.3, and hence also Theorem 1.3, continue to hold when replacing "generically d-rigid" by the notion "d-hypperconnected", introduced by Kalai [7]. Both of these assertions have an interpretation in terms of algebraic shifting, introduced by Kalai (see e.g. his survey [9]), namely: for both the exterior and symmetric shifting operators over the field $\mathbb{R}$, denoted by Δ, $\{d, n\} \in \Delta(K)$. The existence of this edge in the shifted complex implies the nonnegativity of $g_2(K)$.

3 Decomposing a 2-CM Complex

Definition 3.1 A simplicial complex K is $2-CM$ (over a fixed field k) if it is Cohen–Macaulay and for every vertex $v \in K$, $K - v$ is Cohen–Macaulay of the same dimension as K.

Here $K - v$ is the simplicial complex $\{T \in K : v \notin T\}$. By a theorem of Reisner [11], a simplicial complex L is Cohen–Macaulay iff it is pure and for every face $T \in L$ (including the empty set) and every $i < \dim(lk_L(T)$, $\tilde{H}_i(lk_L(T); k) = 0$ where $lk_L(T) = \{S \in L : T \cap S = \emptyset, T \cup S \in L\}$ and $\tilde{H}_i(M; k)$ is the reduced i-th homology of M over k. The proof of Theorem 3.4 is by induction on $\dim(K)$. Let us first consider the case where K is 1-dimensional.

A (simple finite) graph is 2-*connected* if after a deletion of any vertex from it, the remaining graph is connected and nontrivial (i.e. is not a single vertex nor empty). Note that a graph is 2-CM iff it is 2-connected.

Lemma 3.2 *A graph G is 2-connected iff there exists a decomposition $G = \bigcup_{i=1}^{m} C_i$ such that each C_i is a simple cycle and for every $1 < i \leq m$, $C_i \cap (\bigcup_{j<i} C_j)$ contains an edge.*

Moreover, for each $i_0 \in [m]$ the C_i's can be reordered by a permutation $\sigma : [m] \to [m]$ such that $\sigma^{-1}(1) = i_0$ and for every $i > 1$, $C_{\sigma^{-1}(i)} \cap (\bigcup_{j<i} C_{\sigma^{-1}(j)})$ contains an edge.

Proof Whitney [15] showed that a graph G is 2-connected iff it has an open ear decomposition, i.e. there exists a decomposition $G = \bigcup_{i=0}^{m} P_i$ such that each P_i is a simple open path, P_0 is an edge, $P_0 \cup P_1$ is a simple cycle and for every $1 < i \leq m$ $P_i \cap (\bigcup_{j<i} P_j)$ equals the 2 end vertices of P_i.

Assume that G is 2-connected and consider an open ear decomposition as above. Let $C_1 = P_0 \cup P_1$. For $i > 1$ choose a simple path $\tilde{P}_i$ in $\bigcup_{j<i} P_j$ that connects the 2 end vertices of P_i, and let $C_i = P_i \cup \tilde{P}_i$. $(C_1, \ldots, C_m)$ is the desired decomposition sequence of G.

Let C be the graph whose vertices are the C_i's and two of them are neighbors iff they have an edge in common. Thus, C is connected, and hence the 'Moreover' part of the Lemma is proved.

The other implication, that such a decomposition implies 2-connectivity, will not be used in the sequel, and its proof is omitted. □

For the induction step we need the following cone lemma. For v a vertex not in the support of a $(d-1)$-chain c, let $v * c$ denote the following d-chain: if $c = \sum\{a_T T : v \notin T \subseteq [n], |T| = d\}$ where $a_T \in k$ for all T, then $v * c = \sum\{\text{sign}(v, T) a_T T \cup \{v\} : v \notin T \subseteq [n], |T| = d\}$ where $\text{sign}(v, T) = (-1)^{|\{t \in T : t < v\}|}$.

Lemma 3.3 *Let s be a minimal $(d-1)$-cycle and let c be a minimal d-chain such that $\partial(c) = s$, i.e. c has no proper subchain c' such that $\partial(c') = s$. For v a vertex not in any face in* $\text{supp}(c)$*, the support of c, define $\tilde{s} = c - v * s$. Then $\tilde{s}$ is a minimal d-cycle.*

Proof $\partial(\tilde{s}) = \partial(c) - \partial(v * s) = s - (s - v * \partial(s)) = 0$ hence $\tilde{s}$ is a d-cycle. To show that it is minimal, let $\hat{s}$ be a subchain of $\tilde{s}$ such that $\partial(\hat{s}) = 0$. Note that $\text{supp}(c) \cap \text{supp}(v * s) = \emptyset$.

Case 1: v is contained in a face in $\text{supp}(\hat{s})$. By the minimality of s, $\text{supp}(v * s) \subseteq \text{supp}(\hat{s})$. Thus, by the minimality of c also $\text{supp}(c) \subseteq \text{supp}(\hat{s})$ and hence $\hat{s} = \tilde{s}$.

Case 2: v is not contained in any face in $\text{supp}(\hat{s})$. Thus, $\text{supp}(\hat{s}) \subseteq \text{supp}(c)$. As $\partial(\hat{s}) = 0$ then $\partial(c - \hat{s}) = s$. The minimality of c implies $\hat{s} = 0$. □

Theorem 3.4 *Let K be a d-dimensional 2-CM simplicial complex over a field k $(d \geq 1)$. Then there exists a decomposition $K = \bigcup_{i=1}^{m} S_i$ such that each S_i is a minimal d-cycle complex over k and for every $i > 1$, $S_i \cap (\bigcup_{j<i} S_j)$ contains a d-face.*

Moreover, for each $i_0 \in [m]$ the S_i's can be reordered by a permutation $\sigma : [m] \to [m]$ such that $\sigma^{-1}(1) = i_0$ and for every $i > 1$, $S_{\sigma^{-1}(i)} \cap (\bigcup_{j<i} S_{\sigma^{-1}(j)})$ contains a d-face.

Proof The proof is by induction on d. For $d = 1$, by Lemma 3.2 $K = \bigcup_{i=1}^{m(K)} C_i$ such that each C_i is a simple cycle and for every $i > 1$ $C_i \cap (\bigcup_{j<i} C_j)$ contains an edge. Define $s_i = \sum\{\text{sign}_e(i) e : e \in (C_i)_1\}$, then s_i is a minimal 1-cycle (orient the edges

properly: $\text{sign}_e(i)$ equals 1 or -1 accordingly) whose support spans the simplicial complex C_i. Moreover, by Lemma 3.2 each C_{i_0}, $i_0 \in [m(K)]$, can be chosen to be the first in such a decomposition sequence.

For $d > 1$, note that the link of every vertex in a 2-CM simplicial complex is 2-CM. For a vertex $v \in K$, as $lk_K(v)$ is 2-CM then by the induction hypothesis $lk_K(v) = \bigcup_{i=1}^{m(v)} C_i$ such that each C_i is a minimal $(d-1)$-cycle complex and for every $i > 1$ $C_i \cap (\bigcup_{j<i} C_j)$ contains a $(d-1)$-face. Let s_i be a minimal $(d-1)$-cycle whose support spans C_i. As $K - v$ is CM of dimension d, $\tilde{H}_{d-1}(K - v; k) = 0$. Hence there exists a d-chain c such that $\partial(c) = s_i$ and $\text{supp}(c) \subseteq K - v$.

Take c_i to be such a chain with a support of minimal cardinality. By Lemma 3.3, $\tilde{s}_i = c_i - v * s_i$ is a minimal d-cycle. Let $S_i(v)$ by the simplicial complex spanned by $\text{supp}(\tilde{s}_i)$; it is a minimal d-cycle complex. By the induction hypothesis, for every $i > 1$ $S_i(v) \cap (\bigcup_{j<i} S_j(v))$ contains a d-face (containing v). Thus, $K(v) := \bigcup_{j=1}^{m(v)} S_j(v)$ has the desired decomposition for every $v \in K$. $K = \bigcup_{v \in \text{Ver}(K)} K(v)$ as $st_K(v) \subseteq K(v)$ for every v, where $st_K(v) = \{T \in K : T \cup \{v\} \in K\}$.

Let v be any vertex of K. Since the 1-skeleton of K is connected, we can order the vertices of K such that $v_1 = v$ and for every $i > 1$ v_i is a neighbor of some v_j where $1 \leq j < i$. Let $v_{l(i)}$ be such a neighbor of v_i. By the induction hypothesis we can order the $S_j(v_i)$'s such that $S_1(v_i)$ will contain $v_{l(i)}$, and hence, as K is pure, will contain a d-face which appears in $K(v_{l(i)})$ (this face contains the edge $\{v_i, v_{l(i)}\}$). The resulting decomposition sequence $(S_1(v_1), \ldots, S_{m(v_1)}(v_1), S_1(v_2), \ldots, S_{m(v_n)}(v_n))$ is as desired.

Moreover, every $S_j(v_{i_0})$ where $i_0 \in [n]$ and $j \in [m(v_{i_0})]$ can be chosen to be the first in such a decomposition sequence. Indeed, by the induction hypothesis $S_j(v_{i_0})$ can be the first in the decomposition sequence of $K(v_{i_0})$, and as mentioned before, the connectivity of the 1-skeleton of K guarantees that each such prefix $(S_1(v_{i_0}), \ldots, S_{m(v_{i_0})}(v_{i_0}))$ can be completed to a decomposition sequence of K on the same $S_j(v_i)$'s. □

Theorem 1.3 follows also from the following corollary combined with Theorem 2.2.

Corollary 3.5 *Let K be a d-dimensional 2-CM simplicial complex over a field k ($d \geq 1$). Then K is a minimal cycle complex over the Abelian group $\tilde{k} = k(x_1, x_2, \ldots)$ whose elements are finite linear combinations of the (variables) x_i's with coefficients in k.*

Proof Consider a decomposition $K = \bigcup_{i=1}^{m} S_i$ as guaranteed by Theorem 3.4, where $S_i = \overline{\text{supp}(c_i)}$ (the closure w.r.t. inclusion of $\text{supp}(c_i)$) for some minimal d-cycle c_i over k. Define $\tilde{c}_i = x_i c_i$, thus $\tilde{c}_i$ is a minimal cycle over $\tilde{k}$. Define $\tilde{c} = \sum_{i=1}^{m} \tilde{c}_i$. Clearly $\tilde{c}$ is a cycle over $\tilde{k}$ whose support spans K. It remains to show that $\tilde{c}$ is minimal. Let $\tilde{c}'$ be a subchain of $\tilde{c}$ which is a cycle, $\tilde{c}' \neq \tilde{c}$. We need to show that $\tilde{c}' = 0$. Denote by $\tilde{\alpha}_T$ ($\tilde{\alpha}'_T$) the coefficient of the set T in $\tilde{c}$ ($\tilde{c}'$) and by $\tilde{\alpha}_T(i)$ the coefficient of the set T in $\tilde{c}_i$. If $\tilde{\alpha}'_T = 0$ then for every i such that $\tilde{\alpha}_T(i) \neq 0$, the minimality of $\tilde{c}_i$ implies that $\tilde{\alpha}'_F = 0$ whenever $\tilde{\alpha}_F(i) \neq 0$. By assumption, there exists a set T_0 such that $\tilde{\alpha}'_{T_0} = 0 \neq \tilde{\alpha}_{T_0}$. In particular, there exists an index i_0 such

that $\tilde{\alpha}_{T_0}(i_0) \neq 0$, hence $\tilde{\alpha}'_F = 0$ whenever $\tilde{\alpha}_F(i_0) \neq 0$. As $S_{i_0} \cap (\bigcup_{j<i_0} S_j)$ contains a d-face in case $i_0 > 1$, repeated application of the above argument implies $\tilde{\alpha}'_F = 0$ whenever $\tilde{\alpha}_F(1) \neq 0$. Repeated application of the fact that $S_i \cap (\bigcup_{j<i} S_j)$ contains a d-face for $i = 2, 3, \ldots$ and of the above argument shows that $\tilde{\alpha}'_F = 0$ whenever $\tilde{\alpha}_F(i) \neq 0$ for some $1 \leq i \leq m$, i.e. $\tilde{c}' = 0$. □

A pure simplicial complex has a *nowhere zero flow* if there is an assignment of integer non-zero wights to all of its facets which forms a $\mathbb{Z}$-cycle. This generalizes the definition of a nowhere zero flow for graphs (e.g. [12] for a survey).

Corollary 3.6 *Let K be a d-dimensional 2-CM simplicial complex over $\mathbb{Q}$ ($d \geq 1$). Then K has a nowhere zero flow.*

Proof Consider a decomposition $K = \bigcup_{i=1}^m S_i$ as guaranteed by Theorem 3.4. Multiplying by a common denominator, we may assume that each $S_i = \overline{\text{supp}(c_i)}$ for some minimal d-cycle c_i over $\mathbb{Z}$ (instead of just over $\mathbb{Q}$). Let N be the maximal $|\alpha|$ over all nonzero coefficients α of the c_i's, $1 \leq i \leq m$. Let $\tilde{c} = \sum_{i=1}^m (N^m)^i c_i$. $\tilde{c}$ is a nowhere zero flow for K; we omit the details. □

Problem 3.7 *Can the S_i's in Theorem* 3.4 *be taken to be homology spheres?*

Yhonatan Iron and I proved (unpublished) the following lemma:

Lemma 3.8 *Let K, L and $K \cap L$ be simplicial complexes of the same dimension $d-1$. Assume that K and L are weak-Lefschetz, i.e. that multiplication by a generic degree-one element g in $H = H(K), H(L)$, $g : H_{i-1} \to H_i$, is injective for all $i \leq \lfloor d/2 \rfloor$. If $K \cap L$ is CM then $K \cup L$ is weak-Lefschetz.*

In view of this lemma, if the intersections $S_i \cap (\bigcup_{j<i} S_j)$ in Theorem 3.4 can be taken to be CM, and the S_i's can be taken to be homology spheres, then Conjecture 1.1 would be reduced to the long standing g-conjecture for homology spheres. Can the intersections be guaranteed to be CM?

Acknowledgements I would like to thank my adviser Gil Kalai, Anders Björner and Ed Swartz for helpful discussions. This research was done during the author's stay at Institut Mittag-Leffler, supported by the ACE network.

References

1. Asimov, L., Roth, B.: The rigidity of graphs. Trans. Am. Math. Soc. **245**, 279–289 (1978)
2. Asimov, L., Roth, B.: The rigidity of graphs: part II. J. Math. Anal. Appl. **68**, 171–190 (1979)
3. Barnette, D.: The minimum number of vertices of a simple polytope. Isr. J. Math. **10**, 121–125 (1971)
4. Barnette, D.: A proof of the lower bound conjecture for convex polytopes. Pac. J. Math. **46**, 349–354 (1973)
5. Billera, L.G., Lee, C.W.: A proof of the sufficiency of McMullen conditions for f-vectors of simplicial convex polytopes. J. Comb. Theory Ser. A **31**, 237–255 (1981)
6. Fogelsanger, A.: The generic rigidity of minimal cycles. PhD dissertation, Cornell University (1988). Also at http://www.people.cornell.edu/pages/alf6/rigidity.htm

7. Kalai, G.: Hyperconnectivity of graphs. Graphs Comb. **1**, 65–79 (1985)
8. Kalai, G.: Rigidity and the lower bound theorem. Invent. Math. **88**, 125–151 (1987)
9. Kalai, G.: Algebraic shifting. Adv. Stud. Pure Math. **33**, 121–163 (2002)
10. Lee, K.W.: Generalized stress and motion. In: Briztriczky, T., et al. (eds.) Polytopes: Abstract, Convex and Computational, pp. 249–271. Kluwer Academic, Dordrecht (1995)
11. Reisner, G.: Cohen–Macaulay quotients of polynomial rings. Adv. Math. **21**, 30–49 (1976)
12. Seymour, P.D.: Nowhere-zero flows. In: Graham, R., et al. (eds.) Handbook of Combinatorics, pp. 289–299. Elsevier, Amsterdam (1995)
13. Stanley, R.P.: The number of faces of simplicial convex polytopes. Adv. Math. **35**, 236–238 (1980)
14. Swartz, E.: g-elements, finite buildings and higher Cohen–Macaulay connectivity. J. Comb. Theory Ser. A **113**, 1305–1320 (2006)
15. Whitney, H.: Non-separable and planar graphs. Trans. Am. Math. Soc. **34**, 339–362 (1932)

Finding the Homology of Submanifolds with High Confidence from Random Samples

Partha Niyogi · Stephen Smale · Shmuel Weinberger

Abstract Recently there has been a lot of interest in geometrically motivated approaches to data analysis in high-dimensional spaces. We consider the case where data are drawn from sampling a probability distribution that has support on or near a submanifold of Euclidean space. We show how to "learn" the homology of the submanifold with high confidence. We discuss an algorithm to do this and provide learning-theoretic complexity bounds. Our bounds are obtained in terms of a condition number that limits the curvature and nearness to self-intersection of the submanifold. We are also able to treat the situation where the data are "noisy" and lie near rather than on the submanifold in question.

1 Introduction

In recent years there has been considerable interest in the possibility of analyzing and processing data in high-dimensional spaces. Following the intuition that naturally

The main results of this paper were first presented at a conference in honor of John Franks and Clark Robinson at Northwestern University in April 2003. These results were formally written as Technical Report No. TR-2004-08, Department of Computer Science, University of Chicago.

P. Niyogi (✉)
Departments of Computer Science and Statistics, University of Chicago, Chicago, IL 60637, USA
e-mail: niyogi@cs.uchicago.edu

S. Smale
Toyota Technological Institute, University Press Building, Chicago, IL 60637, USA
e-mail: smale@tti-c.org

S. Weinberger
Department of Mathematics, University of Chicago, Chicago, IL 60637, USA
e-mail: shmuel@math.uchicago.edu

occurring data may be generated by structured systems with possibly much fewer degrees of freedom than the ambient dimension would suggest, various researchers (see [3, 10, 16, 17, 20]) have considered the case when the data live on or close to a submanifold of the ambient space. One hopes then to estimate geometrical and topological properties of the submanifold from random points ("scattered data") lying on this unknown submanifold. These questions belong to a class of problems that have come to be known as *manifold learning*.

In this paper we consider the particular question of identifying the homology of the submanifold from random samples. The homology of the submanifold (see [15] for definitions) are natural topological invariants that provide a good characterization of many aspects of it. For example, the dimensions of the homology groups, the Betti numbers $(b_0, b_1, \ldots)$, have natural interpretations. b_0, the dimension of the zeroth homology group is the number of connected components of the submanifold. In data analysis situations, the number of clusters of the data may sometimes be understood in terms of the number of components of an underlying manifold (or other geometric object). If the dimension of the submanifold is d, then one sees that $b_j = 0$ for all $j > d$. Thus the largest non-trivial homology gives us the dimension of the submanifold. If the submanifold is two-dimensional, then b_0 and b_1 are related to the number of connected components and number of holes, respectively, of the submanifold.

We show that it is possible to identify the homology from random samples and discuss an algorithm to do this. There are a few aspects of the developments in this paper that are worth emphasizing. First, we provide sample complexity estimates on the number of examples that are needed to identify the homology with high confidence. Our results are in the style of learning–theoretic treatments (for example, the *Probably Approximately Correct* framework [18]) where unknown objects (typically functions in learning theory) are "learned" from random samples and confidence estimates are provided. Second, we treat the situation where data might be drawn from a distribution that is concentrated *around* the manifold rather than precisely on it. Under specific models of noise, we show that our algorithm can work even with noisy data. In all cases, estimates are provided in terms of a condition number that limits the curvature and nearness to self-intersection of the submanifold.

Our results may also be of interest to researchers in computational geometry and topology who have considered the question of computing homology from simplicial complexes in the past (see [8, 14] for details and further references). A number of researchers in these computational geometry and topology fields have considered the problem of manifold reconstruction from point cloud data. Such work has typically focused on the case of surfaces in $\mathbb{R}^3$ and examples include algorithms associated with the frameworks of alpha shapes [11], CRUST [1] and its variants, and COCONE [2] and its generalizations. CRUST and COCONE provably recover a simplicial 2-manifold that is homeomorphic to the surface. In [6] (written after the results of our current paper were declared), it was shown how to extend these ideas to the general setting of a k-manifold embedded in $\mathbb{R}^N$. In much of this work the medial axis plays a central role in characterizing the conditioning of the manifold (see our later remarks in Sect. 2). It is also worth noting that none of the works mentioned above considers the probabilistic setting where examples are drawn at random—so no high confidence guarantees are provided. The theorems in [1, 2, 6] are analogous to our Proposition 3.1. No version of our main theorem (Theorem 3.1) exists in the

literature. Finally, it is also worth noting that there is a body of work on persistence homology [7, 20] that seeks alternative topological characterizations of the manifold and its homology. See the discussion after Proposition 3.1.

In conclusion, we hope that researchers in graphics, pattern recognition, solid modeling, molecular biology, finance, and other areas where large amounts of high-dimensional data are available may find some use for the topological perspective on data analysis embodied in the algorithms and analyses of this paper.

2 Preliminaries

Consider a compact Riemannian submanifold $\mathcal{M}$ of a Euclidean space $\mathbb{R}^N$. Sample the manifold according to a uniform probability measure on it. Thus points $x_1, \ldots, x_n \in \mathcal{M}$ are generated. This set of points $\bar{x} = \{x_1, \ldots, x_n\}$ is the data set on the basis of which homology groups will be calculated. In later sections we consider the case when the data are drawn from a probability measure with support close to the manifold.

Throughout our discussion, we associate to $\mathcal{M}$ a condition number $(1/\tau)$ where τ is defined as the largest number having the property: The open normal bundle about $\mathcal{M}$ of radius r is embedded in $\mathbb{R}^N$ for every $r < \tau$. Its image Tub_τ is a tubular neighborhood of $\mathcal{M}$ with its canonical projection map

$$\pi_0 : \mathrm{Tub}_\tau \to \mathcal{M}.$$

Note that τ encodes both local curvature considerations as well as global ones: If $\mathcal{M}$ is a union of several components, then τ bounds their separation. For example, if $\mathcal{M}$ is a sphere, then τ is equal to its radius. If $\mathcal{M}$ is an annulus, then τ is the separation of its components. In Sect. 6 we relate the condition number $1/\tau$ to classical notions of curvature in differential geometry via the second fundamental form.

Finally, it is also useful to relate τ to the notions of medial axis and local feature size that have been developed in the computational geometry community. Given $\mathcal{M}$, one may define the set

$$G = \left\{ x \in \mathbb{R}^N \text{ such that } \exists \text{ distinct } p, q \in \mathcal{M} \text{ where } d(x, \mathcal{M}) = \|x - p\| = \|x - q\| \right\},$$

where $d(x, \mathcal{M}) = \inf_{y \in \mathcal{M}} \|x - y\|$ is the distance of x to $\mathcal{M}$. The closure of G is called the medial axis and for any point $p \in \mathcal{M}$ the local feature size $\sigma(p)$ is the distance of p to the medial axis. Then it is easy to check that

$$\tau = \inf_{p \in \mathcal{M}} \sigma(p).$$

3 An Outline of Our Main Results

Ultimately we wish to compute the homology of the manifold $\mathcal{M} \subset \mathbb{R}^N$ from the randomly sampled datapoints $\bar{x} = \{x_1, \ldots, x_n\} \subset \mathcal{M}$. We first begin by considering

Euclidean balls (in the ambient space $\mathbb{R}^N$) of radius ϵ and center x_i. We denote these balls as $B_\epsilon(x_i)$. We can now define the open set $U \subset \mathbb{R}^N$ given by

$$U = \bigcup_{x \in \bar{x}} B_\epsilon(x).$$

Our first proposition states that if $\bar{x} = \{x_1, \ldots, x_n\}$ is $\epsilon/2$ dense in $\mathcal{M}$, then $\mathcal{M}$ is a deformation retract of U.

Proposition 3.1 *Let $\bar{x}$ be any finite collection of points $x_1, \ldots, x_n \in \mathbb{R}^N$ such that it is $(\epsilon/2)$ dense in $\mathcal{M}$, i.e., for every $p \in \mathcal{M}$, there exists an $x \in \bar{x}$ such that $\|p - x\|_{\mathbb{R}^N} < \epsilon/2$. Then for any $\epsilon < \sqrt{\frac{3}{5}}\tau$, we have that U deformation retracts to $\mathcal{M}$. Therefore the homology of U equals the homology of $\mathcal{M}$.*

We prove this proposition in Sect. 4. Subsequent to our work, the authors of [7] presented a different type of calculation of the homology of $\mathcal{M}$ based on their homology approximation theorem together with the method of computing persistent homology (e.g., [20]). Their method does not give the homotopy type of $\mathcal{M}$. On the other hand, it does apply to a class of metric spaces more general than well-conditioned manifolds. A related approach appears in [5].

In the case under consideration here, the points $x_1, \ldots, x_n$ are sampled in i.i.d. fashion from the uniform probability distribution on $\mathcal{M}$. By probabilistic considerations, we will then prove (in Sect. 5) the following proposition.

Proposition 3.2 *Let $\bar{x}$ be drawn by sampling $\mathcal{M}$ in i.i.d. fashion according to the uniform probability measure on $\mathcal{M}$. Then with probability greater than $1 - \delta$, we have that $\bar{x}$ is $(\epsilon/2)$-dense $(\epsilon < \tau/2)$ in $\mathcal{M}$ provided*

$$|\bar{x}| > \beta_1 \left(\log(\beta_2) + \log\left(\frac{1}{\delta}\right)\right),$$

where

$$\beta_1 = \frac{vol(\mathcal{M})}{(\cos^k(\theta_1))vol(B^k_{\epsilon/4})} \quad \text{and} \quad \beta_2 = \frac{vol(\mathcal{M})}{(\cos^k(\theta_2))vol(B^k_{\epsilon/8})}.$$

Here k is the dimension of the manifold $\mathcal{M}$ and $vol(B^k_\epsilon)$ denotes the k-dimensional volume of the standard k-dimensional ball of radius ϵ. Finally, $\theta_1 = \arcsin(\epsilon/8\tau)$ and $\theta_2 = \arcsin(\epsilon/16\tau)$.

Putting these two propositions together, we see that we are able to provide a finite sample estimate for how many times we need to sample $\mathcal{M}$ so that we are guaranteed with high confidence that the homology of the random set U equals the homology of $\mathcal{M}$. Thus our main theorem is

Theorem 3.1 *Let $\mathcal{M}$ be a compact submanifold of $\mathbb{R}^N$ with condition number τ. Let $\bar{x} = \{x_1, \ldots, x_n\}$ be a set of n points drawn in i.i.d. fashion according to the*

uniform probability measure on $\mathcal{M}$. *Let* $0 < \epsilon < \tau/2$. *Let* $U = \bigcup_{x \in \bar{x}} B_\epsilon(x)$ *be a correspondingly random open subset of* $\mathbb{R}^N$. *Then for all*

$$n > \beta_1 \left(\log(\beta_2) + \log\left(\frac{1}{\delta}\right)\right),$$

the homology of U *equals the homology of* $\mathcal{M}$ *with high confidence* (*probability* $> 1 - \delta$).

Remark Note that no version of our main theorem exists in the literature so far. However, versions of our Proposition 3.1 do exist. We have characterized Proposition 3.1 in terms of τ but one may obtain an alternate characterization in terms of the medial axis and the local feature size. In fact, if one considers the union of balls centered at the data points given by $U = \bigcup_{x \in \bar{x}} B_{\epsilon_x}(x)$ where $\epsilon_x = r\sigma(x)$, then it is possible to show that the homology of U coincides with that of $\mathcal{M}$ if $\bar{x}$ is $(\epsilon_x/2)$-dense in $\mathcal{M}$ and for all $r < 0.21$. For the case of surfaces in $\mathbb{R}^3$, a similar result is obtained by Amenta et al. [2] for $r < 0.06$. The set $\bar{x}$ is said to be $(\epsilon_x/2)$-dense if for every $p \in \mathcal{M}$ there exists some $x \in \bar{x}$ such that $\|p - x\| < \epsilon_x/2$. We will prove this in a later paper. It is not obvious, however, how to obtain a version of our main theorem in terms of the local feature size. Finally, we recall the recent results of [7] that we have already alluded to.

3.1 Computing the Homology of U

One now needs to consider algorithms to compute the homology of U. Noting that the $B_\epsilon(x_i)$'s form a cover of U, one can construct the *nerve* of the cover. The nerve is an abstract simplicial complex constructed as follows: One puts in a k-simplex for every $(k+1)$-tuple of intersecting elements of the cover. The Nerve Lemma (see [4]) applies in our case, as balls are convex, to show that the homology of U is the same as the homology of this complex. The algorithm consists of the following components:

1. Given an ϵ, and a set of points $\bar{x} = \{x_1, \ldots, x_n\}$ in $\mathbb{R}^N$, each j-simplex is given by a subset of the n points that have non-zero intersection. Thus we may define L_j to be the collection of all j-simplices. Each simplex $\sigma \in L_j$ is associated with a set of $j+1$ points $(p_0(\sigma), \ldots, p_j(\sigma) \in \bar{x})$ such that

$$\bigcap_{i=0}^{j} B_\epsilon(p_i(\sigma)) \neq \emptyset.$$

   An orientation for the simplex is chosen by picking an ordering and we denote the oriented simplex by $|p_0(\sigma), \ldots, p_j(\sigma)|$.
2. A very crude upper bound on the size of L_j (denoted by $|L_j|$) is given by $\binom{n}{j+1}$. However, it is clear that if two points x_m and x_l are more than 2ϵ apart, they cannot be associated to a simplex. Therefore, there is a locality condition that the $p_i(\sigma)$'s must obey, which results in $|L_j|$ being much smaller than this crude number. The simplicial complex $K_j = \bigcup_{i=0}^{j} L_j$ together with face relations. The simplicial complex corresponding to the nerve of U is $K = K_N$.

3. A basic subroutine for computing the simplicial complex (steps 1 and 2 above) involves the decision problem: for any set of j points, determine whether balls of radius ϵ around each of these points have non-empty intersection. This problem is related to the smallest ball problem defined as follows: Given a set of j points, find the ball with the smallest radius enclosing all these points. One can check that $\bigcap_{i=1}^{j} B_\epsilon(p_i) \neq \emptyset$ if and only if this smallest radius $< \epsilon$. Fast algorithms for the smallest ball problem exist. See [12] for theoretical discussion and [14] for downloadable algorithms from the web.
4. We work in the field of coefficients $\mathbb{R}$. Then a j-chain is a function $c\colon L_j \to \mathbb{R}$ and can be written as a formal sum

$$c = \sum_{\sigma \in L_j} c(\sigma)\sigma.$$

   By adding j-chains componentwise, one gets the vector space of j-chains denoted by C_j.
5. The boundary operator ∂_j is a linear operator from C_j to C_{j-1} defined as follows. For each (oriented) simplex $\sigma \in L_j$,

$$\partial_j \sigma = \sum_{i=0}^{j} (-1)^i \sigma_i,$$

   where σ_i is a $j-1$ face of σ (facing point $p_i(\sigma)$) and the orientation of σ_i is given by $|p_0, \ldots, p_{i-1}, p_{i+1}, \ldots, p_j|$. Now ∂_j is defined on j chains by additivity as

$$\partial_j \left(\sum_{\sigma \in L_j} c(\sigma)\sigma \right) = \sum_{\sigma \in L_j} c(\sigma)\partial_j \sigma.$$

   Thus, ∂_j can be represented as an $n_{j-1} \times n_j$ matrix where $n_{j-1} = |L_{j-1}|$ and $n_j = |L_j|$, respectively. The matrix is usually sparse in our setting.
6. This defines the chain complex

$$\cdots C_{j+1} \xrightarrow{\partial_{j+1}} C_j \xrightarrow{\partial_j} C_{j-1} \cdots.$$

   One can finally define the *image* and *kernel* of the boundary operator given by

$$\operatorname{Im} \partial_j = \{c \in C_{j-1} \mid \exists c' \in C_j \text{ where } \partial_j c' = c\}$$

   and

$$\operatorname{Ker} \partial_j = \{c \in C_j \mid \partial_j c = 0\}.$$

   Now $\operatorname{Im} \partial_{j+1}$ is the vector space of j-boundaries and $\operatorname{Ker} \partial_j$ is the vector space of j cycles. Then the jth homology group is the quotient of $\operatorname{Ker} \partial_j$ over $\operatorname{Im} \partial_{j+1}$, i.e.,

$$H_j = \operatorname{Ker} \partial_j / \operatorname{Im} \partial_{j+1}.$$

The calculation of H_j is seen to be an exercise in linear algebra given the matrix representation of the boundary operators. In our exposition here, we have been working over a field resulting in vector spaces which are characterized purely by their ranks (the Betti numbers). One approach to this is also via the combinatorial Laplacian as outlined in [13]. More generally, one can work over a ring and H_j would then be an Abelian group.

4 The Deformation Retract Argument

In this section we prove Proposition 3.1. Recall that $\epsilon < \sqrt{3/5}\tau$. Consider the canonical map $\pi : U \to \mathcal{M}$ given by (π is the restriction of π_0 to U)

$$\pi(x) = \arg \min_{p \in \mathcal{M}} \|x - p\|.$$

Then we see that the fibers $\pi^{-1}(p)$ are given by $T_p^{\perp} \cap U \cap B_\tau(p)$. The intersection with $B_\tau(p)$ is necessary to eliminate distant regions of U that may intersect with T_p (because the manifold may curve around over great distances) but do not belong to the fiber. For example, for the standard circle in $\mathbb{R}^2$, at any point p on the circle, $T_p^{\perp}$ intersects the circle at two points. One of these is in $B_\tau(p)$ and the other is not. Therefore,

$$\pi^{-1}(p) = \bigcup_{x \in \bar{x}} B_\epsilon(x) \cap T_p^{\perp} \cap B_\tau(p),$$

where $T_p^{\perp}$ is the normal subspace at $p \in \mathcal{M}$ orthogonal to the tangent space T_p. Let us also define $st(p)$ as

$$st(p) = \bigcup_{\{x \in \bar{x}; x \in B_\epsilon(p)\}} B_\epsilon(x) \cap T_p^{\perp} \cap B_\tau(p).$$

It is immediately clear that

$$st(p) \subseteq \pi^{-1}(p).$$

Then the following simple proposition is true.

Proposition 4.1 *$st(p)$ is star shaped relative to p and therefore contracts to p.*

Proof Consider arbitrary $v \in st(p)$. Then $v \in B_\epsilon(x) \cap T_p^{\perp}$ for some $x \in \bar{x}$ such that $x \in B_\epsilon(p)$. Since $x \in B_\epsilon(p)$, we immediately have $p \in B_\epsilon(x)$. Since v, p are both in $B_\epsilon(x)$, by convexity of Euclidean balls, we have that the line segment $v\bar{p}$ joining v to p is entirely contained in $B_\epsilon(x)$. At the same time, $v\bar{p}$ is entirely contained in $T_p^{\perp}$ and it follows therefore that $v\bar{p}$ is contained in $st(p)$. □

We next show that the inclusion of $st(p)$ in $\pi^{-1}(p)$ is an equality proving that $\pi^{-1}(p)$ contracts to p.

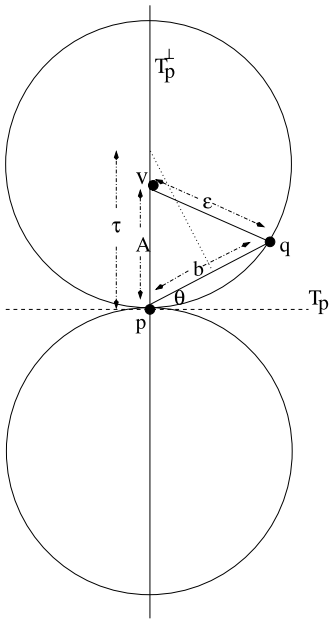


Fig. 1 A picture showing the worst case. The picture shows the plane passing through points v, p, and q. T_p and $T_p^{\perp}$ are shown intersecting with this plane and are represented by the *dotted horizontal line* and the *solid vertical line*, respectively. On the plane of interest, one may then draw two *circles* (of radius τ each) that are tangent to T_p and are on *either side* of T_p as shown. Clearly, v lies on $T_p^{\perp}$ and is marked in the figure. On the other hand, q could potentially lie anywhere *outside* the *two circles*. A moment's reflection shows that $\|v - p\|$ is greatest when q lies on *one* of the *two circles*. Without loss of generality one may consider it to lie on the *top circle* as shown. Over all choices of such q, the worst case is derived in Lemma 4.1

Proposition 4.2

$$st(p) = \pi^{-1}(p).$$

Proof We need to show that $\pi^{-1}(p) \subseteq st(p)$. Consider an arbitrary $v \in B_\epsilon(q) \cap T_p^{\perp} \cap B_\tau(p)$ where $q \in \bar{x}$ and $q \notin B_\epsilon(p)$. For such v the picture of Fig. 1 can be drawn. Following Lemma 4.1, we see that the distance of v to p is at most ϵ^2/τ. Now by the fact that $\bar{x}$ is $(\epsilon/2)$-dense, we have that there is some point $x \in \bar{x}$ which is within $\epsilon/2$ of p. The worst-case picture of this is shown in Fig. 2. From Lemma 4.2, we see that $v \in B_\epsilon(x)$ for this x. The proposition is proved. □

These two propositions taken together show that $\mathcal{M}$ is a deformation retract of U. We see that $\mathcal{M} \subset U$. Further let $F(x,t) : U \times [0,1] \to U$ be given by $F(x,t) = tx + (1-t)\pi(x)$. Then F is continuous, $F(x,0) = \pi$, and $F(x,1)$ is the identity map.

Lemma 4.1 *Consider any $q \notin B_\epsilon(p)$. Let $v \in B_\epsilon(q) \cap T_p^{\perp} \cap B_\tau(p)$. Then the Euclidean distance from v to p is less than ϵ^2/τ.*

Proof We need to consider which configuration of v, q, and p makes the distance $\|v - p\|$ as large as possible. It is easiest to reason about this in the plane passing through these points. It suffices to consider q on the curve as shown in Fig. 1. See the caption for further explanation. Following the symbols on the figure, we have

$$A = b\sin(\theta) + \sqrt{\epsilon^2 - b^2\cos^2(\theta)},$$

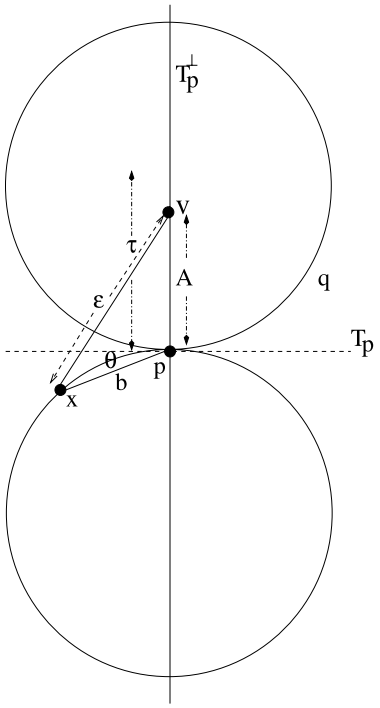


Fig. 2 A picture showing the worst case. The picture is of the plane containing the points p, v, and x. The *two circles* are each of radius τ and tangent to T_p. T_p and $T_p^{\perp}$ are represented by their intersection with the plane of interest as *dotted horizontal* and *solid vertical lines*, respectively

where $b = 2\tau \sin(\theta)$. Therefore, we have

$$A = 2\tau \sin^2(\theta) + \sqrt{\epsilon^2 - 4\tau^2 \sin^2(\theta)\cos^2(\theta)}.$$

From this we see that

$$\begin{aligned}\frac{dA}{d\theta} &= 2\tau \sin(2\theta) - \frac{4\tau^2 \sin(2\theta)\cos(2\theta)}{2\sqrt{\epsilon^2 - \tau^2 \sin^2(2\theta)}} \\ &= 2\tau \sin(2\theta)\left(1 - \frac{\tau \cos(2\theta)}{\sqrt{\epsilon^2 - \tau^2 \sin^2(2\theta)}}\right).\end{aligned}$$

It is easy to check that if $\epsilon < \tau$, then $dA/d\theta < 0$, i.e., A is monotonically decreasing with θ. Therefore the worst-case situation is when $b = 2\tau \sin(\theta) = \epsilon$. For this value of θ, we see that $A = \epsilon^2/\tau$. □

The following lemma ensures that there is an $x \in \bar{x} \cap B_\epsilon(p)$ such that $v \in B_\epsilon(x) \cap T_p^{\perp}$.

Lemma 4.2 *Let $\bar{x}$ be $(\epsilon/2)$-dense in $\mathcal{M}$. For any $p \in \mathcal{M}$, let $v \in \pi^{-1}(p)$. Then for $0 < \epsilon < \sqrt{3/5}\tau$, we have that $v \in B_\epsilon(x) \cap T_p^{\perp}$ for some $x \in B_\epsilon(p) \cap \bar{x}$.*

Proof By the $(\epsilon/2)$-dense property, we know that there is an $x \in \bar{x}$ such that $x \in B_{\epsilon/2}(p)$. Consider the picture in Fig. 2. This represents the most unfavorable position that such an x might have for the current context. The picture shows the plane passing through the points x, v, and p. By the same argument of Lemma 4.1 we see that

$$A = \sqrt{\epsilon^2 - b^2 \cos^2(\theta)} - b \sin(\theta),$$

where $b = 2\tau \sin(\theta) = \epsilon/2$. Putting this value in, we have

$$A = \sqrt{\epsilon^2 - \frac{\epsilon^2}{4}\left(1 - \frac{\epsilon^2}{16\tau^2}\right)} - 2\tau \frac{\epsilon^2}{16\tau^2}.$$

Simplifying, we see that $A > \epsilon^2/\tau$ (needed by Lemma 4.1) if

$$\sqrt{\epsilon^2 - \frac{\epsilon^2}{4}\left(1 - \frac{\epsilon^2}{16\tau^2}\right)} > \frac{9}{8}\frac{\epsilon^2}{\tau}.$$

Squaring both sides, we have

$$\frac{3}{4}\epsilon^2 + \frac{\epsilon^4}{64\tau^2} > \frac{81\epsilon^4}{64\tau^2}.$$

This simplifies to

$$\frac{\epsilon^2}{\tau^2} < \frac{3}{5}.$$

Therefore, as long as $\epsilon < \sqrt{\frac{3}{5}}\tau$, we will have that $v \in B_\epsilon(x)$ for a suitable x. □

5 Probability Bounds

Following our assumption, that the points x_i are drawn at random, we now provide a bound on how many examples need to be drawn so that the empirically constructed complex has the same homology as the manifold. We begin with a basic probability lemma.

Lemma 5.1 *Let $\{A_i\}$ for $i = 1, \ldots, l$ be a finite collection of measurable sets and let μ be a probability measure on $\bigcup_{i=1}^{l} A_i$ such that for all $1 \leq i \leq l$, we have $\mu(A_i) > \alpha$. Let $\bar{x} = \{x_1, \ldots, x_n\}$ be a set of n i.i.d. draws according to μ. Then if*

$$n \geq \frac{1}{\alpha}\left(\log l + \log\left(\frac{1}{\delta}\right)\right)$$

we are guaranteed that with probability $> 1 - \delta$, the following is true:

$$\forall i, \quad \bar{x} \cap A_i \neq \emptyset.$$

Proof This follows from a simple application of the union bound. Let E_i be the event that $\bar{x} \cap A_i$ is empty. The probability with which this happens is given by

$$\mathbb{P}E_i = \big(1 - \mu(A_i)\big)^n \leq (1-\alpha)^n.$$

Therefore, by the union bound, we have

$$\mathbb{P}\bigcup\nolimits_i E_i \leq \sum_{i=1}^{l} \mathbb{P}E_i \leq l(1-\alpha)^n.$$

It remains to show that for $n \geq (1/\alpha)(\log l + \log(1/\delta))$, we have

$$l(1-\alpha)^n \leq \delta.$$

To see this, simply note that $f(x) = xe^x - e^x + 1 \geq 0$ for all $x \geq 0$. This is seen by noting that $f(0) = 0$ and $f'(x) = xe^x \geq 0$ for all $x \geq 0$. Putting $x = \alpha$ in the above function, we have

$$(1-\alpha) \leq e^{-\alpha}$$

and therefore it is easily seen that

$$l(1-\alpha)^n \leq le^{-n\alpha} \leq \delta$$

for the appropriate choice of n. □

Applying this to our setting, we consider a cover of the manifold $\mathcal{M}$ by balls of radius $\epsilon/4$. Let $\{y_i;\ 1 \leq i \leq l\}$ be the centers of such balls that constitute a minimal cover. Therefore, we can choose $A_i = B_\epsilon/4(y_i) \cap \mathcal{M}$. Applying the above lemma, we immediately have an estimate on the number of examples we need to collect. This is given by

$$\frac{1}{\alpha}\left(\log l + \log\left(\frac{1}{\delta}\right)\right),$$

where

$$\alpha = \min_i \frac{vol(A_i)}{vol(\mathcal{M})}$$

and l is the $\epsilon/4$ covering number. These may be expressed entirely in terms of natural invariants of the manifold and we derive these quantities below.

First, we note that the covering number may be bounded in terms of the packing number, i.e., the maximum number of sets of the form $N_i = B_r \cap \mathcal{M}$ (at scale r) that may be packed into $\mathcal{M}$ without overlap. In particular, if $C(\epsilon)$ is the ϵ-covering number of $\mathcal{M}$ and $P(\epsilon)$ is the ϵ-packing number, then the following simple lemma is true.

Lemma 5.2

$$P(2\epsilon) \leq C(2\epsilon) \leq P(\epsilon).$$

Proof The fact that $P(2\epsilon) \leq C(2\epsilon)$ follows from the definition. To see that $C(2\epsilon) \leq P(\epsilon)$, begin by letting $B_\epsilon(x_1), \ldots, B_\epsilon(x_N)$ be a realization of an optimal ϵ-packing so that $N = P(\epsilon)$. We claim that $B_{2\epsilon}(x_1), \ldots, B_{2\epsilon}(x_N)$ form a 2ϵ-cover. If not, there exists an $x \in \mathcal{M}$ such that $B_\epsilon(x) \cap B_\epsilon(x_i)$ is empty for all i. In that case, one can add $B_\epsilon(x)$ to the collection to increase the packing number by 1 leading to a contradiction. Since $B_{2\epsilon}(x_1), \ldots, B_{2\epsilon}(x_N)$ is a valid 2ϵ-cover, we have $C(2\epsilon) \leq N = P(\epsilon)$. □

Since l is the $\epsilon/4$ covering number, we see that $l \leq P(\epsilon/8)$ from Lemma 5.2. Now we need to bound the packing number. To do so, we need the following result.

Lemma 5.3 *Let $p \in \mathcal{M}$. Now consider $A = \mathcal{M} \cap B_\epsilon(p)$. Then $vol(A) \geq (\cos(\theta))^k vol(B_\epsilon^k(p))$ where $B_\epsilon^k(p)$ is the k-dimensional ball in T_p centered at p, $\theta = \arcsin(\epsilon/2\tau)$. All volumes are k-dimensional volumes where k is the dimension of $\mathcal{M}$.*

Proof Consider the tangent space at p given by T_p and let f be the projection of $\mathbb{R}^N$ to T_p. Let $B_r^k(p)$ be the k-dimensional ball of radius $r = \epsilon\cos(\theta)$ (where $\theta = \arcsin(\epsilon/2\tau)$) centered at p lying in T_p. Let $f_A = \{f(q) \mid q \in A\}$ be the image of A under f. We will show that $B_r^k(p) \subset f_A$. Since f is a projection we have

$$vol(A) \geq vol(f_A) \geq vol\big(B_r^k(p)\big) = \big(\cos^k(\theta)\big) vol\big(B_\epsilon^k(p)\big).$$

To see that $B_r^k(p) \subset f_A$, notice that f is an open map whose derivative is non-singular for all $q \in A$ (by Lemma 5.4). Therefore f is locally invertible and there exists a ball $B_s^k(p)$ of radius s such that $f^{-1}(B_s^k(p)) \subset A$. One can keep increasing s until it happens for the first time (say at $s = s'$) that $f^{-1}(B_s^k(p)) \not\subset A$. At this stage, there exists a point q in the closure of A such that either (i) f is singular at q or (ii) $q \notin A$. By Lemma 5.4, we see that (i) is impossible. Therefore, $q \notin A$ but q is in the closure of A implying that $\|q - p\| = \epsilon$. We see that $s' = \epsilon\cos(\phi)$ where ϕ is the angle between the line $\bar{qp}$ (the line joining q to p) and the line $f(\bar{q})p$ (the line joining $f(q)$ to p). By the curvature bound implied by τ, we see that $|\phi| \leq |\theta|$ and therefore $s' = \epsilon\cos(\phi) \geq \epsilon\cos(\theta) = r$. □

Lemma 5.4 *Let $p \in \mathcal{M}$, let $A = \mathcal{M} \cap B_\epsilon(p)$, and let f be the projection to the tangent space at p (T_p). Then for all $\epsilon < \tau/2$, the derivative df is non-singular at all points $q \in A$.*

Proof Suppose $d\,f$ was singular for some $q \in A$. That means that the tangent space at $q(T_q)$ is oriented so that the vector with origin q and endpoint $f(q)$ lies in T_q. Since $q \in B_\epsilon(p)$, we have that $d = \|q - p\| < \tau/2$. Putting Propositions 6.2 and 6.3 together, we get that

$$\cos(\phi) \geq \sqrt{1 - \frac{2d}{\tau}} > 0,$$

where ϕ is the angle between T_p and T_q. From this we see that $\phi < \pi/2$ leading to a contradiction. □

Using Lemma 5.3, we see that a simple bound on the packing number is obtained. We obtain immediately that

$$P(\epsilon) \leq \frac{vol(\mathcal{M})}{(\cos^k(\theta))vol(B^k_\epsilon(p))}.$$

Therefore, we have

$$l \leq P\left(\frac{\epsilon}{8}\right) \leq \frac{vol(\mathcal{M})}{(\cos^k(\theta_2))vol(B^k_{\frac{\epsilon}{8}}(p))},$$

where $\theta_2 = \arcsin(\epsilon/16\tau)$. Similarly, we have that

$$\frac{1}{\alpha} \leq \frac{vol(\mathcal{M})}{(\cos^k(\theta_1))vol(B^k_{\frac{\epsilon}{4}}(p))},$$

where $\theta_1 = \arcsin(\epsilon/8\tau)$.

6 Curvature and the Condition Number $1/\tau$

In this section[1] we examine the consequences of the condition number $1/\tau$ for the submanifold $\mathcal{M}$. As we have mentioned before, τ controls the curvature of the manifold at every point. This fact has been exploited in our earlier proofs. For submanifolds, one may formally study curvature through the second fundamental form (see, e.g., [9]). Here we show formally that the norm of the second fundamental form is bounded by $1/\tau$. Thus a large τ corresponds to a well-conditioned submanifold that has low curvature.

Proposition 6.1 states the bound on the norm of the second fundamental form. Proposition 6.2 states a bound on the maximum angle between tangent spaces at different points in $\mathcal{M}$. Proposition 6.3 states a bound on the maximum difference between the geodesic distance and the ambient distance for neighboring points in $\mathcal{M}$.

We begin by recalling the second fundamental form. Fix a point $p \in \mathcal{M}$. Following standard accounts (see, e.g., [9]), there exists a symmetric bilinear form $B: T_p \times T_p \to T_p^\perp$ that maps any two vectors in the tangent space $(u, v \in T_p)$ into a vector $B(u, v)$ in the normal space. Thus for any normal vector (unit norm) $\eta \in T_p^\perp$, one can define the following:

$$B_\eta(u, v) = \big\langle \eta, B(u, v) \big\rangle = \langle u, L_\eta v \rangle,$$

where the inner product $\langle \cdot, \cdot \rangle$ is the usual inner product in the tangent space of the ambient manifold (in our case $\mathbb{R}^N$). Since B_η: $T_p \times T_p \to \mathbb{R}$ is symmetric and bilinear, we see that $L_\eta: T_p \to T_p$ is a linear self-adjoint operator. The norm of the second fundamental form in direction η is now given by

$$\lambda_\eta = \sup_{u \in T_p} \frac{\langle u, L_\eta u \rangle}{\langle u, u \rangle}.$$

[1] Thanks to Nat Smale for discussions leading to the writing of this section.

It is seen that λ_η is the largest eigenvalue of L_η. (In general, the eigenvalues are also known as the principal curvatures in the normal direction η.) Given this, we can prove the following proposition that characterizes the relation between the curvature through the second fundamental form and the condition number of the submanifold.

Proposition 6.1 *If $\mathcal{M}$ is a submanifold of $\mathbb{R}^N$ with condition number $1/\tau$, then the norm of the second fundamental form is bounded by $1/\tau$ in all directions. In other words, for all points $p \in \mathcal{M}$ and for all (unit norm) $\eta \in T_p^\perp$, we have*

$$\lambda_\eta = \sup_{u \in T_p} \frac{\langle u, L_\eta u \rangle}{\langle u, u \rangle} \leq \frac{1}{\tau}.$$

Proof We prove by contradiction. Suppose the proposition is false. Then there exists a point $p \in \mathcal{M}$, a tangent vector (unit norm) $u \in T_p$, and a normal vector (unit norm) η such that

$$\big\langle \eta, B(u,u) \big\rangle > \frac{1}{\tau}.$$

Consider a geodesic curve $c(t) \in \mathcal{M}$ parametrized by arc length such that $c(0) = p$ and $\dot{c}(0) = (dc/dt)(0) = u$. For convenience, we place the origin at p so that $c(0) = 0 = p$. With this (ambient) coordinate system, consider the point given by $\tau\eta$, i.e., the point a distance τ from p in the direction η. By our hypothesis on the condition number of the submanifold, we see that $p \in \mathcal{M}$ is the closest point on the manifold to the center of the τ-ball given by $\tau\eta$:

$$\text{for all } t, \quad \big\| c(t) - \tau\eta \big\|^2 \geq \tau^2$$

from which we get

$$\text{for all } t, \quad \big\langle c(t), c(t) \big\rangle - 2\tau \big\langle c(t), \eta \big\rangle \geq 0.$$

Consider the function $g(t) = \langle c(t), c(t) \rangle - 2\tau \langle c(t), \eta \rangle$. Since $c(0) = 0$, we see that $g(0) = 0$. Further, we have $g'(t) = 2\langle c(t), \dot{c}(t) \rangle - 2\tau \langle \dot{c}(t), \eta \rangle$. Since $c(0) = 0$ and $\langle \dot{c}(0), \eta \rangle = 0$, we see that $g'(0) = 0$. Finally, $g''(t) = 2\langle \dot{c}(t), \dot{c}(t) \rangle + 2\langle c(t), \ddot{c}(t) \rangle - 2\tau \langle \ddot{c}(t), \eta \rangle$. Since c is parametrized by arc length, we have $\langle \dot{c}(t), \dot{c}(t) \rangle = 1$ and $g''(0) = 2 - 2\tau \langle \ddot{c}(0), \eta \rangle$.

Noting that the tangent vector field dc/dt is parallel (see the proof of Proposition 6.2), we see that $B(dc/dt, dc/dt) = \ddot{c}(t)$. Therefore, by assumption, we have that

$$\big\langle \eta, B(u,u) \big\rangle = \left\langle \eta, B\left(\frac{dc}{dt}, \frac{dc}{dt} \right) \right\rangle = \big\langle \eta, \ddot{c}(0) \big\rangle > \frac{1}{\tau}.$$

Therefore, $g''(0) < 2 - 2\tau(1/\tau) = 0$. By continuity, there exists a t^* such that $g(t^*) < 0$. However, this leads to a contradiction since $g(t) \geq 0$ for all t. □

Since the norm of the second fundamental form is bounded, we see that the manifold cannot curve too much locally. As a result, the angle between tangent spaces at nearby points cannot be too large. Let p and q be two points in the submanifold $\mathcal{M}$

with associated tangent spaces T_p and T_q. Since T_p and T_q are affine subspaces of $\mathbb{R}^N$, one can compare them in the ambient space in a standard way.

Formally, one may transport the tangent spaces to the origin (according to the standard connection defined in the ambient space $\mathbb{R}^N$) and then compare vectors in each of these tangent spaces with each other. Thus for any (unit norm) vectors $u \in T_p$ and $v \in T_q$, we may define the angle θ between them by

$$\cos(\theta) = \big|\langle u', v' \rangle\big|,$$

where $\langle \cdot, \cdot \rangle$ is the usual inner product in $\mathbb{R}^N$, and u', v' are the vectors obtained by parallel transport (in $\mathbb{R}^N$) of u and v, respectively, to the origin. Hereafter, we always take this construction as standard. We drop the prime notation and use $\langle u, v \rangle$ to denote $\langle u', v' \rangle$ in what follows.

We can now state the following proposition.

Proposition 6.2 *Let $\mathcal{M}$ be a submanifold of $\mathbb{R}^N$ with condition number $1/\tau$. Let $p, q \in \mathcal{M}$ be two points with geodesic distance given by $d_{\mathcal{M}}(p, q)$. Let ϕ be the angle between the tangent spaces T_p and T_q defined by* $\cos(\phi) = \min_{u \in T_p} \max_{v \in T_q} |\langle u, v \rangle|$. *Then* $\cos(\phi)$ *is greater than* $1 - (1/\tau) d_{\mathcal{M}}(p, q)$.

Consider two points $p, q \in \mathcal{M}$ connected by a geodesic curve $c(t) \in \mathcal{M}$. Let $c(t)$ be parametrized (proportional to arc length) so that $c(0) = p$, and $c(1) = q$.

Now let $v_p \in T_p$ be a tangent vector (unit norm) and let $v(t)$ be the parallel transport of this vector along the curve $c(t)$. Thus we have $v(0) = v_p$, $v(1) = v_q \in T_q$. Clearly, $\langle v(t), v(t) \rangle = 1$ for all t since v is parallel.

Notice that

$$\big\langle v(0), v(1) \big\rangle = \big\langle v(0), v(0) + w \big\rangle = 1 + \big\langle v(0), w \big\rangle, \tag{1}$$

where

$$w = \int_0^1 \left(\frac{dv}{dt} \right) dt. \tag{2}$$

Combining (1) and (2), we see

$$\cos(\theta) = \big|\big\langle v(0), v(1) \big\rangle\big| \geq 1 - \big|\big\langle v(0), w \big\rangle\big| \geq 1 - \|w\|, \tag{3}$$

where θ is the angle between the vectors $v(0)$ and $v(1)$. Since $v_p = v(0)$ was arbitrary, it is easy to check that $\cos(\phi) \geq \cos(\theta)$.

Now

$$\frac{dv}{dt} = \bar{\nabla}_{dc/dt} v(t),$$

where $\bar{\nabla}$ denotes the connection in Euclidean space. At the same time

$$\nabla_{dc/dt} v(t) = \big(\bar{\nabla}_{dc/dt} v(t)\big)^T,$$

where for any $r \in \mathcal{M}$ and $v \in \bar{T}_r$ (here $\bar{T}_r$ is the tangent space of $\mathbb{R}^N$ at r) we denote by $(v)^T$ the projection of v onto T_r (here T_r is the tangent space to $\mathcal{M}$ at r

viewed as an affine space with origin r). However, since $v(t)$ is parallel, we have that $\nabla_{dc/dt} v(t) = 0$. Therefore, $\bar{\nabla}_{dc/dt} v(t)$ is entirely in the space normal to $T_{c(t)}$, but the component of $\bar{\nabla}_{dc/dt} v(t)$ in the normal direction is precisely given by the second fundamental form. Hence, we have that

$$\frac{dv}{dt} = B\left(\frac{dc}{dt}, v(t)\right),$$

where B is a symmetric, bilinear form (the second fundamental form). Letting η be a unit norm vector in the direction dv/dt, i.e., $\eta = (1/\|dv/dt\|)(dv/dt)$, we see that

$$\left\|\frac{dv}{dt}\right\| = \left\langle \eta, \frac{dv}{dt} \right\rangle = \left\langle \eta, B\left(\frac{dc}{dt}, v(t)\right)\right\rangle = \left\langle \frac{dc}{dt}, L_n v(t)\right\rangle,$$

where L_n is a self-adjoint linear operator. By Proposition 6.1, the norm of L_η is bounded by $1/\tau$. Therefore, we have

$$\left\|\frac{dv}{dt}\right\| \leq \left\|\frac{dc}{dt}\right\| \|L_n v\| \leq \left\|\frac{dc}{dt}\right\| \|L_\eta\|,$$

and

$$\|w\| = \left\|\int_0^1 \frac{dv}{dt}\right\| \leq \int_0^1 \left\|\frac{dv}{dt}\right\| \leq \|L_n\| \int_0^1 \left\|\frac{dc}{dt}\right\| dt \leq \frac{1}{\tau} d_{\mathcal{M}}(p, q). \tag{4}$$

Combining (3) and (4), we get $\cos(\phi) \geq 1 - \frac{1}{\tau} d_{\mathcal{M}}(p, q)$.

We next show a relationship between the geodesic distance $d_{\mathcal{M}}(p, q)$ and the ambient distance $\|p - q\|_{\mathbb{R}^N}$ for any two points p and q on the submanifold $\mathcal{M}$.

Proposition 6.3 *Let $\mathcal{M}$ be a submanifold of $\mathbb{R}^N$ with condition number $1/\tau$. Let p and q be two points in $\mathcal{M}$ such that $\|p - q\|_{\mathbb{R}^N} = d$. Then for all $d \leq \tau/2$, the geodesic distance $d_{\mathcal{M}}(p, q)$ is bounded by*

$$d_{\mathcal{M}}(p, q) \leq \tau - \tau\sqrt{1 - \frac{2d}{\tau}}.$$

Consider two points $p, q \in \mathcal{M}$ and let $c(t)$ be a geodesic curve joining them such that $c(0) = p$ and $c(s) = q$. Let c be parametrized by arc length so that $\|\dot{c}(t)\| = 1$ for all t and $d_{\mathcal{M}}(p, q) = s$.

Noting that the tangent vector field $\dot{c}$ along the curve is parallel, we have $\ddot{c} = B(\dot{c}, \dot{c})$ and from Proposition 6.1 we see that for all t,

$$\|\ddot{c}\| = \left\|B(\dot{c}, \dot{c})\right\| \leq \frac{1}{\tau}.$$

The chord length between p and q is given by $\|c(s) - c(0)\|$ and we now relate this to the geodesic distance $d_{\mathcal{M}}(p, q)$. Observe that

$$c(s) - c(0) = \int_0^s \dot{c}(t)\, dt.$$

Now

$$\dot{c}(t) = \dot{c}(0) + \int_0^t \ddot{c}(r)\, dr.$$

Thus $\dot{c}(t) = \dot{c}(0) + u(t)$ where $u(t) = \int_0^t \ddot{c}(r)\, dr$. We see that

$$\|u(t)\| \leq \int_0^t \left\|\ddot{c}(r)\, dr\right\| \leq \frac{t}{\tau}.$$

Therefore,

$$\left\|c(s) - c(0)\right\| = \left\| \int_0^s \dot{c}(0)\, dt + \int_0^s u(t)\, dt \right\| \geq s\left\|\dot{c}(0)\right\| - \int_0^s \left\|u(t)\right\| dt \geq s - \int_0^s \frac{t}{\tau}\, dt.$$

Therefore we get

$$\left\|c(s) - c(0)\right\| = d \geq s - \frac{s^2}{2\tau}, \tag{5}$$

where d is the ambient distance between the points p and q while s is the geodesic distance between these same points. The inequality in (5) is satisfied only if $s \leq \tau - \tau\sqrt{1 - 2d/\tau}$ or $s \geq \tau + \tau\sqrt{1 - 2d/\tau}$. Since $s = 0$ when $d = 0$, we know that the second inequality does not apply. Therefore, from the first inequality, we have $s \leq \tau - \tau\sqrt{1 - \frac{2d}{\tau}}$.

7 Handling Noisy Data

In this section we show that if our data are noisy in the sense that they are drawn from a probability distribution that is concentrated around (rather than on) the manifold, the homology of the manifold can still be computed from noisy data.

7.1 The Model of Noise

Consider a probability measure μ concentrated around the manifold. We assume that μ satisfies the following two regularity conditions:

1. The support of μ ($\operatorname{supp}\mu$) is contained in the tubular neighborhood of radius r around $\mathcal{M}$. Thus $\operatorname{supp}\mu \subset \operatorname{Tub}_r(\mathcal{M})$.
2. For every $0 < s < r$, we have that

$$\inf_{p \in \mathcal{M}} \mu(B_s(p)) > k_s,$$

where k_s is a constant depending on s and independent of p.

In what follows we assume the data are drawn in an i.i.d. fashion according to a P that satisfies the above properties.

7.2 Main Topological Lemma: Sufficient Conditions

We proceed by constructing ϵ-balls centered on our data points. If these data are s-dense on the manifold, then the homology of the union of these balls will equal that of the manifold $\mathcal{M}$ *even if* the data are drawn from a noisy distribution. In order to see that this might be the case, we provide a simple argument. This argument works with non-optimal choices of ϵ and s and later sections enter into the considerations of choosing better values for these parameters and therefore providing more natural complexity estimates.

Let $\bar{x} = \{x_1, \ldots, x_n\}$ be a set of n points in the tubular neighborhood of radius r around $\mathcal{M}$. Let U be given by

$$U = \bigcup_{x \in \bar{x}} B_\epsilon(x).$$

Proposition 7.1 *If $\bar{x}$ is r-dense in $\mathcal{M}$, then $\mathcal{M}$ is a deformation retract of U for all $r < (\sqrt{9} - \sqrt{8})\tau$ and*

$$\epsilon \in \left(\frac{(r+\tau) - \sqrt{r^2 + \tau^2 - 6\tau r}}{2}, \frac{(r+\tau) + \sqrt{r^2 + \tau^2 - 6\tau r}}{2} \right).$$

Proof We show that for each $p \in \mathcal{M}$, it is the case that $\pi^{-1}(p)$ contracts to p. Consider a $v \in \pi^{-1}(p)$. Consider the line segment, $v\bar{p}$, joining v to p. We claim that this line segment is entirely contained in $\pi^{-1}(p)$. Clearly, if $v \in B_\epsilon(x)$ for some $x \in \bar{x} \cap B_\epsilon(p)$, this is immediate by the convexity of balls in Euclidean space. So we only need to consider the situation where $v \in B_\epsilon(x)$ for some $x \notin \bar{x} \cap B_\epsilon(p)$. So let $v \in B_\epsilon(q) \cap T_p^\perp$. Let

$$u = \arg \min_{x \in v\bar{p} \cap \bar{B_\epsilon(q)}} \|x - p\|.$$

As long as $u \in B_\epsilon(x)$ for some $x \in \bar{x} \cap B_\epsilon(p)$, we see that the line segment $u\bar{p}$ is contained in $\pi^{-1}(p)$ and therefore v contracts to p.

Since we choose $r < \epsilon$, we are guaranteed that there is an $x \in \bar{x} \cap B_r(p) \subset B_\epsilon(p)$. The worst-case picture is shown in Fig. 3. Following the symbols of the figure, as long as

$$\tau - A < \epsilon - r,$$

we have that v contracts to p. Thus we need

$$\big(\tau - (\epsilon - r)\big)^2 < A^2 = (\tau - r)^2 - \epsilon^2. \tag{6}$$

Expanding the squares, this reduces to

$$\epsilon^2 - \epsilon(\tau + r) + 2\tau r < 0.$$

This is a quadratic in ϵ and is satisfied for

$$\epsilon \in \left(\frac{(r+\tau) - \sqrt{r^2 + \tau^2 - 6\tau r}}{2}, \frac{(r+\tau) + \sqrt{r^2 + \tau^2 - 6\tau r}}{2} \right) \tag{7}$$

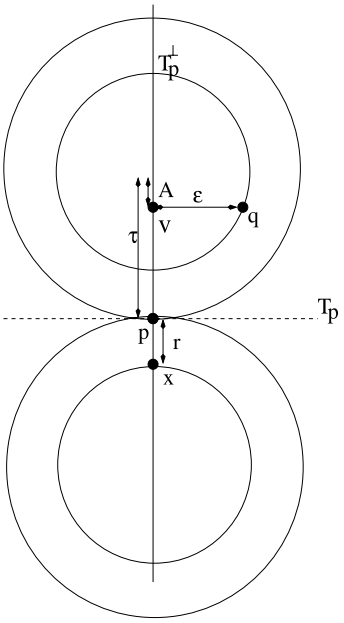


Fig. 3 A picture showing the worst case. As before, we draw the picture in the plane connecting points v, p, and q. T_p and $T_p^\perp$ are intersected with this plane in the picture and shown by the *dotted horizontal line* and *solid vertical line*, respectively. The *concentric circles* have the same center and are of radius τ and $\tau - r$, respectively, and follow our usual construction in earlier figures and arguments. All lengths are marked by *arrows*

provided

$$r^2 - 6\tau r + \tau^2 > 0.$$

This, in turn, is a quadratic in r and it is easy to check that it is satisfied as long as

$$r < \left(3 - 2\sqrt{2}\right)\tau = \left(\sqrt{9} - \sqrt{8}\right)\tau. \tag{8}$$

Thus we see that for r, ϵ satisfying (7) and (8), we have that v contracts to p. □

We now need to compute the probability of drawing a random $\bar{x}$ that is guaranteed to be r-dense. The following proposition is true.

Proposition 7.2 *Let $N_{r/2}$ be the $(r/2)$-covering number of the manifold. Let $p_1, \ldots, p_{N_{r/2}} \in \mathcal{M}$ be points on the manifold such that $B_{r/2}(p_i)$ realize an $(r/2)$-cover of the manifold. Let $\bar{x}$ be generated by i.i.d. draws according to a probability measure μ that satisfies the regularity properties described earlier. Then if $|\bar{x}| > (1/k_{r/2})(\log(N_{r/2}) + \log(1/\delta))$, with probability greater than $1 - \delta$, $\bar{x}$ will be r-dense in $\mathcal{M}$.*

Proof Take $A_i = B_{r/2}(p_i)$ and apply Lemma 5.1. By the conclusion of that lemma, we have that with high probability each of the A_i's is occupied by at least one $x \in \bar{x}$. Therefore it follows that for any $p \in \mathcal{M}$, there is at least one $x \in \bar{x}$ such that $\|p - x\| < r$. Thus with high probability $\bar{x}$ is r-dense on the manifold. □

Putting these together, our main conclusion is:

Theorem 7.1 *Let $N_{r/2}$ be the $(r/2)$-covering number of the submanifold $\mathcal{M}$ of $\mathbb{R}^N$. Let $\bar{x}$ be generated by i.i.d. draws according to a probability measure μ that satisfies the regularity properties described earlier. Let $U = \bigcup_{x\in\bar{x}} B_\epsilon(x)$. Then if $|\bar{x}| > (1/k_{r/2})(\log(N_{r/2}) + \log(1/\delta))$, with probability greater than $1-\delta$, $\mathcal{M}$ is a deformation retract of U as long as* (i) $r < (\sqrt{9}-\sqrt{8})\tau$ *and* (ii)

$$\epsilon \in \left(\frac{(r+\tau) - \sqrt{r^2+\tau^2-6\tau r}}{2}, \frac{(r+\tau) + \sqrt{r^2+\tau^2-6\tau r}}{2} \right).$$

7.3 Main Topological Lemma—General Considerations

In general, we may demand points that are s-dense. Putting ϵ-balls around these points we construct U in the usual way. The condition number τ and the noise bound r are additional parameters that are outside our control and determined externally. We now ask what is the feasible space (s, ϵ, r, τ) that will guarantee that U is homotopy equivalent to $\mathcal{M}$?

Following our usual logic, we see that the worst-case situation is given by Fig. 4. An arbitrary $v \in B_\epsilon(q) \cap T_p^\perp \cap B_\tau(p)$ will contract to p if

$$B_\epsilon(q) \cap B_\epsilon(x) \cap v\bar{p} \neq \phi.$$

This is the same as requiring

$$(\tau - w)^2 < (\tau - r)^2 - \epsilon^2. \tag{9}$$

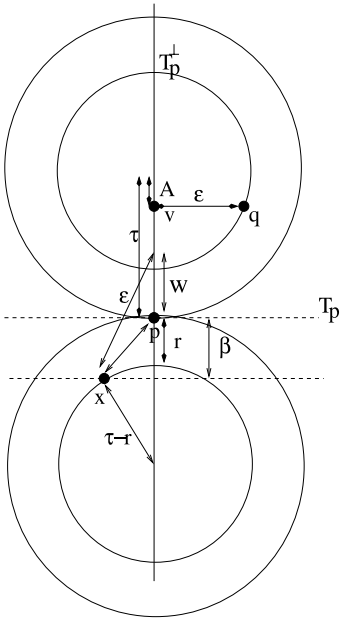


Fig. 4 A picture showing the worst case. As before, we draw the picture in the plane connecting points v, p, and q. T_p and $T_p^\perp$ are intersected with this plane in the picture and shown by the *dotted horizontal line* and *solid vertical line*, respectively. The *concentric circles* have the same center and are of radius τ and $\tau - r$, respectively, and follow our usual construction in earlier figures and arguments. All lengths are marked by *arrows*

Additionally, we have the following equations that need to be satisfied (following Fig. 4):

$$(\tau - r)^2 - (\tau - \beta)^2 = s^2 - \beta^2, \tag{10}$$

$$s^2 - \beta^2 + (\beta + w)^2 = \epsilon^2. \tag{11}$$

If one eliminates w and β from the above equations, one will get a single inequality relating s, ϵ, τ, r that describes for each τ, r the feasible set of possible choices of s, ϵ that are sufficient to guarantee homotopy equivalence. Let us see how our earlier theorems follow from particular choices of this general set of equations.

7.3.1 The Case when $s = r$

We have already examined the case when the points $\bar{x}$ are chosen to be r-dense in $\mathcal{M}$. Putting $s = r$ in (9)–(11), we see the following:

From (10), we have (for $s = r$)

$$(\tau - r)^2 - (\tau - \beta)^2 = r^2 - \beta^2.$$

This simplifies to give $\beta = r$.

Putting $\beta = r$ and $s = r$ in (11), we get

$$r^2 - r^2 + (r + w)^2 = \epsilon^2,$$

giving us $w = \epsilon - r$.

Finally, putting $w = \epsilon - r$ in inequality (9), we get

$$\big(\tau - (\epsilon - r)\big)^2 < (\tau - r)^2 - \epsilon^2,$$

which is the same as inequality (6) whose solution was examined in the previous section.

7.3.2 The Case when $r = 0$

We can recover our main theorem for the noise-free case by considering the case $r = 0$. We proceed to do this now.

The fundamental inequality of (9) gives us (for $r = 0$)

$$(\tau - w)^2 < \tau^2 - \epsilon^2.$$

This is the same as requiring

$$w^2 - 2\tau + \epsilon^2 < 0.$$

Using standard analysis for quadratic functions, we see that the following condition is required:

$$w > \tau - \sqrt{\tau^2 - \epsilon^2}. \tag{12}$$

We can eliminate w using (10) and (11). Thus, from (10), we get $\beta = s^2/2\tau$ and substituting in (11), we get a quadratic equation in w whose positive solution is given by $w = -s^2/2\tau + \sqrt{s^4/4\tau^2 + (\epsilon^2 - s^2)}$. This gives rise to the following condition:

$$-\frac{s^2}{2\tau} + \sqrt{\frac{s^4}{4\tau^2} + \left(\epsilon^2 - s^2\right)} > \tau - \sqrt{\tau^2 - \epsilon^2}. \tag{13}$$

Inequality (13) gives the feasible region for s and ϵ for the homotopy equivalence of U and $\mathcal{M}$. Let us consider the special case when $s = \epsilon/2$—a choice we made in Sect. 3 without any attention to optimality. Putting in this value, after several simplifying steps, one obtains that

$$\epsilon^4 + 51\epsilon^2\tau^2 - 48\tau^4 < 0. \tag{14}$$

This is satisfied for all $0 < \epsilon^2 < 0.9244\tau^2$ or $0 < \epsilon < 0.96\tau$.

Remark 1 Note that in our original proof of our main noise free theorem (Theorem 3.1), the deformation retract argument of Sect. 3 passes through the construction of $st(p)$ and shows contraction of $\pi^{-1}(p)$ by equating it with $st(p)$. This condition is stronger than we require. Here we see that the condition $B_\epsilon(q) \cap B_\epsilon(x) \cap \bar{vp} \neq \phi$ is sufficient. This latter condition is weaker and therefore gives us a slightly stronger version of Theorem 3.1 in the sense that it holds for a larger range of ϵ.

Remark 2 If we assume that τ, r are beyond our control, the sample complexity depends entirely upon s. Therefore if we wish to proceed by drawing the fewest number of examples, then it is necessary to maximize s subject to the condition of (13).

Remark 3 The total complexity of finding the homology depends both upon s and ϵ in a more complicated way. The size of $\bar{x}$ depends entirely upon s and nothing else. However, the number of k-tuples to consider in the simplicial complex depends both upon the size of $\bar{x}$ as well as ϵ because ϵ determines how many balls will have non-empty intersections. We leave this more nuanced complexity analysis for future consideration.

References

1. Amenta, N., Bern, M.: Surface reconstruction by Voronoi filtering. Discrete Comput. Geom. **22**, 481–504 (1999)
2. Amenta, N., Choi, S., Dey, T.K., Leekha, N.: A simple algorithm for homeomorphic surface reconstruction. Int. J. Comput. Geom. Appl. **12**, 125–141 (2002)
3. Belkin, M., Niyogi, P.: Semisupervised learning on Riemannian manifolds. Mach. Learn. **56**, 209–239 (2004)
4. Bjorner, A.: Topological methods. In: Graham, R., Grotschel, M., Lovasz, L. (eds.) Handbook of Combinatorics, pp. 1819–1872. North-Holland, Amsterdam (1995)
5. Chazal, F., Lieutier, A.: Weak feature size and persistent homology: computing homology of solids in $\mathbb{R}^n$ from noisy data samples. Preprint
6. Cheng, S.W., Dey, T.K., Ramos, E.A.: Manifold reconstruction from point samples. In: Proceedings of ACM-SIAM Symposium on Discrete Algorithms, pp. 1018–1027 (2005)

7. Cohen-Steiner, D., Edelsbrunner, H., Harer, J.: Stability of persistence diagrams. In: Proceedings of the 21st Symposium on Computational Geometry, pp. 263–271 (2005)
8. Dey, T.K., Edelsbrunner, H., Guha, S.: Computational topology. In: Chazelle, B., Goodman, J.E., Pollack, R. (eds.) Advances in Discrete and Computational Geometry, Contemporary Mathematics, vol. 223, pp. 109–143. AMS, Providence (1999)
9. Do Carmo, M.P.: Riemannian Geometry. Birkhäuser, Basel (1992)
10. Donoho, D., Grimes, C.: Hessian eigenmaps: new locally-linear embedding techniques for high-dimensional data. Preprint. Department of Statistics, Stanford University (2003)
11. Edelsbrunner, H., Mucke, E.P.: Three-dimensional alpha shapes. ACM Trans. Graph. **13**, 43–72 (1994)
12. Fischer, K., Gaertner, B., Kutz, M.: Fast smallest-enclosing-ball computation in high dimensions. In: Proceedings of the 11th Annual European Symposium on Algorithms (ESA), pp. 630–641 (2003)
13. Friedman, J.: Computing Betti numbers via combinatorial laplacians. Algorithmica **21**, 331–346 (1998)
14. Kaczynski, T., Mischaikow, K., Mrozek, M.: Computational Homology. Springer, New York (2004)
15. Munkres, J.: Elements of Algebraic Topology. Addison-Wesley, Menlo Park (1984)
16. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**, 2323–2326 (2000)
17. Tenenbaum, J.B., De Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**, 2319–2323 (2000)
18. Valiant, L.G.: A theory of the learnable. Commun. ACM **27**(11), 1134–1142 (1984)
19. Website for smallest enclosing ball algorithm. http://www2.inf.ethz.ch/personal/gaertner/miniball.html
20. Zomorodian, A., Carlsson, G.: Computing persistent homology. Discrete Comput. Geom. **33**, 249–274 (2005)

Odd Crossing Number and Crossing Number Are Not the Same

Michael J. Pelsmajer · Marcus Schaefer · Daniel Štefankovič

Abstract The *crossing number* of a graph is the minimum number of edge intersections in a plane drawing of a graph, where each intersection is counted separately. If instead we count the number of pairs of edges that intersect an odd number of times, we obtain the *odd crossing number*. We show that there is a graph for which these two concepts differ, answering a well-known open question on crossing numbers. To derive the result we study drawings of maps (graphs with rotation systems).

1 A Confusion of Crossing Numbers

Intuitively, the crossing number of a graph is the smallest number of edge crossings in any plane drawing of the graph. As it turns out, this definition leaves room for interpretation, depending on how we answer the questions: what is a drawing, what is a crossing, and how do we count crossings? The papers by Pach and Tóth [7] and Székely [9] discuss the historical development of various interpretations and definitions—often implicit—of the crossing number concept.

A *drawing* D of a graph G is a mapping of the vertices and edges of G to the Euclidean plane, associating a distinct point with each vertex, and a simple plane curve with each edge so that the ends of an edge map to the endpoints of the corresponding curve. For simplicity, we also require that

M.J. Pelsmajer (✉)
Department of Applied Mathematics, Illinois Institute of Technology, Chicago, IL 60616, USA
e-mail: pelsmajer@iit.edu

M. Schaefer
Department of Computer Science, DePaul University, Chicago, IL 60604, USA
e-mail: mschaefer@cs.depaul.edu

D. Štefankovič
Computer Science Department, University of Rochester, Rochester, NY 14627-0226, USA
e-mail: stefanko@cs.rochester.edu

- A curve does not contain any endpoints of other curves in its interior
- Two curves do not touch (that is, intersect without crossing), and
- No more than two curves intersect in a point (other than at a shared endpoint)

In such a drawing the intersection of the interiors of two curves is called a *crossing*. Note that by the restrictions we placed on a drawing, crossings do not involve endpoints, and at most two curves can intersect in a crossing. We often identify a drawing with the graph it represents. For a drawing D of a graph G in the plane we define

- $\mathrm{cr}(D)$ - the total number of crossings in D
- $\mathrm{pcr}(D)$ - the number of pairs of edges which cross at least once; and
- $\mathrm{ocr}(D)$ - the number of pairs of edges which cross an odd number of times

Remark 1 For any drawing D, we have $\mathrm{ocr}(D) \leq \mathrm{pcr}(D) \leq \mathrm{cr}(D)$.

We let $\mathrm{cr}(G) = \min \mathrm{cr}(D)$, where the minimum is taken over all drawings D of G in the plane. We define $\mathrm{ocr}(G)$ and $\mathrm{pcr}(G)$ analogously.

Remark 2 For any graph G, we have $\mathrm{ocr}(G) \leq \mathrm{pcr}(G) \leq \mathrm{cr}(G)$.

The question (first asked by Pach and Tóth [7]) is whether the inequalities are actually equalities.[1] Pach [6] called this "perhaps the most exciting open problem in the area." The only evidence for equality is an old theorem by Chojnacki, which was later rediscovered by Tutte—and the absence of any counterexamples.

Theorem 1.1 (Chojnacki [4], Tutte [10]) *If* $\mathrm{ocr}(G) = 0$, *then* $\mathrm{cr}(G) = 0$.[2]

In this paper we will construct a simple example of a graph with $\mathrm{ocr}(G) < \mathrm{pcr}(G) = \mathrm{cr}(G)$. We derive this example from studying what we call weighted maps on the annulus. Section 2 introduces the notion of weighted maps on arbitrary surfaces and gives a counterexample to $\mathrm{ocr}(M) = \mathrm{pcr}(M)$ for maps on the annulus. In Section 3 we continue the study of crossing numbers for weighted maps, proving in particular that $\mathrm{cr}(M) \leq c_n \cdot \mathrm{ocr}(M)$ for maps on a plane with n holes. One of the difficulties in dealing with the crossing number is that it is **NP**-complete [2]. In Section 4 we show that the crossing number can be computed in polynomial time for maps on the annulus. Finally, in Section 5 we show how to translate the map counterexample from Section 2 into an infinite family of simple graphs for which $\mathrm{ocr}(G) < \mathrm{pcr}(G)$.

2 Map Crossing Numbers

A *weighted map* M is a surface S and a set $P = \{(a_1, b_1), \ldots, (a_m, b_m)\}$ of pairs of distinct points on ∂S with positive weights $w_1, \ldots, w_m$. A *realization* R of the map

[1] Doug West lists the problem on his page of open problems in graph theory [12]. Dan Archdeacon even conjectured that equality holds [1].

[2] In fact they proved something stronger, namely that in any drawing of a non-planar graph there are two non-adjacent edges that cross an odd number of times. Also see [8].

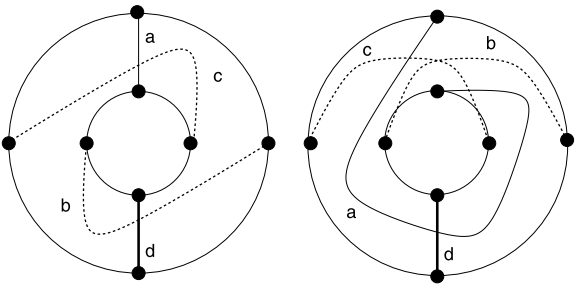


Fig. 1 Optimal drawings: pcr and cr (*above left*), ocr (*above right*)

$M = (S, P)$ is a set of m properly embedded arcs $\gamma_1, \ldots, \gamma_m$ in S where γ_i connects a_i and b_i.[3]

Let

$$\mathrm{cr}(R) = \sum_{1 \le k < \ell \le m} \iota(\gamma_k, \gamma_\ell) w_k w_\ell,$$

$$\mathrm{pcr}(R) = \sum_{1 \le k < \ell \le m} [\iota(\gamma_k, \gamma_\ell) > 0] w_k w_\ell,$$

$$\mathrm{ocr}(R) = \sum_{1 \le k < \ell \le m} [\iota(\gamma_k, \gamma_\ell) \equiv 1 \pmod 2] w_k w_\ell,$$

where $\iota(\gamma, \gamma')$ is the geometric intersection number of γ and γ' and $[x]$ is 1 if the condition x is true, and 0 otherwise. We define $\mathrm{cr}(M) = \min \mathrm{cr}(R)$, where the minimum is taken over all realizations R of M. We define $\mathrm{pcr}(M)$ and $\mathrm{ocr}(M)$ analogously.

Remark 3 For every map M, $\mathrm{ocr}(M) \le \mathrm{pcr}(M) \le \mathrm{cr}(M)$.

Conjecture 1 *For every map* M, $\mathrm{cr}(M) = \mathrm{pcr}(M)$.

Lemma 2.1 *If Conjecture* 1 *is true, then* $\mathrm{cr}(G) = \mathrm{pcr}(G)$ *for every* graph G.

Proof Let D be a drawing of G with minimal pair crossing number. Drill small holes at the vertices. We obtain a drawing R of a weighted map M. If Conjecture 1 is true, there exists a drawing of M with the same crossing number. Collapse the holes to vertices to obtain a drawing D' of G with $\mathrm{cr}(D') \le \mathrm{pcr}(G)$. □

However, we show below that we can separate the odd crossing number from the crossing number for weighted maps, even in the annulus (a disk with a hole).

When analyzing crossing numbers of drawings on the annulus, we describe curves with respect to an initial drawing of the curve and a number of *Dehn twists*. Consider, for example, the four curves in the left part of Figure 1. Comparing them to the

[3] If we take a realization R of a map M, and contract each boundary component to a vertex, we obtain a drawing of a graph with a given rotation system [3]. For our purposes, maps are a more visual way to look at graphs with a rotation system.

corresponding curves in the right part, we see that the curves labeled c and d have not changed, but the curves labeled a and b have each undergone a single clockwise twist.

Two curves are *isotopic rel boundary* if they can be obtained from each other by a continuous deformation which does not move the boundary ∂M. Isotopy rel boundary is an equivalence relation, its equivalence classes are called *isotopy classes*. An *isotopy class* on the annulus is determined by a properly embedded arc connecting the endpoints, together with the number of twists performed.

Lemma 2.2 *Let $a \le b \le c \le d$ be such that $a + c \ge d$. For the weighted map M in Figure* 1 *we have* $\mathrm{cr}(M) = \mathrm{pcr}(M) = ac + bd$ *and* $\mathrm{ocr}(M) = bc + ad$.

Proof The upper bounds follow from the drawings in Figure 1, the left drawing for crossing and pair crossing number, the right drawing for odd crossing number.

Claim $\mathrm{pcr}(M) \ge ac + bd$.

Proof of the Claim Let R be a drawing of M minimizing $\mathrm{pcr}(R)$. We can apply twists so that the thick edge d is drawn as in the left part of Figure 1. Let α, β, γ be the number of clockwise twists applied to the ends of arcs a, b, c on the inner boundary to obtain the drawing R, where $\alpha = \beta = \gamma = 0$ corresponds to the drawing shown in the left part of Figure 1. Then,

$$\begin{aligned}\mathrm{pcr}(R) = {} & cd[\gamma \ne 0] + bd[\beta \ne -1] + ad[\alpha \ne 0] + bc[\beta \ne \gamma] \\ & + ab[\alpha \ne \beta] + ac[\alpha \ne \gamma + 1]. \end{aligned} \tag{1}$$

If $\gamma \ne 0$, then $\mathrm{pcr}(R) \ge cd + ab$ because at least one of the last five conditions in (1) must be true; the last five terms contribute at least ab (since $d \ge c \ge b \ge a$), and the first term contributes cd. Since $d(c - b) \ge a(c - b)$, $cd + ab \ge ac + bd$, and the claim is proved in the case that $\gamma \ne 0$.

Now assume that $\gamma = 0$. Equation (1) becomes

$$\mathrm{pcr}(R) = bd[\beta \ne -1] + bc[\beta \ne 0] + ad[\alpha \ne 0] + ac[\alpha \ne 1] + ab[\alpha \ne \beta]. \tag{2}$$

If $\beta \ne -1$, then $\mathrm{pcr}(R) \ge bd + ac$ because either $\alpha \ne 0$ or $\alpha \ne 1$. Since $bd + ac \ge bc + ad$, the claim is proved in the case that $\beta \ne -1$.

This leaves us with the case that $\beta = -1$. Equation (2) becomes

$$\mathrm{pcr}(R) = bc + ad[\alpha \ne 0] + ac[\alpha \ne 1] + ab[\alpha \ne -1]. \tag{3}$$

The right-hand side of Equation (3) is minimized for $\alpha = 0$. In this case $\mathrm{pcr}(R) = bc + ac + ab \ge ac + bd$ because we assume that $a + c \ge d$. □

Claim $\mathrm{ocr}(M) \ge bc + ad$.

Proof of the Claim Let R be a drawing of M minimizing $\mathrm{ocr}(R)$. Let α, β, γ be as in the previous claim. We have

$$\operatorname{ocr}(R) = cd[\gamma]_2 + bd[\beta+1]_2 + ad[\alpha]_2 + bc[\beta+\gamma]_2 + ab[\alpha+\beta]_2 + ac[\alpha+\gamma+1]_2, \tag{4}$$

where $[x]_2$ is 0 if $x \equiv 0 \pmod 2$, and 1 otherwise.

If $\beta \not\equiv \gamma \pmod 2$, then the claim clearly follows unless $\gamma = 0$, $\beta = 1$, and $\alpha = 0$ (all modulo 2). In that case $\operatorname{ocr}(R) \geq bc + ab + ac \geq bc + ad$. Hence, the claim is proved if $\beta \not\equiv \gamma \pmod 2$.

Assume then that $\beta \equiv \gamma \pmod 2$. Equation (4) becomes

$$\operatorname{ocr}(R) = cd[\beta]_2 + bd[\beta+1]_2 + ad[\alpha]_2 + ab[\alpha+\beta]_2 + ac[\alpha+\beta+1]_2. \tag{5}$$

If $\alpha \equiv 1 \pmod 2$, then the claim clearly follows because either cd or bd contributes to the ocr. Thus we can assume $\alpha \equiv 0 \pmod 2$. Equation (5) becomes

$$\operatorname{ocr}(R) = (cd + ab)[\beta]_2 + (bd + ac)[\beta+1]_2. \tag{6}$$

For both $\beta \equiv 0 \pmod 2$ and $\beta \equiv 1 \pmod 2$ we get $\operatorname{ocr}(R) \geq bc + ad$. □

We get a separation of pcr and ocr for maps with small integral weights.

Corollary 2.3 *There is a weighted map M on the annulus with edges of weight $a = 1$, $b = c = 3$, and $d = 4$ for which* $\operatorname{cr}(M) = \operatorname{pcr}(M) = 15$ *and* $\operatorname{ocr}(M) = 13$.

Optimizing the gap over the reals yields $b = c = 1$, $a = (\sqrt{3}-1)/2$, and $d = 1+a$, giving us the following separation of $\operatorname{pcr}(M)$ and $\operatorname{ocr}(M)$.

Corollary 2.4 *There exists a weighted map M on the annulus with* $\operatorname{ocr}(M) \leq \sqrt{3}/2 \operatorname{pcr}(M)$.

Conjecture 2 *For every weighted map M on the annulus,* $\operatorname{ocr}(M) \geq \frac{\sqrt{3}}{2} \operatorname{pcr}(M)$.

3 Upper Bounds on Crossing Numbers

In Section 5 we will transform the separation of ocr and pcr on maps into a separation on graphs. In particular, we will show that for every $\varepsilon > 0$ there is a graph G so that

$$\operatorname{ocr}(G) < \left(\sqrt{3}/2 + \varepsilon\right) \operatorname{cr}(G).$$

The gap cannot be arbitrarily large, as Pach and Tóth showed.

Theorem 3.1 (Pach and Tóth [7]) *Let G be a graph. Then* $\operatorname{cr}(G) \leq 2(\operatorname{ocr}(G))^2$.[4]

This result suggests the question whether the linear separation can be improved. We do not believe this to be possible:

[4] Better upper bounds on $\operatorname{cr}(G)$ in terms of $\operatorname{pcr}(G)$ are known [5, 11].

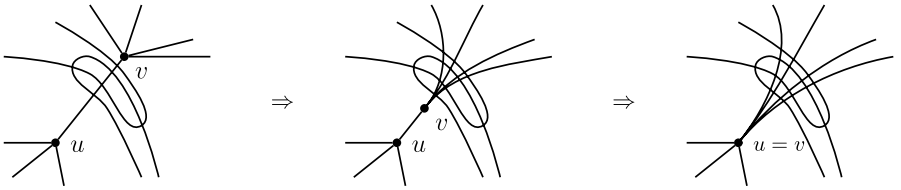


Fig. 2 Pulling an endpoint (*left*) and contracting the edge (*right*)

Conjecture 3 *There is a* $c > 0$ *so that* $\mathrm{cr}(G) < c \cdot \mathrm{ocr}(G)$.

In this section, we will show that our approach of comparing the different crossing numbers for maps with a fixed number of holes will not lead to a super-linear separation. Namely, for a (weighted) map M on a plane with n holes, we always have

$$\mathrm{cr}(M) \leq \mathrm{ocr}(M)\binom{n+4}{4}/5, \tag{7}$$

with strict inequality if $n > 1$. It follows that for fixed n, there *is* only a constant factor separating $\mathrm{cr}(M)$ and $\mathrm{ocr}(M)$. And only fixed, small n are computationally feasible in analyzing potential counterexamples.

Observe that as a special case of Equation (7), if M is a (weighted) map on the annulus ($n = 2$) we get that $\mathrm{cr}(M) < 3\,\mathrm{ocr}(M)$, which comes reasonably close to the $\sqrt{3}/2$ lower bound from the previous section.

Before proving Equation (7) in full generality, we first consider the case of unit weights.

For this section only, we will switch our point of view from maps as curves between holes on a plane to maps as graphs with a rotation system; that is, we contract each hole to a vertex, and record the order, in which the curves (edges) leave the vertex. Our basic operation will be the contraction of an edge by pulling one of its endpoints along the edge, until it coincides with the other endpoint (the rotations of the vertices merge). Figure 2 illustrates pulling v towards u along uv.[5]

Consider a drawing of G with the minimum number of odd pairs (edge pairs that cross an odd number of times), $\mathrm{ocr}(G)$. We want to contract edges without creating too many new odd pairs. For each edge e, let o_e be the number of edges that cross e an odd number of times. Then $\sum_{e \in E(G)} o_e = 2\,\mathrm{ocr}(G)$, and since each edge is incident to exactly two vertices,

$$\sum_{v \in V(G)} \sum_{e \ni v} o_e = 4\,\mathrm{ocr}(G).$$

Applying the pigeonhole principle twice, there must be a vertex $v \in V(G)$ with $\sum_{e \ni v} o_e \leq 4\,\mathrm{ocr}(G)/n$, and there is a non-loop edge e incident to v with $o_e \leq 4\,\mathrm{ocr}(G)/(n \cdot d^*(v))$ (where $d^*(v)$ counts the number of non-loop edges incident to v). Contracting v to its neighbor along e creates at most $o_e(d^*(v) - 1) <$

[5]The illustration is taken from [8], where we investigate some other uses of this operation for graph drawings.

$4\,\mathrm{ocr}(G)/n$ odd pairs (only edges intersecting e oddly will lead to odd intersections, and the parity of intersection along loops with endpoint v does not change; self-intersections can be removed). Repeating this operation $n-1$ times, we transform G into a bouquet of loops at a single vertex with at most $\mathrm{ocr}(G)\prod_{i=0}^{n-2}(1+4/(n-i))$ odd pairs (strictly less if $n>1$). Without changing the rotation of the vertex, we can redraw all loops so that each odd pair intersects exactly once, and other pairs do not cross at all. We can then undo the contractions of the edges in reverse order without creating any new crossings. This yields a drawing of G with the original vertex rotations with at most $\mathrm{ocr}(G)\prod_{i=0}^{n-2}(1+4/(n-i))$ crossings. Since the product term equals $\binom{n+4}{4}/5$, we have shown that $\mathrm{cr}(G)\le \mathrm{ocr}(G)\binom{n+4}{4}/5$ (strict inequality for $n>1$), as desired.

This argument proves Equation (7) for maps with unit weights. The next step is to extend this lemma to maps with arbitrary weights.

Consider two curves γ_1,γ_2 whose endpoints are adjacent and in the same order. In a drawing minimizing one of the crossing numbers we can always assume that the two curves are routed in parallel, following the curve that minimizes the total number of intersections with all curves other than γ_1 and γ_2. The same argument holds for a block of curves with adjacent endpoints in the same order. This allows us to claim Equation (7) for maps with integer weights: a curve with integral weight w is replaced by w parallel duplicates of unit weight.

If we scale all the weights in a map M by a factor α, all the crossing numbers will change by a factor of α^2. Hence, the case of rational weights can be reduced to integer weights. Finally, we observe that if we consider any of the crossing numbers as a function of the weights of M, this function is continuous: This is obvious for a fixed drawing of M, so it remains true if we minimize over a finite set of drawings of M. The maximum difference in the number of twists in an optimal drawing is bounded by a function of the crossing number; and thus it suffices to consider a finite set of drawings of M. We have shown:

Theorem 3.2 $\mathrm{cr}(M)\le \mathrm{ocr}(M)\binom{n+4}{4}/5$ *for weighted maps M on the plane with n holes.*

4 Computing Crossing Numbers on the Annulus

Let M be a map on the annulus. We explained earlier that as far as crossing numbers are concerned, we can describe a curve in the realization of M by a properly embedded arc γ_{ab} connecting endpoints a and b on the inner and outer boundary of the annulus, and an integer $k\in\mathbb{Z}$, counting the number of twists applied to the curve γ_{ab}. Our goal is to compute the number of intersections between two arcs after applying a number of twists to each one of them. Since twists can be positive and negative and cancel each other out, we need to count crossings more carefully. Let us orient all arcs from the inner boundary to the outer boundary. Traveling along an arc α, a crossing with β counts as $+1$ if β crosses from right to left, and as -1 if it crosses from left to right. Summing up these numbers over all crossings for two arcs α and β yields

$\hat{\iota}(\alpha, \beta)$, the *algebraic crossing number* of α and β. Tutte [10] introduced the notion

$$\operatorname{acr}(G) = \min_{D} \sum_{\{e,f\} \in \binom{E}{2}} |\hat{\iota}(\gamma_e, \gamma_f)|,$$

the *algebraic crossing number* of a graph, a notion that apparently has not drawn any attention since.

Let $D^k(\gamma)$ denote the result of adding k twists to the curve γ. For two curves α and β connecting the inner and outer boundary we have:

$$\hat{\iota}\big(D^k(\alpha), D^\ell(\beta)\big) = k - \ell + \hat{\iota}(\alpha, \beta). \tag{8}$$

Note that $\iota(\alpha, \beta) = |\hat{\iota}(\alpha, \beta)|$ for any two curves α, β on the annulus.

Let π be a permutation of $[n]$. A map M_π corresponding to π is constructed as follows. Choose $n+1$ points on each of the two boundaries and number them $0, 1, \ldots, n$ in the clockwise order. Let a_i be the vertex numbered i on the outer boundary and b_i be the vertex numbered π_i on the inner boundary, $i = 1, \ldots, n$. We ask a_i to be connected to b_i in M_π.

We will encode a drawing R of M_π by a sequence of n integers $x_1, \ldots, x_n$ as follows. Fix a curve β connecting the a_0 and b_0 and choose γ_i so that $\iota(\beta, \gamma_i) = 0$ (for all i). We will connect a_i, b_i with the arc $D^{x_i}(\gamma_i)$ in R. Note that for $i < j$, $\hat{\iota}(\gamma_i, \gamma_j) = [\pi_i > \pi_j]$ and hence

$$\hat{\iota}(D^{x_i}(\gamma_i), D^{x_j}(\gamma_j)) = x_i - x_j + [\pi_i > \pi_j].$$

We have

$$\operatorname{acr}(M_\pi) = \operatorname{cr}(M_\pi) = \min\left\{\sum_{i<j} |x_i - x_j + [\pi_i > \pi_j]| w_i w_j : x_i \in \mathbb{Z}, i \in [n]\right\}, \tag{9}$$

$$\operatorname{pcr}(M_\pi) = \min\left\{\sum_{i<j} [x_i - x_j + [\pi_i > \pi_j] \neq 0] w_i w_j : x_i \in \mathbb{Z}, i \in [n]\right\}, \tag{10}$$

and

$$\operatorname{ocr}(M_\pi) = \min\left\{\sum_{i<j} [x_i - x_j + [\pi_i > \pi_j] \not\equiv 0 \pmod 2] w_i w_j : x_i \in \mathbb{Z}, i \in [n]\right\}. \tag{11}$$

Consider the relaxation of the integer program for $\operatorname{cr}(M_\pi)$:

$$\operatorname{cr}'(M_\pi) = \min\left\{\sum_{i<j} |x_i - x_j + [\pi_i > \pi_j]| w_i w_j : x_i \in \mathbb{R}, i \in [n]\right\}. \tag{12}$$

Since (12) is a relaxation of (9), we have $\operatorname{cr}'(M_\pi) \leq \operatorname{cr}(M_\pi)$. The following lemma shows that $\operatorname{cr}'(M_\pi) = \operatorname{cr}(M_\pi)$.

Lemma 4.1 *Let n be a positive integer. Let $b_{ij} \in \mathbb{Z}$ and let $a_{ij} \in \mathbb{R}$ be non-negative, $1 \le i < j \le n$. Then*

$$\min\left\{\sum_{i<j} a_{ij}|x_i - x_j + b_{ij}| : x_i \in \mathbb{R}, i \in [n]\right\}$$

has an optimal solution with $x_i \in \mathbb{Z}$, $i \in [n]$.

Proof Let $\overline{x}^*$ be an optimal solution which satisfies the maximum number of $x_i - x_j + b_{ij} = 0$, $1 \le i < j \le n$. Without loss of generality, we can assume $x_1^* = 0$. Let G be a graph on the vertex set $\{1, \ldots, n\}$ with an edge between vertices i, j if $x_i^* - x_j^* + b_{ij} = 0$. Note that if i, j are connected by an edge and one of x_i^*, x_j^* is an integer, then both x_i^* and x_j^* are integers. It is then enough to show that G is connected.

Suppose that G is not connected. There exists a non-empty $A \subsetneq V(G)$ so that there are no edges between A and $V(G) - A$. Let χ_A be the characteristic vector of the set A, that is, $(\chi_A)_i = [i \in A]$. Let $f(\lambda)$ be the value of the objective function on $\overline{x} = \overline{x}^* + \lambda \cdot \chi_A$. Let I be the interval on which the signs of the $x_i - x_j + b_{ij}$, $1 \le i < j \le n$ are the same as for $\overline{x}^*$. Then I is not the entire line (otherwise G would be connected). Since f is linear on I, f is optimal at $\lambda = 0$, and I contains a neighborhood of 0, it must be that f is constant on I. Choosing $x = x^* + \lambda\chi_A$ for λ an endpoint of I gives an optimal solution satisfying more $x_i - x_j + b_{ij} = 0$, $1 \le i < j \le n$, a contradiction. □

Theorem 4.2 *The crossing number of maps on the annulus can be computed in polynomial time.*

Proof Note that $\mathrm{cr}'(M_\pi)$ is computed by the following linear program L_π:

$$\begin{aligned} \min \quad & \sum_{i<j} y_{ij} w_i w_j, \\ & y_{ij} \ge x_i - x_j + [\pi_i > \pi_j], \quad 1 \le i < j \le n, \\ & y_{ij} \ge -x_i + x_j - [\pi_i > \pi_j], \quad 1 \le i < j \le n. \end{aligned}$$

□

Question 1 *Let M be a map on the annulus. Can* $\mathrm{ocr}(M)$ *be computed in polynomial time?*

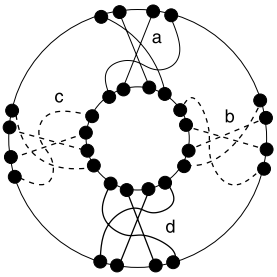
We conjectured earlier that crossing number and odd crossing number agree on maps. A more moderate goal would be to establish the following conjecture.

Conjecture 4 *For any map M on the annulus* $\mathrm{cr}(M) = \mathrm{pcr}(M)$.

5 Separating Crossing Numbers of Graphs

We modify the map from Lemma 2.2 to obtain a graph G separating $\mathrm{ocr}(G)$ and $\mathrm{pcr}(G)$. The graph G will have integral weights on edges. From G we can get an

Fig. 3 The inside flipped



unweighted graph G' with $\mathrm{ocr}(G') = \mathrm{ocr}(G)$ and $\mathrm{pcr}(G') = \mathrm{pcr}(G)$ by replacing an edge of weight w by w parallel edges of weight 1 (this does not change any of the crossing numbers). If needed we can get rid of parallel edges by subdividing edges, which does not change any of the crossing numbers.

We start with the map M from Lemma 2.2 with the following integral weights:

$$a = \left\lfloor \frac{\sqrt{3}-1}{2} m \right\rfloor, \qquad b = c = m, \qquad d = \left\lfloor \frac{\sqrt{3}+1}{2} m \right\rfloor,$$

where $m \in \mathbb{N}$ will be chosen later.

We replace each pair (a_i, b_i) of M by w_i pairs $(a_{i,1}, b_{i,1}), \ldots, (a_{i,w_i}, b_{i,w_i})$ where the $a_{i,j}$ ($b_{i,j}$) occur on ∂S in clockwise order in a small interval around of a_i (b_i). As before, we can argue that all the curves corresponding to (a_i, b_i) can be routed in parallel in an optimal drawing and, therefore, the resulting map N with unit weights will have the same crossing numbers as M.

We then replace the boundaries of the annulus by cycles (using one vertex for each $a_{i,j}$ and $b_{i,j}$), obtaining a graph G. We assign weight $W = 1 + \mathrm{pcr}(N)$ to the edges in the cycles. This ensures that in a drawing of G minimizing any of the crossing numbers the boundary cycles are embedded without any intersections. Consequently, a drawing of G on the sphere that minimizes any one of the crossing numbers looks very much like the drawing of a map on the annulus. With one subtle difference: one of the boundaries may flip.

Given the map N on the annulus, the *flipped map* N' is obtained by flipping the order of the points on one of the boundaries. In other words, there are essentially two different ways of embedding the two boundary cycles of G on the sphere without intersections depending on the relative orientation of the boundaries. In one of the cases the drawing D of G gives a drawing of N, in the other case it gives a drawing of the flipped map N'. Fortunately, in the flipped case the group of edges corresponding to the weighted edge from a_i to b_i must intersect often with each other (as illustrated in Figure 3).

Now we know that

$$\begin{aligned} \mathrm{ocr}(G) &\leq \mathrm{ocr}(N) \quad \text{(since every drawing of } N \text{ is a drawing of } G) \\ &\leq w_1 w_3 + w_2 w_4 \quad \text{(by Lemma 2.2)} \\ &\leq \frac{3}{2} m^2 \quad \text{(by the choice of weights).} \end{aligned}$$

We will presently prove the following estimate on the flipped map.

Lemma 5.1 $\operatorname{ocr}(N') \geq 2m^2 - 4m$.

With that estimate and our discussion of flipped maps, we have

$$\begin{aligned}\operatorname{pcr}(G) &= \min\{\operatorname{pcr}(N), \operatorname{pcr}(N')\} \\ &\geq \min\{\operatorname{pcr}(N), \operatorname{ocr}(N')\} \quad \text{(since ocr} \leq \text{cr)} \\ &\geq \min\{\sqrt{3}m^2 - 2m, 2m^2 - 4m\} \quad \text{(choice of } w\text{, and Lemma 5.1).}\end{aligned}$$

By making m sufficiently large, we can make the ratio of $\operatorname{ocr}(G)$ and $\operatorname{pcr}(G)$ arbitrarily close to $\sqrt{3}/2$.

Theorem 5.2 *For any $\varepsilon > 0$ there is a graph G such that*

$$\operatorname{ocr}(G) < (\sqrt{3}/2 + \varepsilon)\operatorname{pcr}(G).$$

The proof of Lemma 5.1 will require the following estimate.

Lemma 5.3 *Let $0 \leq a_1 \leq a_2 \leq \cdots \leq a_n$ be such that $a_n \leq a_1 + \cdots + a_{n-1}$. Then*

$$\max_{|y_i| \leq a_i} \left(\left(\sum_{i=1}^{n} y_i \right)^2 - 2 \sum_{i=1}^{n} y_i^2 \right) = \left(\sum_{i=1}^{n} a_i \right)^2 - 2 \sum_{i=1}^{n} a_i^2.$$

Proof of Lemma 5.1 Let $w_1 = a, w_2 = b, w_3 = d, w_4 = c$ (with a, b, c, d as in the definition of N). In any drawing of N' each group of the edges split into two classes, those with an even number of twists and those with an odd number of twists (two twists make the same contribution to $\operatorname{ocr}(M')$ as no twists). Consequently, we can estimate $\operatorname{ocr}(N')$ as follows.

$$\begin{aligned}\operatorname{ocr}(N') &= \min_{k_i \in \{0,1,\ldots,w_i\}} \left(\sum_{i=1}^{4} \binom{k_i}{2} + \sum_{i=1}^{4} \binom{w_i - k_i}{2} + \sum_{i \neq j} k_i (w_j - k_j) \right) \\ &\geq -\frac{1}{2} \sum_{i=1}^{4} w_i + \min_{0 \leq x_i \leq w_i} \left(\sum_{i=1}^{4} \frac{x_i^2}{2} + \sum_{i=1}^{4} \frac{(w_i - x_i)^2}{2} + \sum_{i \neq j} x_i (w_j - x_j) \right) \\ &= -\frac{1}{2} \sum_{i=1}^{4} w_i + \frac{1}{4} \left(\sum_{i=1}^{4} w_i \right)^2 + \min_{|y_i| \leq w_i/2} \left(2 \sum_{i=1}^{4} y_i^2 - \left(\sum_{i=1}^{4} y_i \right)^2 \right) \\ &\geq \frac{1}{2} \sum_{i=1}^{4} w_i^2 - \frac{1}{2} \sum_{i=1}^{4} w_i \quad \text{(using Lemma 5.3)} \\ &\geq \frac{1}{2} \left(\left(\frac{\sqrt{3}+1}{2} m - 1 \right)^2 + 2m^2 + \left(\frac{\sqrt{3}-1}{2} m - 1 \right)^2 - 4m \right) \\ &\geq 2m^2 - 4m. \end{aligned} \tag{13}$$

The equality between the second and third line can be verified by substituting $y_i = x_i - w_i/2$. □

Proof of Lemma 5.3 Let $y_1, \ldots, y_n$ achieve the maximum value. Replacing the y_i by $|y_i|$ does not decrease the objective function. Without loss of generality, we can assume $0 \le y_1 \le y_2 \le \cdots \le y_n$. Note that if $y_i < y_j$, then $y_i = a_i$ (otherwise increasing y_i by ε and decreasing y_j by ε increases the objective function for small ε).

Let k be the largest i such that $y_i = a_i$. Let $k = 0$ if no such i exists. We have $y_i = a_i$ for $i \le k$ and $y_{k+1} = \cdots = y_n$. If $k = n$ we are done.

We conclude the proof by showing that $k < n$ is not possible. Let t be the common value of $y_{k+1} = \cdots = y_n$. Note that we have $t = y_{k+1} \le a_{k+1}$.

Let

$$f(t) = \left(\left(\sum_{i=1}^{k} a_i\right) + (n-k)t\right)^2 - 2\left(\left(\sum_{i=1}^{k} a_i^2\right) + (n-k)t^2\right).$$

We have

$$f'(t) = 2(n-k)\left(\left(\sum_{i=1}^{k} a_i\right) + (n-k-2)t\right).$$

Note that $f'(t) > 0$ for $t < a_{k+1}$. (This is easy to see when $k < n-1$; for $k = n-1$ we make use of the assumption that $a_n \le \sum_{i=1}^{n-1} a_i$.) Therefore, $f(t)$ will be maximized by $t = a_{k+1}$ over values $t \le a_{k+1}$. Hence, $y_{k+1} = a_{k+1}$, contradicting our choice of k. □

6 Conclusion

The relationship between the different crossing numbers remains mysterious, and we have already mentioned several open questions and conjectures. Here we want to revive a question first asked by Tutte (in slightly different form). Recall the definition of the algebraic crossing number from Section 4:

$$\operatorname{acr}(G) = \min_{D} \sum_{\{e,f\} \in \binom{E}{2}} |\hat{\imath}(\gamma_e, \gamma_f)|,$$

where γ_e is a curve representing edge e in a drawing D of G. It is clear that

$$\operatorname{acr}(G) \le \operatorname{cr}(G).$$

Does equality hold?

Acknowledgement Thanks to the anonymous referee for helpful comments.

References

1. Archdeacon, D.: Problems in topological graph theory. http://www.emba.uvm.edu/~archdeac/problems/altcross.html (accessed April 7th, 2005)
2. Garey, M.R., Johnson, D.S.: Crossing number is NP-complete. SIAM J. Algebr. Discrete Methods **4**(3), 312–316 (1983)
3. Gross, J.L., Tucker, T.W.: Topological Graph Theory. Dover, Mineola (2001). Reprint of the 1987 original
4. Chojnacki, C. (Haim Hanani): Über wesentlich unplättbare Kurven im drei-dimensionalen Raume. Fundam. Math. **23**, 135–142 (1934)
5. Kolman, P., Matoušek, J.: Crossing number, pair-crossing number, and expansion. J. Comb. Theory Ser. B **92**(1), 99–113 (2004)
6. Pach, J.: Crossing numbers. In: Discrete and Computational Geometry (Tokyo, 1998). Lecture Notes in Comput. Sci., vol. 1763, pp. 267–273. Springer, Berlin (2000)
7. Pach, J., Tóth, G.: Which crossing number is it anyway? J. Comb. Theory Ser. B **80**(2), 225–246 (2000)
8. Pelsmajer, M.J., Schaefer, M., Štefankovič, D.: Removing even crossings. In: EuroComb, April 2005
9. Székely, L.A.: A successful concept for measuring non-planarity of graphs: the crossing number. Discrete Math. **276**(1–3), 331–352 (2004). 6th International Conference on Graph Theory
10. Tutte, W.T.: Toward a theory of crossing numbers. J. Comb. Theory **8**, 45–53 (1970)
11. Valtr, P.: On the pair-crossing number. In: Combinatorial and Computational Geometry. MSRI Publications, vol. 52, pp. 545–551 (2005)
12. West, D.: Open problems—graph theory and combinatorics. http://www.math.uiuc.edu/~west/openp/ (accessed April 7th, 2005)

Visibility Graphs of Point Sets in the Plane

Florian Pfender

Abstract The visibility graph $\mathcal{V}(X)$ of a discrete point set $X \subset \mathbb{R}^2$ has vertex set X and an edge xy for every two points $x, y \in X$ whenever there is no other point in X on the line segment between x and y. We show that for every graph G, there is a point set $X \in \mathbb{R}^2$, such that the subgraph of $\mathcal{V}(X \cup \mathbb{Z}^2)$ induced by X is isomorphic to G. As a consequence, we show that there are visibility graphs of arbitrary high chromatic number with clique number 6 settling a question by Kára, Pór and Wood.

1 Introduction

The concept of a visibility graph is widely studied in discrete geometry. You start with a set of objects in some metric space, and the visibility graph of this configuration contains the objects as vertices, and two vertices are connected by an edge if the corresponding objects can "see" each other, i.e., there is a straight line not intersecting any other part of the configuration from one object to the other. Often, there are extra restrictions on the objects and on the direction of the lines of visibility.

Specific classes of visibility graphs which are well studied include bar visibility graphs (see [3]), rectangle visibility graphs (see [6]) and visibility graphs of polygons (see [1]). In this paper we consider visibility graphs of point sets.

Let $X \subset \mathbb{R}^2$ be a discrete point set in the plane. The *visibility graph of* X is the graph $\mathcal{V}(X)$ with vertex set X and edges xy for every two points $x, y \in X$ whenever there is no other point in X on the line segment between x and y, i.e., when the point x is visible from the point y and vice versa.

Supported by the DFG Research Center MATHEON (FZT86).

F. Pfender (✉)
MA 6-2, TU Berlin, 10623 Berlin, Germany
e-mail: fpfender@math.tu-berlin.de

Kára et al. discuss these graphs [4], and make some observations regarding the chromatic number $\chi(\mathcal{V}(X))$ and the clique number $\omega(\mathcal{V}(X))$, the order of the largest clique. In particular, they characterize all visibility graphs with $\chi(\mathcal{V}(X)) = 2$ and $\chi(\mathcal{V}(X)) = 3$, and in both cases, $\omega(\mathcal{V}(X)) = \chi(\mathcal{V}(X))$. Similarly, they show the following proposition.

Proposition 1 *Let $\mathbb{Z}^2$ be the integer lattice in the plane, then $\omega(\mathcal{V}(\mathbb{Z}^2)) = \chi(\mathcal{V}(\mathbb{Z}^2)) = 4$.*

Note that $\mathcal{V}(\mathbb{Z}^2)$ is not perfect as it contains induced 5-cycles. Further, it is not true in general that $\omega(\mathcal{V}(X)) = \chi(\mathcal{V}(X))$—there are point sets with as few as nine points with $\omega(\mathcal{V}(X)) = 4$ and $\chi(\mathcal{V}(X)) = 5$.

For general graphs, there are examples with $\chi(G) = k$ and $\omega(G) = 2$ for any k, one famous example is the sequence of graphs M_{k-2} by Mycielski [5]. No similar construction is known for visibility graphs with a bounded clique number. As their main result, Kára et al. construct a family of point sets with $\chi(\mathcal{V}(X)) \geq (c_1 \log\ \omega(\mathcal{V}(X_i)))^{c_2 \log\ \omega(\mathcal{V}(X_i))}$ for some constants c_1 and c_2 and with $\omega(\mathcal{V}(X_i))$ getting arbitrarily large. Our main result is the following theorem.

Theorem 2 *For every graph G, there is a set of points $X \subset \mathbb{R}^2$ such that the subgraph of $\mathcal{V}(X \cup \mathbb{Z}^2)$ induced by X is isomorphic to G.*

Let G_k be a graph with $\chi(G_k) = k$ and $\omega(G_k) = 2$, and let X_k be the corresponding set given by Theorem 2. Let $Y_k \subset X_k \cup \mathbb{Z}^2$ be the subset of points contained in the convex hull of X_k. Then $\chi(\mathcal{V}(Y_k)) \geq \chi(G_k) = k$ and $\omega(\mathcal{V}(Y_k)) \leq \omega(G_k) + \omega(\mathcal{V}(\mathbb{Z}^2)) = 6$, so we get the following corollary settling the question from above raised by Kára et al.

Corollary 3 *For every k, there is a finite point set $Y \subset \mathbb{R}^2$, such that $\chi(\mathcal{V}(Y)) \geq k$ and $\omega(\mathcal{V}(Y)) = 6$.*

2 Proof of Theorem 2

Let G be a graph with vertex set $V(G) = \{1, 2, \ldots, n\}$ and edge set $E(G)$. We prove the following lemma in Sect. 3.

Lemma 4 *For M large enough, there is a set of prime numbers $\{p_{ij} : 1 \leq i < j \leq n\}$ with the following properties*:

1. $2^M < p_{ij} < 2^{M+1}$.
2. *For $1 \leq k \leq n$, let $P_k = 2^{n_k} \prod_{i=1}^{k-1} p_{ik} \prod_{j=k+1}^{n} p_{kj}$, and choose $n_k \in \mathbb{Z}$ such that $\lfloor \log_2 P_k \rfloor = nM + 2k$. Then $p_{k\ell}$ is the only number in $\{p_{ij} : 1 \leq i < j \leq n\}$ which divides $P_\ell - P_k$ for $1 \leq k < \ell \leq n$.*

Note that $\prod_{i=1}^{k-1} p_{ik} \prod_{j=k+1}^{n} p_{kj} < 2^{(n-1)(M+1)} < 2^{nM}$, and thus $n_k > 0$ and $P_k \in \mathbb{Z}$ for all k. From this, we can construct the set of points X in Theorem 2:

$$X = \{x_i : 1 \le i \le n\} \subset \mathbb{R}^2, \quad \text{with } x_i = \left(2^{-nM} P_i,\ i\ \frac{\prod_{k<j}(P_j - P_k)}{\prod_{kj \in E(G)} p_{kj}}\right).$$

Before we prove the lemma, we show that this point set has the properties stated in the theorem. For $1 \le i < \ell \le n$, let $m_{i\ell}$ be the slope of the line through x_i and x_ℓ. Then

$$m_{i\ell} = \frac{\ell - i}{P_\ell - P_i} \cdot \frac{2^{nM} \prod_{k<j}(P_j - P_k)}{\prod_{kj \in E(G)} p_{kj}}.$$

There are no three collinear points in X, as

$$2^{nM+2i+1} \le P_{i+1} - P_i < 2^{nM+2i+3},$$

thus $m_{i(i+1)} > m_{(i+1)(i+2)}$, and therefore $m_{i\ell} > m_{ik}$ for $i < \ell < k$. Thus, $\mathcal{V}(X)$ is complete, and it remains to show that there is an integer point on the line segment between x_i and x_ℓ if and only if $i\ell \notin E(G)$. To establish this goal, we look at the intersections of the line segment from x_i to x_ℓ $(i < \ell)$ with the integer gridlines parallel to the y-axis.

Let $s \in \mathbb{Z}$, with $2^{-nM} P_i < s < 2^{-nM} P_\ell < 2^{2n+1}$. As $2^{2j} \le 2^{-nM} P_j < 2^{2j+1}$ for every j, such an s exists. Let $z_{i\ell}^s = (s, y_{i\ell}^s)$ be a point on the line segment from x_i to x_ℓ. Then

$$\begin{aligned} y_{i\ell}^s &= i\,\frac{\prod_{k<j}(P_j - P_k)}{\prod_{kj \in E(G)} p_{kj}} + (s - 2^{-nM} P_i) m_{i\ell} \\ &= \underbrace{i\,\frac{\prod_{k<j}(P_j - P_k)}{\prod_{kj \in E(G)} p_{kj}}}_{(1)} + \underbrace{s\,\frac{\ell - i}{P_\ell - P_i} \cdot \frac{2^{nM} \prod_{k<j}(P_j - P_k)}{\prod_{kj \in E(G)} p_{kj}}}_{(2)} \\ &\quad + \underbrace{P_i\,\frac{\ell - i}{P_\ell - P_i} \cdot \frac{\prod_{k<j}(P_j - P_k)}{\prod_{kj \in E(G)} p_{kj}}}_{(3)}. \end{aligned}$$

Expression (1) is an integer since p_{kj} divides $P_j - P_k$. By the same argument, (3) is an integer—just note further that $p_{i\ell}$ divides P_i. It remains the analysis of (2).

If $i\ell \notin E(G)$, then (2) is an integer. Therefore, $z_{i\ell}^s \in \mathbb{Z}^2$, and $x_i x_\ell \notin E(\mathcal{V}(X \cup \mathbb{Z}^2))$. If $i\ell \in E(G)$, observe that $p_{i\ell} > 2^M > \max\{\ell - i, s\}$, so $p_{i\ell}$ does not divide s or $\ell - i$. Clearly, $p_{i\ell}$ does not divide 2^{nM}, and, by Lemma 4, it does not divide any of the $P_j - P_k$ other than $P_\ell - P_i$. Thus, (2) is not an integer, $z_{i\ell}^s \notin \mathbb{Z}^2$ for all s considered, and $x_i x_\ell \in E(\mathcal{V}(X \cup \mathbb{Z}^2))$, proving Theorem 2.

3 Proof of Lemma 4

By an inequality of Finsler [2], there are more than $2^M/(3(M+1)\ln 2) > 2n^3$ prime numbers in the interval from 2^M to 2^{M+1}.

We pick the p_{ij} sequentially in the order $p_{12}, p_{13}, \ldots, p_{1n}, p_{23}, \ldots, p_{(n-1)n}$, with the following conditions given by the lemma:

(a) p_{ij} is a prime number with $2^M < p_{ij} < 2^{M+1}$.
(b) p_{ij} is different from all the primes picked before.
(c) p_{ij} does not divide $P_k - P_\ell$ for all $1 \le \ell < k < i$.
(d) If $j = n$, no $p_{k\ell}$ divides $P_i - P_r$ for $\{k, \ell\} \neq \{i, r\}$.

Assume that we have picked numbers up to but not including p_{ij} according to (a)–(d), and we want to pick p_{ij}. Consider first the case that $j < n$. There were less than $\binom{n}{2}$ primes selected before, and each $P_k - P_\ell$ has at most n prime divisors greater than 2^M, thus at most

$$\binom{n}{2} + n\binom{n}{2} < n^3$$

of the choices are blocked, and we can find p_{ij} according to (a)–(c).

If $j = n$, pick p_{ij} according to (a)–(c), and assume that $p_{k\ell}$ divides $P_i - P_r$ for some $\{k, \ell\} \neq \{i, r\}$ (i.e., condition (d) is violated). We have $k \neq i$ as all $p_{i\ell}$ divide P_i, otherwise $p_{i\ell}$ also divides P_r and thus $r = \ell$, a contradiction. Similarly, $\ell \neq i$.

Pick another number p'_{ij} according to (a)–(c). If $p_{k\ell}$ divides $P'_i - P_r$, then $p_{k\ell}$ divides $P'_i - P_i = (p'_{ij} - p_{ij})P_i/p_{ij}$, and thus $p_{k\ell}$ divides $p'_{ij} - p_{ij}$. However, this is impossible since $|p'_{ij} - p_{ij}| < 2^M < p_{k\ell}$. Therefore, each $p_{k\ell}$ can block at most one choice for p_{ij} this way, so in total at most $\binom{n}{2}$ further choices are blocked by condition (d), and we can always find a number p_{ij} with (a)–(d). This concludes the proof of the lemma.

4 Further Questions

We have shown that there are visibility graphs with $\chi(\mathcal{V}(X)) \ge k$ and $\omega(\mathcal{V}(X)) = 6$ for every k. For all visibility graphs with $\omega(\mathcal{V}(X)) \le 3$, we know that $\chi(\mathcal{V}(X)) = \omega(\mathcal{V}(X))$. The only cases left to consider are $\omega(\mathcal{V}(X)) = 4$ and $\omega(\mathcal{V}(X)) = 5$. A similar technique of combining a visibility graph with $\omega(\mathcal{V}(X)) = 3$ with a graph G with $\omega(G) = 2$ and a large chromatic number will not work, since the visibility graphs with $\omega(\mathcal{V}(X)) = 3$ are too simple (all but at most two of their vertices are collinear unless $\mathcal{V}(X)$ is a special graph on six vertices). It would be no surprise to us if the chromatic number of visibility graphs with $\omega(\mathcal{V}(X)) = 5$ is bounded.

Finally, one could look for smaller point sets with $\chi(\mathcal{V}(X)) \ge k$ and $\omega(\mathcal{V}(X)) = 6$, as our sets tend to be very large.

References

1. Abello, J., Kumar, K.: Visibility graphs and oriented matroids. Discrete Comput. Geom. **28**, 449–465 (2002)

2. Finsler, P.: Über die Primzahlen zwischen n und $2n$. In: Festschrift zum 60. Geburtstag von Prof. Dr. Andreas Speiser, pp. 118–122. Füssli, Zürich (1945)
3. Hutchinson, J.P.: A note on rectilinear and polar visibility graphs. Discrete Appl. Math. **148**, 263–272 (2005)
4. Kára, J., Pór, A., Wood, D.R.: On the chromatic number of the visibility graph of a set of points in the plane. Discrete Comput. Geom. **34**, 497–506 (2005)
5. Mycielski, J.: Sur le coloriage des graphs. Colloquium Math. **3**, 161–162 (1955)
6. Streinu, I., Whitesides, S.: Rectangle visibility graphs: characterization, construction, and compaction. In: Proc. of the STACS 2003. Lecture Notes in Computer Science, vol. 2607, pp. 26–37. Springer, Berlin (2003)

Decomposability of Polytopes

Krzysztof Przesławski · David Yost

Abstract A known characterization of the decomposability of polytopes is reformulated in a way which may be more computationally convenient, and a more transparent proof is given. New sufficient conditions for indecomposability are then deduced, and illustrated with some examples.

1 Introduction

This paper is concerned with criteria for the indecomposability of polytopes. We recall that a polytope P is *decomposable* if it is equal to a Minkowski sum $Q + R$ of two polytopes Q and R which are not homothetic to P. Naturally, all other polytopes are described as indecomposable. The concept of decomposability is due to Gale [1] although he used a different name. The concept is also interesting for more general convex bodies, but we do not consider them here. It is not surprising to learn [1] that triangles are indecomposable, and, conversely, that any two-dimensional polygon is the sum of triangles and segments. Gale also announced that any pyramid, i.e. the convex hull of a facet and a single point, is indecomposable. Shephard made perhaps the next serious study of it, showing amongst other things that a polytope is indecomposable if all of its 2-faces are triangles [5, (13)]. A number of papers have subsequently found progressively weaker sufficient conditions for indecomposability and we are continuing this tradition.

K. Przesławski (✉)
Faculty of Mathematics, Computer Science and Econometrics, University of Zielona Góra,
65-246 Zielona Góra, Poland
e-mail: K.Przeslawski@im.uz.zgora.pl

D. Yost
School of Information Technology and Mathematical Sciences, University of Ballarat,
P.O. Box 663, Ballarat, Victoria 3353, Australia
e-mail: d.yost@ballarat.edu.au

The latter result is a special case of (12) of [5], which asserts that a polytope is indecomposable if there is an edge to which all vertices are connected by a strong chain of indecomposable faces. A simple reformulation of this statement is that a polytope is indecomposable if there is a strong chain of indecomposable faces which contains all the vertices.

By a strong chain of faces is meant a finite sequence of faces in which each successive pair shares an edge. McMullen [3] showed that the hypothesis of Shephard's result could be relaxed in the following way: the union of this chain need contain only one vertex from each facet, not all of them. His proof, like Shephard's, was geometric in character, although the statement of the hypothesis is graph theoretic. By that, we mean that the 1-skeleton of a polytope is clearly a graph, and the hypothesis is just a statement about this graph.

Earlier Kallay [2] had weakened the hypothesis in several other ways. One was to consider collections of vertices which did not necessarily form a face. For example, three vertices can be pairwise adjacent, whilst their centroid is an interior point of the polytope. (Blissfully unaware of [2], the second author used a similar approach in [7] for examining irreducibility of centrally symmetric polytopes.) Another weakening was to show that each successive pair of the chain could share just two vertices, not necessarily an edge. He adopted a strictly graph theoretic approach, defining the concept of indecomposability for geometric graphs, and showing that a polytope is indecomposable if and only if its 1-skeleton is indecomposable in this sense.

Our aim is to extend some of the results obtained in these works. Although similar to [2], our approach is simpler and more general; in particular, we require no knowledge of spherical complexes. Implicit in [2] is the use of a mapping from the vertices of the polytope into the ambient space. This was explicit in [7] and we carry on with it here.

2 Basic Notions

All graphs considered here are assumed to have a finite number of vertices. Let $G = (V, E)$ be a graph with set of vertices V and set of edges $E \subset \binom{V}{2}$. Mostly we are interested in the 1-skeleton of a polytope, but it is practical to consider this more abstract situation. Let $f, g \in (\mathbf{R}^d)^V$. Let I be a non-empty subset of $\mathbf{R}$. We say that g is *edgewise I-dominated* by f if for any pair $u, v \in V$ of adjacent vertices there exists $\alpha \in I$ such that

$$g(u) - g(v) = \alpha\big(f(u) - f(v)\big).$$

In the case $I = \mathbf{R}$, we say simply that g is *edgewise dominated* by f, and write $g \preceq f$. (In case f is the identity mapping, g is an isomorphism and $I = (0, \infty)$, this coincides with the concept of local similarity defined in [2].) If I is a non-zero singleton, then we also say that g is *similar* to f.

Observe that the sets $E(f) := \{g\colon\ g \preceq f\}$ and $S(f) := \{g\colon\ g \text{ is similar to } f\}$ are vector spaces. Clearly, $S(f)$ is the direct sum of the d-dimensional subspace of translations and the one-dimensional subspace of multiples of the identity. If these spaces are equal, then f is said to be *indecomposable*. Thus the quotient space

$D(f) := E(f)/S(f)$ relates to "decomposability" of f. The dimension of $D(f)$ is called the *index of decomposability* of f. We denote this index by $\operatorname{dec} f$. Hence f is indecomposable if and only if $\operatorname{dec} f = 0$. These notions find a natural interpretation when we discuss decomposability of polytopes.

Suppose that a function $\varphi\colon V \to \mathbf{R}$ is given. We say that φ attains a (local!) *maximum* at $v \in V$ if for any u adjacent to v we have $\varphi(v) \geq \varphi(u)$. The set of all maximizers of φ is denoted throughout by $\operatorname{argmax} \varphi$.

We begin with an auxiliary result. For ease of notation, we prefer to talk about linear functionals on $\mathbf{R}^d$, rather than the scalar product in $\mathbf{R}^d$.

Lemma 1 *Let V be the set of vertices of a graph G. Let $f, g \in (\mathbf{R}^d)^V$, and let g be edgewise $(0, +\infty)$-dominated by f. Then for any $y \in (\mathbf{R}^d)^*$*

$$\operatorname{argmax} y \circ g = \operatorname{argmax} y \circ f.$$

Proof If $v \notin \operatorname{argmax} y \circ f$, then there exists a vertex w adjacent to v such that $y \circ f(v) < y \circ f(w)$. By our assumptions, there exists $\alpha > 0$ such that $\alpha(f(w) - f(v)) = g(w) - g(v)$. Applying y to this equation, we get readily from the linearity of y that $y \circ g(w) - y \circ g(v) > 0$. Thus $v \notin \operatorname{argmax} y \circ g$. The symmetrical relationship between f and g completes the proof. □

Recall that a graph G is called a *cycle* if $|V| = k \geq 3$ and V can be ordered as $\{x_1, \ldots, x_k\}$, so that $E = \{\{x_1, x_2\}, \ldots, \{x_{k-1}, x_k\}, \{x_k, x_1\}\}$. The number k is said to be the *length* of the cycle. The following result is simple, but it does lead us to new examples of indecomposable polytopes.

Proposition 2 *Let C_k be a cycle of length k. Let $f\colon C_k \to \mathbf{R}^d$ be an injection for which $f(C_k)$ is an affinely independent set, that is, elements of $f(C_k)$ are vertices of a simplex. Then* $\operatorname{dec} f = 0$.

Proof Let $g \preceq f$. For each $i \leq k$, let $u_i = g(x_{i+1}) - g(x_i)$ and $v_i = f(x_{i+1}) - f(x_i)$ (we let here $x_{k+1} = x_1$). By definition, for each i there exists α_i such that $u_i = \alpha_i v_i$. From $\sum u_i = 0$, we obtain $\sum \alpha_i v_i = 0$. This equation and the fact that elements x_i are affinely independent readily imply that all numbers α_i are equal. □

We do not make any use of the next result. However, we include it, as it helps to understand the situation.

Proposition 3 *Let $G = (V, E)$ be a graph. If $|V| > 2$ and there exists an injection $f\colon V \to \mathbf{R}^d$ such that* $\operatorname{dec} f = 0$, *then G is 2-connected.*

Proof It is clear that G is connected. If G were not 2-connected, then there would exist an edge $\{u, v\}$ whose removal would disconnect the graph. Let A and B be the components of u and v, respectively. Defining the function g by $g_{|A} = f_{|A}$ and $g_{|B} = f_{|B} + f(v) - f(u)$, it is clear that $g \in E(f) \setminus S(f)$. Consequently, $\operatorname{dec} f \neq 0$. □

3 Decomposability and Indecomposability

We use standard notation which will not surprise anyone [8]. By $|P$ we mean the set of *vertices* of P. A set $F \subset P$ is a *face* of P if there exists $y \in (\mathbf{R}^d)^*$ such that $F = \{v \in P\colon\ y(v) = h(P, y)\}$, where $h(P, y) = \max y(P)$. The mapping $y \mapsto h(P, y)$ is called the *support function* of P. We mean by the 1-*skeleton* of P the graph $G_P = (V, E)$ such that $V = |P$ and E consists of all these pairs $\{u, v\}$ for which the line segment $[u, v]$ is a one-dimensional face of P.

As an immediate consequence of Lemma 1 we have

Lemma 4 *Let G_P be the 1-skeleton of a polytope P. For $y \in (\mathbf{R}^d)^*$, let $C = \{v \in |P\colon\ y(v) = h(P, y)\}$. If $g\colon\ |P \to \mathbf{R}^d$ is edgewise $(0, \infty)$-dominated by $\mathrm{id}_{|P}$, then $g(C)$ is equal to*

$$\big\{w \in g(|P)\colon\ y(w) = \max y \circ g(|P)\big\}.$$

Proof It suffices to observe that $C = \operatorname{argmax} y \circ \mathrm{id}_{|P}$. □

Let conv A denote the *convex hull* of $A \subset \mathbf{R}^d$. Let $Q = \operatorname{conv} g(|P)$, where g is as in the lemma. It follows that g is a one-to-one correspondence between $|P$ and $|Q$ and that g^{-1} is edgewise $(0, \infty)$-dominated by $\mathrm{id}_{|Q}$ (g^{-1} relates here to the 1-skeleton of Q). Moreover, the induced mapping $\widetilde{g}$ defined on faces of P by the formula

$$\widetilde{g}(F) = \operatorname{conv} g(F)$$

is an isomorphism of the facial structures of P and Q.

Corollary 5 *Suppose that the mapping g is edgewise $[0, \infty)$-dominated by $\mathrm{id}_{|P}$, and denote again $Q = \operatorname{conv} g(|P)$. Let F be a face of P and let $y \in (\mathbf{R}^d)^*$ be such that $y(v) = h(P, y)$, whenever $v \in F$. Then*

$$g(|F) \subset \big\{w \in g(|P)\colon\ y(w) = h(Q, y)\big\}.$$

Proposition 6 *If g is edgewise $[0, 1]$-dominated by $\mathrm{id}_{|P}$, then Q, defined as before, is a summand of P, that is, there exists a polytope R such that $P = Q + R$.*

Proof Let $k(u) = u - g(u)$. The function k is also edgewise $[0, 1]$-dominated by $\mathrm{id}_{|P}$. Let $R = \operatorname{conv} k(|P)$. For $y \in (\mathbf{R}^d)^*$, choose $v \in |P$ such that $y(v) = h(P, y)$.

By the preceding corollary, $y \circ g(v) = h(Q, y)$ and $y \circ k(v) = h(R, y)$. Moreover, by the definition of k, $y(v) = y \circ g(v) + y \circ k(v)$. Thus, $h(P, y) = h(Q, y) + h(R, y)$, which implies $P = Q + R$. □

The next theorem is essentially Corollary 5 of [2]. It is formulated there in a different but equivalent form.

Theorem 7 *P is decomposable if and only if $\operatorname{dec} \mathrm{id}_{|P} \neq 0$. Moreover, any non-similar function g that is edgewise $[0, 1]$-dominated by $\mathrm{id}_{|P}$ defines a non-homothetic summand of P.*

Proof If P is decomposable, there exist polytopes Q and R, which are non-homothetic to P, such that $P = Q + R$. Thus, for any $v \in |P$ there exists a unique element $g(v) \in |Q$ for which we have $v \in g(v) + R$. It is easy to see that g is edgewise $[0, 1]$-dominated by $\mathrm{id}_{|P}$. Since g is onto $|Q$ and $Q = \operatorname{conv} g(|P)$, g cannot belong to $S(\mathrm{id}_{|P})$, for otherwise Q would be a homothetic copy of P.

Conversely, suppose that $\operatorname{dec} \mathrm{id}_{|P} \neq 0$. Then there exists some $f \in E(\mathrm{id}_{|P}) \setminus S(\mathrm{id}_{|P})$. If $\alpha > 0$ is sufficiently small, αf will be edgewise $(0, 1)$-dominated by $\mathrm{id}_{|P}$. Put $g = \mathrm{id}_{|P} - \alpha f$. Then g is edgewise $(0, 1)$-dominated by but not similar to $\mathrm{id}_{|} P$. By Proposition 6, $Q = \operatorname{conv} g(P)$ is a summand of P. The fact that it is not a homothetic copy of P is clear.

The second part of the theorem is rather obvious. □

For further use we need a graph theoretic consequence of the above result, essentially Proposition 8 of [2]. We note that the subgraph G here need not be the 1-skeleton of any polytope.

Theorem 8 *Let P be a d-dimensional polytope in $\mathbf{R}^d$. Then P is indecomposable if and only if there exists a subgraph $G = (V, E)$ of the 1-skeleton of P such that id_V is indecomposable (as a mapping related to G), and V meets every facet of P.*

Proof We have to show the "if" part only, as the "only if" part is a consequence of the preceding theorem. (It suffices to let G be the 1-skeleton of P.)

Suppose that Q is a non-trivial summand of P, that is, Q contains more than one element. Take the function $g\colon |P \to |Q$ defined as in the preceding proof. Since id_V is indecomposable, there is a number α and a vector $x \in \mathbf{R}^d$ such that

$$g_{|V} = \alpha \,\mathrm{id}_V + x. \tag{1}$$

It is clear that shifting Q if necessary we may assume $x = 0$. We may also assume that 0 belongs to the interior of P.

Let $y \in (\mathbf{R}^d)^*$ be any outer normal of a facet F of P and let $v \in V \cap F$. By (1) and Corollary 5

$$h(Q, y) = y \circ g(v) = \alpha y(v) = \alpha h(P, y).$$

Obviously, for at least one of the normals we have $h(Q, y) > 0$. Hence $\alpha > 0$ and $h(Q, y) = h(\alpha P, y)$ for each normal y. Since Q is a summand of P, we deduce that $Q = \alpha P$, which implies the indecomposability of P. □

Our next notion relates to the notion of strongly connected family of polytopes which is useful in formulating sufficient conditions for indecomposability (see [3, 6], and [7]).

Let $\mathcal{G}$ be a family of subgraphs of a graph G. We say that $\mathcal{G}$ is *strongly connected* if, for any pair of graphs $G, K \in \mathcal{G}$, there exists a sequence $G_1, \ldots, G_k$ of graphs in $\mathcal{G}$ with sets of vertices $V_1, \ldots, V_k$, respectively, such that $G_1 = G$, $G_k = K$ and $|V_i \cap V_{i+1}| \geq 2$ for $i = 1, \ldots, k - 1$. Such a sequence is called a *strong chain of graphs*.

Now, as a simple consequence of the previous result we obtain

Theorem 9 *Let P be a polytope in $\mathbf{R}^d$. Let $\mathcal{G}$ be a strongly connected family of subgraphs of the 1-skeleton G_P. If for each $(V, E) \in \mathcal{G}$ the identity map id_V is indecomposable and $W := \bigcup\{V\colon (V, E) \in \mathcal{G}\}$ meets every facet of P, then P is indecomposable.*

Proof Let $D = \bigcup\{E\colon (V, E) \in \mathcal{G}\}$. It suffices to show that id_W, as a mapping related to $G := (W, D)$, is indecomposable. Let $g\colon W \to \mathbf{R}^d$ be similar to id_W. Fix $u \in W$. For any $w \in W$ there exists a strong chain $G_1, \ldots, G_k$ of graphs in $\mathcal{G}$ such that $u \in V_1$ and $w \in V_k$. Let g_i be the restriction of g to V_i. By our assumptions, for each i there exist $\alpha_i \in \mathbf{R}$ and $z_i \in \mathbf{R}^d$ such that $g_i(x) = \alpha_i x + z_i$. By the definition of a strong chain, there exist two different elements s and t which belong to $V_i \cap V_{i+1}$. Therefore,

$$g(s) - g(t) = \alpha_i(s - t) = \alpha_{i+1}(s - t),$$

which implies that $\alpha_i = \alpha_{i+1}$ and also $z_i = z_{i+1}$. In consequence, g is similar to id_W. □

Previous workers [3–6] have usually assumed that each graph (V, E) belonging to $\mathcal{G}$ has its vertices V contained in a proper face of the polytope P. We emphasize that this assumption is not necessary. This point is implicit in [2] and explicit in [7, p. 137], although the latter deals only with triangles.

Applicability of Theorem 9 depends on the existence of a reasonable class of graphs embedded into $\mathbf{R}^d$ for which the identity is indecomposable. As is shown by Proposition 2, the simplest graphs that conform to these demands are cycles. We make use of the following:

Corollary 10 *Let P be a polytope in $\mathbf{R}^d$. Let $\mathcal{G}$ be a strongly connected family of subgraphs of the 1-skeleton G_P. If each $(V, E) \in \mathcal{G}$ is a cycle with an affinely independent set of vertices and each facet of P has a vertex that belongs to a certain graph from $\mathcal{G}$, then P is indecomposable.*

4 Some Applications

Meyer [4] and Kallay [2] gave examples of decomposable three-dimensional polytopes possessing combinatorially equivalent copies which are indecomposable. It is known [6, p. 47] that any such polytope must have at least eight vertices. Kallay's polytope has ten vertices while Meyer's has even more. Smilansky [6, Theorem 6.11(b)] announced the existence of a three-dimensional polytope of this kind with exactly eight vertices, and referred the reader to his thesis for the details. As an application, we now give an example of this kind. We have not had access to Smilanky's thesis but we would not be surprised if his example is equivalent to ours.

Example 11 A conditionally decomposable polyhedron with eight vertices.

Let P be the convex hull of the following points: $A_1 = (2, 1, 0)$, $A_2 = (1, 2, 0)$, $A_3 = (-2, -1, 0)$, $A_4 = (-1, -2, 0)$, $B_1 = (-1, -1, 1)$, $B_2 = (1, 1, 1)$, $C_1 =$

$(1, 1, -1)$ and $C_2 = (-1, -1, -1)$. One may visualize P as the union of two roves, one in the half-space $z \geq 0$, the other in the half-space $z \leq 0$, with a common base $A_1A_2A_3A_4$. It can then be seen that P is the Minkowski sum of the standard octahedron and a segment parallel to $(1, 1, 0)$, i.e. the line segment $[C_1, C_2]$ is a summand of P. Next we define another polytope Q, whose vertices are labeled in the same way as for P. The idea is to tilt the face $A_1B_2B_1A_4$ a little about the edge B_1A_4. To do so, we replace A_1 by $(2 + 3\varepsilon, 1, 3\varepsilon)$ and replace B_2 by $(1 + 2\varepsilon, 1, 1 + 2\varepsilon)$ (where ε need not be too small), and let the other vertices be the same as in P. Thus Q is obtained simply by perturbing two vertices of P. We must verify that the labeling induces a one-to-one correspondence between the facial structures of P and Q. In detail, A_1 is still in the plane $x - y - z = 1$, B_2 is still in the plane $x - y - z = -1$ and both of them are in the plane

$$(\varepsilon - 1)x + (\varepsilon + 1)y - (\varepsilon + 1)z = -1 - 3\varepsilon,$$

as are the original points A_4 and B_1. So these three planes contain the faces $A_1A_4C_1C_2$, $A_2A_3B_1B_2$ and $A_1B_2B_1A_4$, respectively, and Q is combinatorially equivalent to P. If $\varepsilon \neq 0$, A_1 is taken out of the xy-plane. Then for Q, the cycle $A_1A_2A_3A_4$ is a subgraph of the 1-skeleton of Q which satisfies the assumptions of Corollary 10; in particular, the vertices A_i are affinely independent. Consequently, Q is indecomposable.

The point of this note is that there are other polytopes which can be shown to be indecomposable by Corollary 10 but not by earlier results. We present some now.

Example 12 There is a combinatorially indecomposable polyhedron with eleven vertices and six triangular faces, no two of which have a common edge. Thus traditional methods of proving indecomposability are not available. However, in any geometric realization, it has two affinely independent 4-cycles, with two vertices in common, whose union touches every face (Fig. 1).

Let $h = \frac{1}{2}$, or any other suitable number. Put $A = (1, 0, -1)$, $B = (0, -1, 1)$, $C = (-1, 1, 0)$, $D = (1, -1, h)$, $E = (1, -1, -h)$, $F = (-1, h, 1)$, $G = (-1, -h, 1)$, $H = (h, 1, -1)$, $J = (-h, 1, -1)$, $N = (1, 1, 1)$ and $S = (-1, -1, -1)$.

Clearly no two triangular faces have a common edge. Still, indecomposability can be proved easily by noting that the connected 4-cycles *NASB* and *NBSC* are affinely independent and their union touches every face.

There exist polytopes combinatorially equivalent to this one, in which the corresponding 4-cycle *NASB* is affinely dependent. (A concrete example is given after Example 13.) Nevertheless, this polytope is combinatorially indecomposable. It suffices to observe that of the three connected 4-cycles *NASB*, *NASC* and *NBSC*, at least two must be affinely independent (in any given geometric realization), otherwise the vertices A, B, C, N and S would be co-planar.

Note that any indecomposable polyhedron must have at least four triangular faces [6, Corollary 6.8].

Example 13 There is a combinatorially indecomposable polytope with nine vertices and only four triangular faces, of which no two have a common edge (Fig. 2).

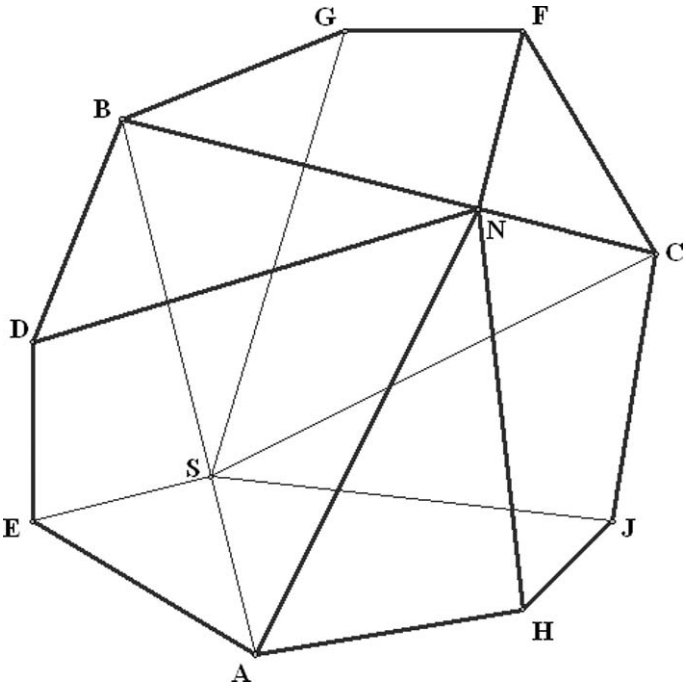


Fig. 1 Example 12.

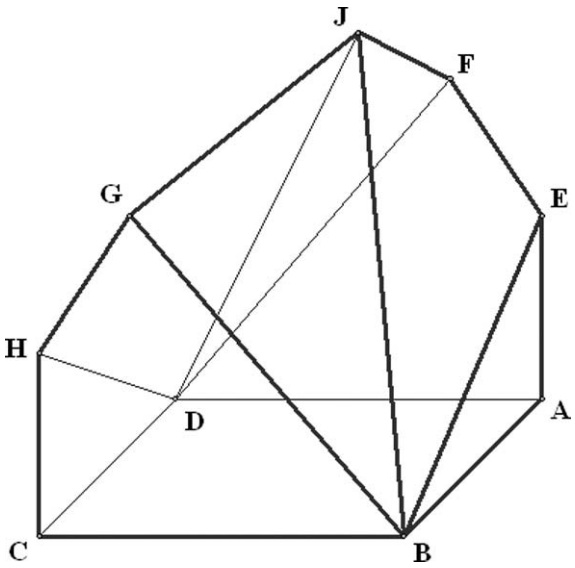


Fig. 2 Example 13.

The same argument using Corollary 10 works but this time it is simpler; we need consider only one 4-cycle, not two. Let $A = (-1, 1, -1)$, $B = (1, 1, -1)$, $C = (1, -1, -1)$, $D = (-1, -1, -1)$, $E = (-1, 1, 0)$, $F = (-1, \frac{1}{2}, \frac{3}{4})$, $G = (1, -\frac{1}{2}, \frac{3}{4})$, $H = (1, -1, 0)$ and $J = (0, 0, 1)$, and let P be their convex hull. We now list the faces of P, together with the equations of the planes containing them:

$$ABCD:\ z = -1, \qquad ADEF:\ x = -1, \qquad BCGH:\ x = 1,$$
$$BEFJ:\ x + 3y + 2z = 2, \qquad DGHJ:\ -x - 3y + 2z = 2,$$
$$ABE:\ y = 1, \qquad CDH:\ y = -1,$$
$$BGJ:\ 5x + 7y + 6z = 6, \qquad DFJ:\ -5x - 7y + 6z = 6.$$

In any polyhedron equivalent to P, the 4-cycle $BCDJ$ must be affinely independent. It clearly touches every face, so P is indecomposable. Again, arguments with triangles will not work.

Let us remark that if we add two extra vertices to this polyhedron, $K = (-\frac{1}{2}, 0, -2)$ and $L = (\frac{1}{2}, 0, -2)$, then the resulting polyhedron is combinatorially equivalent to Example 12.

References

1. Gale, D.: Irreducible convex sets. In: Proc. International Congress of Mathematicians, Amsterdam, vol. 2, pp. 217–218 (1954)
2. Kallay, M.: Indecomposable polytopes. Isr. J. Math. **41**, 235–243 (1982)
3. McMullen, P.: Indecomposable convex polytopes. Isr. J. Math. **58**, 321–323 (1987)
4. Meyer, W.J.: Indecomposable polytopes. Trans. Am. Math. Soc. **190**, 77–86 (1974)
5. Shephard, G.C.: Decomposable convex polytopes. Mathematika **10**, 89–95 (1963)
6. Smilansky, Z.: Decomposability of polytopes and polyhedra. Geom. Dedicata **24**, 29–49 (1987)
7. Yost, D.: Irreducible convex sets. Mathematika **38**, 134–155 (1991)
8. Ziegler, G.M.: Lectures on Polytopes. Graduate Texts in Mathematics, vol. 152. Springer, New York (1995)

An Inscribing Model for Random Polytopes

Ross M. Richardson · Van H. Vu · Lei Wu

Abstract For convex bodies K with $\mathcal{C}^2$ boundary in $\mathbb{R}^d$, we explore random polytopes with vertices chosen along the boundary of K. In particular, we determine asymptotic properties of the volume of these random polytopes. We provide results concerning the variance and higher moments of this functional, as well as an analogous central limit theorem.

1 Introduction

Let X be a set in $\mathbb{R}^d$ and let $t_1, \ldots, t_n$ be independent random points chosen according to some distribution μ on X. The convex hull of the t_i's is called a *random polytope* and its study is an active area of research which links together combinatorics, geometry and probability. This study traces its root to the middle of the nineteenth century with Sylvester's famous question about the probability of four random points in the plane forming a convex quadrangle [17], and has become a mainstream research area since the mid 1960s, following the investigation of Rényi and Sulanke [13] and Efron [8].

The research of V.H. Vu was done under the support of A. Sloan Fellowship and an NSF Career Grant. The research of L. Wu is done while the author was at University of California San Diego.

R.M. Richardson (✉) · L. Wu
Department of Mathematics, UCSD, 9500 Gilman Drive, La Jolla, CA 92093-0112, USA
e-mail: rmrichardson@math.ucsd.edu

L. Wu
e-mail: lei.berners.wu@googlemail.com

V.H. Vu
Department of Mathematics, Rutgers University, Piscataway, NJ 08854, USA
e-mail: vanvu@math.rutgers.edu

Throughout this paper, if not otherwise mentioned, we fix a convex body $K \in \mathcal{K}^2_+$, where $\mathcal{K}^2_+$ is the set of compact, convex bodies in $\mathbb{R}^d$ which have non-empty interior and whose boundaries are $\mathcal{C}^2$ and have everywhere positive Gauß–Kronecker curvature. The reader who is interested in the case of general K, e.g. when K is a polytope, is referred to [7, 18, 19]. Without loss of generality, we also assume K has volume 1. For a set $X \subset \mathbb{R}^d$ we define $[X]$ to be the convex hull of X.

A standard definition for the notion of a random polytope is as follows. Let $t_1, \ldots, t_n$ be independent random points chosen according to the uniform distribution on K. We let $K_n = [\{t_1, \ldots, t_n\}]$. Here and later we write $K_n = \{t_1, \ldots, t_n\}$ instead to simplify notations without causing much confusion. Another one, which we call the "inscribing polytope" model, also begins with a convex body K, but the points are chosen from the surface of K with respect to a properly defined measure. The main goal of the theory of random polytopes is to understand the asymptotic behavior ($n \to \infty$) of certain key functionals on K_n, such as the volume or the number of faces.

For most of these functionals, the expectations have been estimated (either approximately or up to a constant factor) for a long time, due to collective results of many researchers (we refer the interested reader to [5, 20] and [15] for surveys). The main open question is thus to understand the distributions of these functionals around their means, as coined by Weil and Wieacker's survey from the Handbook of Convex Geometry (see the concluding paragraph of [20])

> *We finally emphasize that the results described so far give mean values hence first-order information on random sets and point processes. This is due to the geometric nature of the underlying integral geometric results. There are also some less geometric methods to obtain higher-order informations or distributions, but generally the determination of variance, e.g., is a major open problem.*

The last few years have seen several developments in this direction, thanks to new methods and tools from modern probability. Let us first discuss the model K_n where the points are chosen inside K. Reitzner [11], using the Efron–Stein inequality, shows that

$$\operatorname{Var} \mathrm{Vol_d}(K_n) = O(n^{-\frac{d+3}{d+1}}),$$
$$\operatorname{Var} f_i(K_n) = O(n^{\frac{d-1}{d+1}}),$$

where $\mathrm{Vol_d}$ is the standard volume measure on $\mathbb{R}^d$, f_i denotes the number of i-dimensional faces. For convenience, we let $Z = \mathrm{Vol_d}(K_n)$. Using martingale techniques, Vu [18] proves the following tail estimate

$$\mathbb{P}\Big(|Z - \mathbb{E}Z| \geq \sqrt{\lambda n^{-\frac{d+3}{d+1}}}\Big) \leq \exp(-c\lambda) + \exp(-c'n)$$

for any $0 < \lambda < n^\alpha$, where c, c' and α are positive constants. A similar bound also holds for f_i with the same proof. From this tail estimate, one can deduce the above variance bound and also bounds for any fixed moments. These moment bounds are sharp, up to a constant, as shown by Reitzner in [10]. Thus, the order of magnitudes of all fixed moments are determined.

Another topic where a significant development has been made is central limit theorems. It has been conjectured that the key functionals such as the volume and number of faces satisfy a central limit theorem.

Conjecture (CLT conjecture) *Let K_n be the random polytope determined by n random points chosen in K. Then there is a function $\epsilon(n)$ tending to zero with n such that for every x*

$$\left| \mathbb{P}\left(\frac{Z - \mathbb{E}Z}{\sqrt{\operatorname{Var} Z}} \leq x \right) - \Phi(x) \right| \leq \epsilon(n),$$

where Φ denotes the distribution function of the standard normal distribution.

Reitzner [10], using an inequality due to Rinott [14] (which proved a central limit theorem for a sum of weakly dependent random variables), showed that a central limit theorem really holds for the volume and number of faces of the so-called Poisson random polytope. This is a variant of K_n, where the number of random points is not n, but a Poisson random variable with mean n. This model has the advantage that the numbers of points found in disjoint regions of K are independent, a fact which is technically useful. Combining the above tail estimate and Reitzner's result, Vu [19] proved the CLT conjecture.

The above results together provide a fairly comprehensive picture about K_n when the points are chosen inside K. We refer the reader to the last section of [19] for a detailed summary. The main goal of this paper is to provide such a picture for the inscribing model, where points are chosen on the surface of K.

Before we may speak about selecting points on the boundary ∂K, we need to specify the probability measure on ∂K. One wants the random polytope to approximate the original convex body K in the sense that the symmetric difference of the volume of K and K_n is as small as possible. Hence, intuitively, a measure that puts more weight on regions of higher curvature is desired. A good discussion on this can be found in [16]. Let μ_{d-1} be the $(d-1)$-dimensional Hausdorff measure restricted to ∂K. We let μ be a probability measure on ∂K such that

$$d\mu = \rho d\mu_{d-1}, \tag{1}$$

where $\rho : \partial K \to \mathbb{R}_+$ is a positive, continuous function with $\int_{\partial K} \rho d\mu_{d-1} = 1$.

Note that the assumption $\rho > 0$ is essential, as otherwise we might have a measure that causes K_n to always lie in at most half (or any portion) of K with probability 1.

With the boundary measure properly defined, we can choose n random points on the boundary of K independently according to μ_ρ on ∂K. Denote the convex hull of these n points by K_n and we call it *random inscribed polytope*. For this model, the volume is perhaps the most interesting functional (as the number of vertices is always n), and it will be the focus of the present work. For notational convenience, we denote Z for $\mathrm{Vol_d}(K_n)$ throughout this paper.

The inscribing model is somewhat more difficult to analyze than the model where points are chosen inside K. Indeed, sharp estimates on the volume were obtained only recently, thanks to the tremendous effort of Schütt and Werner, in a long and

highly technical paper [16]. We have

$$\mathbb{E}Z = 1 - (c_K + o(1))n^{-\frac{2}{d-1}} \tag{2}$$

where c_K is a constant depending on K (the 1 here represents the volume of K).

It is worth recalling that in the model where points are chosen uniformly inside K it is known that $\mathbb{E}\mathrm{Vol_d}(K - K_n) = O(n^{-\frac{2}{d+1}})$. Observe that by inserting $n^{\frac{d+1}{d-1}}$ for n in this result we obtain a function $O(n^{-\frac{2}{d-1}})$, which is the correct growth rate found in (2). We can explain this (at least intuitively) by noting that in the uniform model, the expected number of vertices is $\Theta(n^{\frac{d-1}{d+1}})$. However, in the inscribing model all points are vertices. Thus we may view the uniform model on n points as yielding the same type of behavior as the inscribing model on $n^{\frac{d-1}{d+1}}$ points. Further evidence for this behavior is given by Reitzner in [12] where he obtains estimates (which are sharp up to a constant factor) for all intrinsic volumes.

Reitzner gives an upper bound on the variance [11]:

$$\mathrm{Var}\, Z = O(n^{-\frac{d+3}{d-1}}).$$

The first result we show in this paper is that the variance estimate is sharp, up to a constant factor.

Theorem 1.1 (Variance) *Given $K \in \mathcal{K}^2_+$,*

$$\mathrm{Var}\, Z = \Omega(n^{-\frac{d+3}{d-1}}),$$

where the implicit constant depends on dimension d and the convex body K only.

The next result in this paper shows that the volume has exponential tail.

Theorem 1.2 (Concentration) *For a given convex body $K \in \mathcal{K}^2_+$, there are positive constants α and c such that the following holds. For any constant $0 < \eta < \frac{d-1}{3d+1}$ and $0 < \lambda \le \frac{\alpha}{4} n^{\frac{d-1}{3d+1} + \frac{2(d+1)\eta}{d-1}} < \frac{\alpha}{4} n$, we have*

$$\mathbb{P}(|Z - \mathbb{E}Z| \ge \sqrt{\lambda V_0}) \le 2\exp(-\lambda/4) + \exp(-cn^{\frac{d-1}{3d+1} - \eta}), \tag{3}$$

where $V_0 = \alpha n^{-\frac{d+3}{d-1}}$.

It is easy to deduce from this theorem the following:

Corollary 1.3 (Moments) *For any given convex body K and $k \ge 2$, the kth moments of Z satisfies*

$$M_k = O((n^{-\frac{d+3}{d-1}})^{k/2}).$$

To emphasize the dependence of $Z = \mathrm{Vol_d} K_n$ on n, we write Z_n instead of Z in the following result:

Corollary 1.4 (Rate of convergence)

$$\lim_{n\to\infty}\left|\left(\frac{Z_n}{\mathbb{E}Z_n}-1\right)f(n)\right|=0$$

almost surely, for

$$f(n)=\delta(n)(n^{-\frac{d+3}{d-1}}\ln n)^{-1/2}$$

where $\delta(n)$ is a function tending to zero arbitrarily slowly as $n\to\infty$.

Finally, we obtain the central limit theorem for the Poisson model. Let $K\in\mathcal{K}^2_+$, and let Pois(n) be a Poisson point process with intensity n. Then the intersection of Pois(n) and ∂K consists of random points $\{t_1,\ldots,t_N\}$ where the number of points N is Poisson distributed with mean $n\mu(\partial K)=n$. We write $\Pi_n=[x_1,\ldots,x_N]$.

Theorem 1.5 *Given $K\in\mathcal{K}^2_+$, we have*

$$\left|\mathbb{P}\left(\frac{\mathrm{Vol_d}(\Pi_n)-\mathbb{E}\mathrm{Vol_d}(\Pi_n)}{\sqrt{\mathrm{Var}\,\mathrm{Vol_d}(\Pi_n)}}\le x\right)-\Phi(x)\right|=o(1),$$

where the $o(1)$ term is of order $O(n^{-\frac{1}{4}}\ln^{\frac{d+2}{d-1}}n)$ as $n\to\infty$.

We hope this result will infer a central limit theorem for K_n, which indeed is the case for random polytopes where the points are chosen inside K, as mentioned earlier (see [10, 19]). However, for random inscribing polytopes, some difficulties remain. We are, however, able to prove that the two models are very close in the sense that the expectations of volume for the two models are asymptotically equivalent, and the variances are only off by constant multiplicative factor (see Theorem 5.5).

In the rest of the paper, we present the proof of the above theorems in Sects. 3, 4, and 5, respectively; Sect. 2 is devoted to notations; we also present proofs of some crucial technical lemmas in the appendix, along with statements of many other lemmas whose proofs can either be found or deduced relatively easily from the literature (see, e.g., [5, 10–12, 18]).

2 Notations

2.1 Geometry

The vectors $e_1,\ldots,e_d$ always represent a fixed orthonormal basis of $\mathbb{R}^d$. The discussions in this paper, unless otherwise specified, are all based on this basis. For a vector x, we denote its coordinate by $x^1,\ldots,x^d$, i.e. $x=(x^1,\ldots,x^d)$. By $B^i(x,r)$ we indicate the i-dimensional Euclidean closed ball of radius r centered at x, i.e.

$$B^i(x,r)=\{y\in\mathbb{R}^i\mid\|x-y\|\le r\}.$$

The norm $\|\cdot\|$ is the Euclidean norm. When the dimension is d, we sometimes simply write $B(x,r)$.

For points $t_1, \ldots, t_n \in \mathbb{R}^d$, the convex hull of them is defined by

$$[t_1, \ldots, t_n] = \left\{ \lambda_1 t_1 + \cdots + \lambda_n t_n \,\middle|\, 0 \le \lambda_i \le 1,\ 1 \le i \le n, \sum_{i=1}^{n} \lambda_i = 1 \right\}.$$

In particular, the closed line segment between two points x and y is

$$[x, y] = \{\lambda x + (1-\lambda)y \mid 0 \le \lambda \le 1\}.$$

To analyze the geometry, it is necessary to introduce the following. For any $y \in \mathbb{R}^d$ write $y = (y^1, \ldots, y^d)$ for the coordinates with respect to some fixed basis $e_1, \ldots, e_d$. For unit vector $u \in \mathbb{R}^d$, let $H(u, h) = \{x \in \mathbb{R}^d \mid \langle x, u\rangle = h\}$, where here $\langle , \rangle$ denotes the standard inner product on $\mathbb{R}^d$. Further, the halfspace associated to this hyperplane we denote by $H^+(u, h) = \{x \in \mathbb{R}^d \mid \langle x, u\rangle \ge h\}$. Since K is smooth, for each point $y \in \partial K$, there is some unique outward normal u_y. We thus may define the *cap* $C = C(y, h)$ of K to be $H^+(u_y, h_K(y) - h) \cap K$, where $h_K(y)$ is the support function such that $H^+(u_y, h_K(y))$ intersects K in the point y only. In general, one should think of a cap as $K \cap H^+$ where H^+ is some closed half space. Throughout this paper, we also use the notion of ϵ-*cap* to emphasize that $\mathrm{Vol_d}(C) = \mathrm{Vol_d}(K \cap H^+) = \epsilon$. Similarly, we call $C = K \cap H^+$ an ϵ-*boundary* cap to emphasize that $\mu(\partial K \cap H^+) = \epsilon$.

We define the ϵ-*wet part* of K to be the union of all caps that are ϵ-boundary caps of K and we denote it by F_ϵ^c. The complement of the ϵ-wet part in K is said to be the ϵ-*floating body* of K, which we denote by F_ϵ. This notion comes from the mental picture that when K is a three dimensional convex body containing ϵ units of water, the floating body is the part that floats above water (see [6]). Finally, consider the floating body F_ϵ and a point $x \in F_\epsilon^c$. We say that x *sees* y if the chord $[x, y]$ does not intersect F_ϵ. Set $S_{x,\epsilon}$ to be the set of those $y \in K$ seen by x. We then define

$$g(\epsilon) = \sup_{x \in F_\epsilon^c} \mathrm{Vol_d}(S_{x,\epsilon}).$$

In particular, we note that $S_{x,\epsilon}$ is the union of all ϵ-boundary caps containing x.

Since K is smooth, it is well known that $g(\epsilon) = \Theta(\mathrm{Vol_d}(\epsilon\text{-boundary cap}))$ (see [6]).

2.2 Asymptotic Notation

We shall always assume n is sufficiently large, without comment. We use the notation Ω, O, Θ etc. with respect to $n \to \infty$, unless otherwise indicated. All constants are assumed to depend on at most the dimension d, the body K, and ρ.

3 Variance

In this section, we provide a proof of Theorem 1.1. It follows an argument first used by Reitzner in [10], which has also been utilized by Bárány and Reitzner [4] to prove a lower bound of the variance in the case where the convex body is a polytope. Essentially, we condition on arrangements of our vertices where they can be perturbed

in such a way that the resulting change in volume is independent for each vertex in question.

Choosing the vertices along the boundary according to a given distribution, as opposed to uniformly in the body, adds technical complication and requires greater use of the boundary structure. The key to the study is the boundary approximation mentioned both in this section and in Appendix 1.

3.1 Small Local Perturbations

We begin by establishing some notation. Define the standard paraboloid E to be

$$E = \{z \in \mathbb{R}^d \mid z^d \geq (z^1)^2 + \cdots + (z^{d-1})^2\}.$$

Hence we have $2E = \{z \in \mathbb{R}^d \mid z^d \geq \frac{1}{2}((z^1)^2 + \cdots + (z^{d-1})^2)\}$ and observe that we have the inclusion

$$E \subset 2E.$$

We now choose a simplex S in the cap $C(0,1)$ of E. Choose the base of the simplex to be a regular simplex with vertices in $\partial E \cap H(e_d, h_d)$ and the origin (h_d to be determined later). We shall denote by $v_0, v_1, \ldots, v_d$ the vertices of this simplex, singling out v_0 to be the apex of S (i.e. the origin). The important point here is that for sufficiently small h_d, the cone $\{\lambda x \in \mathbb{R}^d \mid \lambda \geq 0, x \in S\}$ contains $2E \cap H(e_d, 1)$. Indeed, as the radius of $E \cap H(e_d, h_d)$ is $\sqrt{h_d}$, the inradius of base of the simplex is $\sqrt{h_d/d^2}$, hence for $h_d < 1/2d^2$ our above inclusion holds.

Now, look at the orthogonal projection of the vertices of the simplex to the plane spanned by $\{e_1, \ldots, e_{d-1}\}$, which we think of as $\mathbb{R}^{d-1}$ and denote the relevant operator as

$$\text{proj} : \mathbb{R}^d \to \mathbb{R}^{d-1}.$$

Around the origin we center a ball B_0 of radius r, and around each projected point (except the origin) we can center a ball in $\mathbb{R}^{d-1}$ of radius r', both to be chosen later. We label these balls $B_1, \ldots, B_d$, where B_i is the ball about $\text{proj}(v_i)$. We can form the corresponding sets B_i' to be the inverse image of these sets on ∂E under the projection operator. In other words, if $b : \mathbb{R}^{d-1} \to \mathbb{R}$ is the quadratic form whose graph defines E, $\tilde{b} : \mathbb{R}^{d-1} \to \partial E$ the map induced by b, then

$$B_i' = \tilde{b}(B_i), \quad i = 0, \ldots, d.$$

We note that if we choose r sufficiently small, then for any choice of random points $Y \in B_0'$ and $x_i \in B_i'$, $i = 1, \ldots, d$ the cone on these points is close to the cone on the simplex in the sense that

$$\{\lambda x \mid x \in [Y, x_1, \ldots, x_d],\ \lambda \geq 0\} \supset 2E \cap H(e_d, 1).$$

We may also think of Y being chosen randomly, according to the distribution induced from the $(d-1)$-dimensional Hausdorff measure on E, say. Then, passing to a smaller r if necessary, we see that for any choice of $x_i \in B_i'$, $i = 1, \ldots, d$, we have

$$\text{Var}_Y(\text{Vol}_d([Y, x_1, \ldots, x_d])) \geq c_0 > 0.$$

All the above follows from continuity. We hope results of this type to be true for arbitrary caps of ∂K, and indeed our current construction will serve both model and computational tool for similar constructions on arbitrary caps.

We now consider the general paraboloid

$$Q = \left\{ z \in \mathbb{R}^d \;\middle|\; z^d \geq \frac{1}{2}(k_1(z^1)^2 + \cdots + k_{d-1}(z^{d-1})^2) \right\},$$

where here $k_i > 0$ for all i and let the curvature be $\kappa = \prod k_i$. We now transform the cap $C(0,1)$ of E to the cap $C(0,h)$ of Q by the (unique) linear map A which preserves the coordinate axes. Let D_i be the image of B_i under this affinity. We find that the volume of the D_i scales to give

$$\mu(D_i) = c_1 h^{\frac{d-1}{2}}, \quad i = 1, \ldots, d, \tag{4}$$

where here c_1 is some positive constant only depending on the curvature $\kappa = \prod k_i$ and our choice of r and r'.

Next, for each point $x \in \partial K$ we identify our general paraboloid Q with the approximating paraboloid Q_x of K at x (in particular, we identify $\mathbb{R}^{d-1}$ with the tangent hyperplane at x and the origin with x). We thus write $D_i(x)$ to indicate the set $D_i, i = 1, \ldots, d$, corresponding to Q_x. Analogously to the construction of the $\{B_i'\}$ we can construct the $\{D_i'(x)\}$ as follows. Let $f^x : \mathbb{R}^{d-1} \to \mathbb{R}$ be the function whose graph locally defines ∂K at x (this exists for h sufficiently small, see Lemma 6.1), $\tilde{f} : \mathbb{R}^{d-1} \to \partial K$ the induced function. Let

$$D_i'(x) = \tilde{f}(D_i(x)).$$

We note here that in general the sets $D_i'(x)$ are *not* the images of B_i' under A as $A(B_i')$ may not lie on the boundary ∂K in general.

Because the curvature is bounded above and below by positive constants, as is ρ, we see that the volume of $D_i(x)$ is given by

$$c_3 h^{\frac{d-1}{2}} \leq \mu(D_i(x)) \leq c_4 h^{\frac{d-1}{2}}, \tag{5}$$

where c_3, c_4 are constants depending only on K.

We now wish to get bounds for $\mathrm{Var}_Y(\mathrm{Vol_d}([Y, x_1, \ldots, x_d]))$ where $x_i \in D_i'(x), i = 1, \ldots, d$, and we choose Y randomly in $D_0'(x)$ according to the distribution on the boundary. To begin with, we'll need the following technical lemma.

Lemma 3.1 *There exists a $r_0 > 0$ and r_0' such that for all $r_0 > r > 0$ and $r_0' > r' > 0$ we have an $h_r > 0$ such that for any choice of $x_i \in D_i'(x), i = 1, \ldots, d$, and $h_r > h > 0$:*

$$c_5 h^{d+1} \leq \mathrm{Var}_Y([Y, x_1, \ldots, x_d]) \leq c_6 h^{d+1}, \tag{6}$$

where c_5, c_6 are positive constants depending only on K and r.

The proof of this lemma is given in Appendix 1. Assuming this lemma is true, we proceed with our analysis as follows.

Fix some choice for $h_d < 1/2d^2$. Let $v_0, \ldots, v_d$ denote the vertices of the simplex S. Then by continuity we know that there is some $\eta > 0$ such that choosing x_i in η-balls $B(v_i, \eta)$ centered at the vertices preserves our desired inclusion, namely

$$\{\lambda x \mid x \in [x_0, x_1, \ldots, x_d], \lambda \geq 0\} \supset 2E \cap H(e_d, 1). \tag{7}$$

We now desire to set $r' > 0$ such that $D_i'(x) \subset A(B(v_i, \eta))$ for all $x \in \partial K$. As a consequence, we will obtain the inclusion, for $x_i \in D_i'(x)$,

$$\{\lambda x \mid x \in [x_0, x_1, \ldots, x_d], \lambda \geq 0\} \supset 2Q_x \cap H(u_x, h) \supset K \cap H(u_x, h).$$

Choose $\epsilon > 0$ such that

$$\begin{aligned} U_i = \{(x, y) \in \mathbb{R}^d \mid x \in B(\text{proj}v_i, \eta/2) \subset \mathbb{R}^{d-1} \text{ and} \\ (1+\epsilon)^{-1} b_E(x) \leq y \leq (1+\epsilon) b_E(x)\} \subset B(v_i, \eta) \end{aligned} \tag{8}$$

for each i, where b_E is the quadratic form defining our standard paraboloid E. Appealing to Lemma 6.1 we take h sufficiently small such that for all $x \in \partial K$,

$$(1+\epsilon)^{-1} b_x(y) \leq f_x(y) \leq (1+\epsilon) b_x(y).$$

Choosing $r' < \eta/2$ forces the B_i to be balls of radius r' about $\text{proj}v_i$, which by the above causes $D_i'(x) \subset A(U_i) \subset A(B(v_i, \eta))$.

With these choices for r, r' and some constant $h_0 > 0$ to enforce the condition that h is sufficiently small above, we now proceed to the body of our argument.

3.2 Proof of Lower Bound on Variance

Choose n points $t_1, \ldots, t_n$ randomly in ∂K according to the probability induced by the distribution. Choose n points $y_1, \ldots, y_n \in \partial K$ and corresponding disjoint caps according to Lemma 6.6. In each cap $C(y_j, h_n)$ (of K) establish sets $\{D_i(y_j)\}$ and $\{D_i'(y_j)\}$ for $i = 0, \ldots, d$ and $j = 1, \ldots, n$ as in the above discussion.

We let A_j, $j = 1, \ldots, n$ be the event that exactly one random point is contained in each of the $D_i(y_j)$, $i = 0, \ldots, d$ and every other point is outside $C(y_j, h_n) \cap \partial K$. We calculate the probability as

$$\begin{aligned} P(A_j) &= n(n-1)\cdots(n-d)\mathbb{P}(t_i \in D_i'(y_j), i = 0, \ldots, d) \\ &\quad \times \mathbb{P}(t_i \notin C(y_j, h_n) \cap \partial K, i \geq d+1) \\ &= n(n-1)\cdots(n-d) \prod_{i=0}^{d} \mu(D_i'(y_j)) \prod_{k=d+1}^{n} (1 - \mu(C(y_j, h_n) \cap \partial K)). \end{aligned}$$

We can give a lower bound for this quantity with (5) and Lemma 6.6 , and noting specifically that $h_n = \Theta(n^{-2/(d-1)})$:

$$\mathbb{P}(A_j) \geq c_7 n^{d+1} n^{-d-1} (1 - c_8 n^{-1})^{n-d-1} \geq c_9 > 0, \tag{9}$$

where c_7, c_8, c_9 are positive constants. In particular, denoting by $\mathbf{1}_A$ the indicator function of event A. We obtain that

$$\mathbb{E}\left(\sum_{j=1}^{n} \mathbf{1}_{A_j}\right) = \sum_{j=1}^{n} \mathbb{P}(A_j) \geq c_9 n. \tag{10}$$

Now we denote by $\mathcal{F}$ the position of all points of $\{t_1, \ldots, t_n\}$ except those which are contained in $D_0'(y_j)$ with $\mathbf{1}_{A_j} = 1$. We then use the conditional variance formula to obtain a lower bound:

$$\operatorname{Var} Z = \mathbb{E} \operatorname{Var}(Z \mid \mathcal{F}) + \operatorname{Var} \mathbb{E}(Z \mid \mathcal{F}) \geq \mathbb{E} \operatorname{Var}(Z \mid \mathcal{F}).$$

Now we look at the case where $\mathbf{1}_{A_j}$ and $\mathbf{1}_{A_k}$ are both 1 for some $j, k \in \{1, \ldots, n\}$. Assume without loss of generality that t_j and t_k are the points in $D_0'(y_j)$ and $D_0'(y_k)$, respectively. We note that by construction there can be no edge between t_j and t_k, so the volume change affected by moving t_j within $D_0'(y_j)$ is independent of the volume change of moving t_k within $D_0'(y_k)$. This independence allows us to write the conditional variance as the sum

$$\operatorname{Var}(Z|\mathcal{F}) = \sum_{j=1}^{n} \operatorname{Var}_{t_j}(Z) \mathbf{1}_{A_j},$$

where here each variance is taken over $t_j \in D_0'(y_j)$. We now invoke Lemma 3.1, equation (10), and the bound $h_n \approx n^{-2/(d-1)}$ to compute

$$\begin{aligned} \mathbb{E} \operatorname{Var}(Z \mid \mathcal{F}) &= \mathbb{E}\left(\sum_{j=1}^{n} \operatorname{Var}_{t_j}(Z) \mathbf{1}_{A_j}\right) \geq c_5 h^{d+1} \mathbb{E}\left(\sum_{j=1}^{n} \mathbf{1}_{A_j}\right) \\ &\geq c_{10} (n^{-2/(d-1)})^{d+1} c_6 n = c_{11} n^{-(d+3)/(d-1)}. \end{aligned}$$

Thus, the above provides the promised lower bound on $\operatorname{Var} Z$.

4 Concentration

Our concentration result shows that $\operatorname{Vol}_d(K_n)$ is highly concentrated about its mean. Namely, we obtain a bound of the form

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq \sqrt{\lambda \operatorname{Var} Z}\,) \leq c_1 \exp(-c_2 \lambda) \tag{11}$$

for positive constants c_1, c_2. Such an inequality indicates that Z has an exponential tail, which proves sufficient to provide information about the higher moments of Z and the rate of convergence of Z to its mean.

4.1 Discrete Geometry

We now set up some basic geometry which will be the subject of our analysis. Let L be a finite collection of points. For a point $x \in K$, define

$$\Delta_{x,L} = \mathrm{Vol_d}([L \cup x]) - \mathrm{Vol_d}([L]).$$

A key property is the following observation.

Lemma 4.1 *Let L be a set whose convex hull contains the floating body F_ϵ. Then for any $x \in K$,*

$$\Delta_{x,L} \leq g(\epsilon).$$

The major geometry result which allows for our analysis is the following lemma quantifying the fact that K_n contains the floating body F_ϵ with high probability.

Lemma 4.2 *There are positive constants c and c' such that the following holds for every sufficiently large n. For any $\epsilon \geq c' \ln n/n$, the probability that K_n does not contain F_ϵ is at most* $\exp(-c\epsilon n)$.

The proof of this result can be done using the notion of VC-dimension, similar details of which can be found in [18].

4.2 A Slightly Weaker Result

The proof of Theorem 1.2 is rather technical. So we will first attempt a simpler one of a slightly weaker result, which represents one of the main methodology used in this paper.

Put $G_0 = 3g(\epsilon)$ and $V_0 = 36ng(\epsilon)^2$, where $g(\epsilon)$ is as defined in the previous subsection. We show:

Theorem 4.3 *For a given $K \in \mathcal{K}_+^2$ there are positive constants α, c, and ϵ_0 such that the following holds: for any $\alpha \ln n/n < \epsilon \leq \epsilon_0$ and $0 < \lambda \leq V_0/4G_0^2$, we have*

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq \sqrt{\lambda V_0}) \leq 2\exp(-\lambda/4) + \exp(-c\epsilon n).$$

We note that the constants used in the definition of G_0 and V_0 are chosen for convenience and can be optimized, though we make no effort to do so.

To compare Theorem 4.3 with Theorem 1.2, we first compute V_0. $V_0 = 36ng(\epsilon)^2 = \Theta(\epsilon^{(d+1)/(d-1)})$, from definition of $g(\epsilon)$ and by Lemma 6.2. So, setting $\epsilon = \alpha \ln n/n$ for some positive constant c satisfying Lemma 6.2 and greater than a given α gives

$$\begin{aligned} V_0 &= 36ng(\epsilon)^2 = 36n\Theta(\epsilon^{(d+1)/(d-1)})^2 = \Theta(nn^{-2(d+1)/(d-1)}(\ln n)^{2(d+1)/(d-1)}) \\ &= \Theta(n^{-(d+3)/(d-1)}(\ln n)^{2(d+1)/(d-1)}). \end{aligned} \tag{12}$$

So, up to a logarithmic factor V_0 is comparable to $\mathrm{Var}\, Z$.

To obtain Theorem 1.2 we utilize a martingale inequality (Lemma 4.4). This inequality, which is a generalization of an earlier result of Kim and Vu [9], appears to be a new and powerful tool in the study of random polytopes. It was first used by Vu in [18], and seems to provide a very general framework for the study of key functionals. The reader who is familiar with other martingale inequalities, most notably that of Azuma [2], will be familiar with the general technique (see also [1]).

Recall $K_n = [t_1, \ldots, t_n]$, where $t_i, i = 1, \ldots, n$, are independent random points in ∂K. Let the sample space be $\Omega = \{t \mid t = (t_1, \ldots, t_n), t_i \in \partial K\}$ and let $Z = Z(t_1, \ldots, t_n) = \mathrm{Vol_d}(K_n)$ a function of these points, we may define the *(absolute) martingale difference sequence*

$$G_i(t) = |\mathbb{E}(Z \mid t_1, \ldots, t_{i-1}, t_i) - \mathbb{E}(Z \mid t_1, \ldots, t_{i-1})|.$$

Thus, $G_i(t)$ is a function of $t = (t_1, \ldots, t_n)$ that only depends on the first i points. We then set

$$V_i(t) = \int G_i^2(t)\partial t_i, \qquad V(t) = \sum_{i=1}^{n} V_i(t),$$

$$G_i'(t) = \sup_{t_i} G_i(t) \quad \text{and} \quad G(t) = \max_i G_i'(t).$$

Note also that $|Z - \mathbb{E}Z| \le \sum_i G_i$. The key to our proof is the following concentration lemma, which was derived using the so-called divide-and-conquer martingale technique (see [18]).

Lemma 4.4 *For any positive λ, G_0 and V_0 satisfying $\lambda \le V_0/4G_0^2$, we have*

$$\mathbb{P}(|Z - \mathbb{E}Z| \ge \sqrt{\lambda V_0}) \le 2\exp(-\lambda/4) + \mathbb{P}(V(t) \ge V_0 \text{ or } G(t) \ge G_0). \tag{13}$$

The proof of this lemma can be found in [18].

Comparing Lemma 4.4 to Theorem 4.3 we find that the technical difficulty comes in bounding the term $\mathbb{P}(V(t) \ge V_0$ or $G(t) \ge G_0)$, which corresponds to the error term p_{NT}.

Set $V' = n^{-1}V_0 = 36g(\epsilon)^2$. We find that we can replace $\exp(-c\epsilon n)$ with $n\exp(-c'\epsilon n)$ by adjusting the relevant constant c' so that $n\exp(-c'\epsilon n) < \exp(-c\epsilon n)$. Thus, we're going to prove that

$$\mathbb{P}(G(t) \ge G_0 \text{ or } V(t) \ge V_0) \le n\exp(-c\epsilon n)$$

for some positive constant c.

To do this, we'll prove the following claim.

Claim 4.5 *There is a positive constant c such that for any $1 \le i \le n$,*

$$\mathbb{P}(G_i'(t) \ge G_0 \text{ or } V_i(t) \ge V') \le \exp(-c\epsilon n).$$

From this claim the trivial union bound gives

$$\mathbb{P}(G(t) \ge G_0 \text{ or } V(t) \ge V_0) \le n\exp(-c\epsilon n),$$

hence quoting Lemma 4.4 finishes our proof of Theorem 4.3.

4.3 Proof of Claim 4.5

Recall that $Z = Z(t_1, \ldots, t_n) = \mathrm{Vol_d}(K_n)$ for points $t_i \in \partial K$.

The triangle inequality gives us

$$\begin{aligned} G_i(t) &= |\mathbb{E}(Z \mid t_1, \ldots, t_{i-1}, t_i) - \mathbb{E}(Z \mid t_1, \ldots, t_{i-1})| \\ &\le \mathbb{E}_x \big|\mathbb{E}(Z \mid t_1, \ldots, t_{i-1}, t_i) - \mathbb{E}(Z \mid t_1, \ldots, t_{i-1}, x)\big|, \end{aligned}$$

where $\mathbb{E}_x$ denotes the expectation over a random point x. The analysis for the two terms in the last inequality is similar, so we will estimate the first one. Let us fix (arbitrarily) $t_1, \ldots, t_{i-1}$. Let L be the union of $\{t_1, \ldots, t_{i-1}\}$ and the random set of points $\{t_{i+1}, \ldots, t_n\}$. Since

$$\mathrm{Vol_d}([L \cup t_i]) = \mathrm{Vol_d}([L]) + \Delta_{t_i, L},$$

we have

$$\mathbb{E}(Z \mid t_1, \ldots, t_{i-1}, t_i) = \mathbb{E}(\mathrm{Vol_d}([L]) \mid t_1, \ldots, t_{i-1}) + \mathbb{E}(\Delta_{t_i, L} \mid t_1, \ldots, t_{i-1}).$$

The key inequality of the analysis is the following:

$$\mathbb{E}(\Delta_{t_i, L} \mid t_1, \ldots, t_{i-1}) \le \mathbb{P}(F_\epsilon \nsubseteq [L] \mid t_1, \ldots, t_{i-1}) + g(\epsilon). \tag{14}$$

The inequality (14) follows from two observations:

- If $F_\epsilon \nsubseteq [L]$, $\Delta_{t_i, L}$ is at most 1.
- If $[L]$ contains F_ϵ, $\Delta_{t_i, L} \le g(\epsilon)$ by Lemma 4.1.

We denote by $\Omega_{(j)}$ and $\Omega^{<j>}$ the spaces spanned by $\{t_1, \ldots, t_j\}$ and $\{t_j, \ldots, t_n\}$, respectively.

Set $\delta = n^{-4}$. We say that the set $\{t_1, \ldots, t_{i-1}\}$ is *typical* if

$$\mathbb{P}_{\Omega^{(i+1)}}(F_\epsilon \subseteq [L] \mid t_1, \ldots, t_{i-1}) \ge 1 - \delta.$$

The rest of the proof has two steps. In the first step, we show that if $\{t_1, \ldots, t_{i-1}\}$ is typical then $G_i'(t) \le G_0$ and $V_i(t) \le V'$. In the second step, we bound the probability that $\{t_1, \ldots, t_{i-1}\}$ is not typical.

First step. Assume that $\{t_1, \ldots, t_{i-1}\}$ is typical, so $\mathbb{P}_{\Omega^{<i+1>}}(F_\epsilon \nsubseteq [L] \mid t_1, \ldots, t_{i-1}) \le \delta = n^{-4}$. We first bound $G_i'(t)$. Observe that

$$\begin{aligned} G_i(t) &\le \mathbb{E}_x |\mathbb{E}(Z \mid t_1, \ldots, t_{i-1}, t_i) - \mathbb{E}(Z \mid t_1, \ldots, t_{i-1}, x)| \\ &\le \mathbb{E}_x |\mathbb{E}(\Delta_{t_i, L} \mid t_1, \ldots, t_{i-1}) - \mathbb{E}(\Delta_{x, L} \mid t_1, \ldots, t_{i-1})| \\ &\le \mathbb{E}(\Delta_{t_i, L} \mid t_1, \ldots, t_{i-1}) + \mathbb{E}_x \mathbb{E}(\Delta_{x, L} \mid t_1, \ldots, t_{i-1}) \\ &\le 2g(\epsilon) + 2n^{-4} \le 3g(\epsilon) = G_0 \quad \text{(by (14))}. \end{aligned}$$

In the last inequality we use the fact that $\epsilon = \Omega(\ln n / n)$, $g(\epsilon) = \Omega(\epsilon^{(d+1)/(d-1)}) \gg n^{-4}$. Thus it follows that

$$G_i'(t) = \max_{t_i} G_i(t) \le G_0.$$

Calculating $V_i(t)$ using the above bound on $G_i(t)$ it follows that

$$V_i(t) = \int G_i(t)^2 d\mu(t_i) \leq \int 9g(\epsilon)^2 d\mu(t_i) = 9g(\epsilon)^2 < V'.$$

Second step. In this step, we bound the probability that $\{t_1, \ldots, t_{i-1}\}$ is not typical. First of all, we will need a technical lemma as follows. Let Ω' and Ω'' be probability spaces and set Ω''' to be their product. Let A be an event in Ω''' which occurs with probability at least $1 - \delta'$, for some $0 < \delta' < 1$.

Lemma 4.6 *For any* $1 > \delta > \delta'$

$$\mathbb{P}_{\Omega'}(\mathbb{P}_{\Omega''}(A \mid x) \leq 1 - \delta) \leq \delta'/\delta,$$

where x is a random point in Ω' and $\mathbb{P}_{\Omega'}$ and $\mathbb{P}_{\Omega''}$ are the probabilities over Ω' and Ω'', respectively.

Proof Recall that $\mathbb{P}_{\Omega'''}(A) \geq 1 - \delta'$. However,

$$\mathbb{P}_{\Omega'''}(A) = \int_{\Omega'} \mathbb{P}_{\Omega''}(A \mid x) \partial x \leq 1 - \delta \mathbb{P}_{\Omega'}(\mathbb{P}_{\Omega''}(A \mid x) \leq 1 - \delta).$$

The claim follows. □

Recall that $L = \{t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n\}$. Lemma 4.2 yields

$$\mathbb{P}(F_\epsilon \nsubseteq [L]) \leq \exp(-c_0 \epsilon n),$$

for some positive constant c_0 depending only on K. Applying Lemma 4.6 with $\Omega' = \Omega_{(i-1)}$, $\Omega'' = \Omega^{\langle i+1 \rangle}$, $\delta' = \exp(-c\epsilon n)$ and $\delta = n^{-4}$, we have

$$\begin{aligned}
&\mathbb{P}_{\Omega_{(i-1)}}(\{t_1, \ldots, t_{i-1}\} \text{ is not typical}) \\
&\quad = \mathbb{P}_{\Omega_{(i-1)}}(\mathbb{P}_{\Omega^{\langle i+1 \rangle}}(F_\epsilon \nsubseteq [L] \mid t_1, \ldots, t_{i-1}) \leq 1 - \delta) \\
&\quad \leq \delta'/\delta = n^4 \exp(-c_0 \epsilon n) \leq \exp(-c\epsilon n)
\end{aligned}$$

for $c = c_0/2$, given $c_0 \epsilon n \geq 8 \ln n$. This final condition can be satisfied by setting the α involved in the lower bound of ϵ to be sufficiently large. Thus, our proof is complete.

4.4 A Better Bound on Deviation

By using more of the smooth boundary structure, we can obtain a better result. As we shall see at the end of the proof, this result implies Theorem 1.2.

Theorem 4.7 *For any smooth convex body K with distribution μ along the boundary, there are constants $c, c', \alpha, \epsilon_0$ such that the following holds. For any $V_0 \geq$*

$\alpha n^{-(d+3)/(d-1)}$, $\epsilon_0 \geq \epsilon > \alpha \ln n/n$, $G_0 \geq 3\epsilon^{(d+1)/(d-1)}$, *and* $0 < \lambda \leq V_0/4G_0^2$, *we have*

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq \sqrt{\lambda V_0}) \leq 2\exp(-\lambda/4) + p_{NT},$$

where

$$p_{NT} = \exp(-c\epsilon n) + \exp(-c' n^{\frac{d-1}{3d+1}-\eta}),$$

and η *is any small positive constant less than* $\frac{d-1}{3d+1}$.

The proof of Theorem 4.7 follows from more careful estimates concerning $\Delta_{x,L}$. An analogous result for random polytopes can be found in Sect. 2.5 of [18].

The key difference between this result and Theorem 4.3 is that here V_0 is independent of ϵ, so we can set $V_0 = \alpha n^{-(d+3)/(d-1)}$ without affecting the tail estimate. If we also set $\epsilon = n^{-\frac{2d+2}{3d+1}-\eta}$, then the two error terms in p_{NT} are the same (up to a constant factor). Since $G_0 = 3g(\epsilon) = 3\Theta(\epsilon^{(d+1)/(d-1)})$, we have $\lambda < V_0/4G_0^2 \leq c'' n^{\frac{d-1}{3d+1}+\frac{2(d+1)\eta}{d-1}}$ for some constant c''. Hence Theorem 1.2.

5 Central Limit Theorem

5.1 Poisson Central Limit Theorem

Before we prove the theorem, we should give a brief review of the Poisson point process. Let $K \in \mathcal{K}^2_+$, and let Pois(n) be a Poisson point process with intensity n concentrated on K. Then applying Pois(n) on K gives us random points $\{x_1, \ldots, x_N\}$ where the number of points N is Poisson distributed with intensity $n\mu(\partial K) = n$. We write $\Pi_n = [x_1, \ldots, x_N]$. Conditioning on N, the points $x_1, \ldots, x_N$ are independently uniformly distributed in ∂K. For two disjoint subsets A and B of ∂K, their intersections with Pois(n), i.e. the point sets $A \cap \text{Pois}(n) = \{x_1, \ldots, x_N\}$ and $B \cap \text{Pois}(n) = \{y_1, \ldots, y_M\}$, are independent. This means N and M are independently Poisson distributed with intensity $n\mu(A)$ and $n\mu(B)$ respectively, and x_i and y_j are chosen independently.

The following standard estimates of the tail of Poisson distribution will be used repeatedly throughout this section. Let X be a Poisson random variable with mean λ. Then

$$\begin{aligned}\mathbb{P}\left(X \leq \frac{\lambda}{2}\right) &= \sum_{k=0}^{\lambda/2} e^{-\lambda}\frac{\lambda^k}{k!} \leq e^{-\lambda} + \sum_{k=1}^{\lambda/2} e^{-\lambda}\left(\frac{e\lambda}{k}\right)^k \\ &\leq \frac{\lambda+1}{2}e^{-\lambda}(2e)^{\lambda/2} \leq \frac{\lambda+1}{2}\left(\frac{e}{2}\right)^{-\lambda/2} = \Theta\left(\left(\frac{e}{2}\right)^{-\lambda/2}\right), \quad (15)\end{aligned}$$

where the last equality holds when λ is large. Similarly,

$$\mathbb{P}(X \geq 3\lambda) \leq \sum_{k=3\lambda}^{\infty} e^{-\lambda}\left(\frac{e\lambda}{k}\right)^k \leq \sum_{k=0}^{\infty} e^{-\lambda}\left(\frac{e}{3}\right)^k = ce^{-\lambda}, \quad (16)$$

where c is a small constant.

The key ingredient of the proof is the following theorem:

Theorem 5.1 (Baldi and Rinott [3]) *Let G be the dependency graph of random variables Y_i's, $i = 1, \ldots, m$, and let $Y = \sum_i Y_i$. Suppose the maximal degree of G is D and $|Y_i| \le B$ a.s., then*

$$\left| \mathbb{P}\left(\frac{Y - \mathbb{E}Y}{\sqrt{\operatorname{Var} Y}} \le x \right) - \Phi(x) \right| = O(\sqrt{S}),$$

where $\Phi(x)$ is the standard normal distribution and $S = \frac{mD^2B^3}{(\sqrt{\operatorname{Var} Y})^3}$.

Here the dependency graph of random variables Y_i's is a graph on m vertices such that there is no edge between any two disjoint subsets, A_1 and A_2, of $\{Y_i\}_{i=1}^m$ if these two sets of random variables are independent.

Because we can divide the convex body K into Voronoi cells according to the cap covering Lemma 6.6, we will study $\mathrm{Vol_d}(\Pi_n)$ as a sum of random variables which are volumes of the intersection of Π_n with each of the Voronoi cell. And the theorem above allows us to prove central limit theorem for sums of random variables that may have small dependency on each other.

First we let

$$m = \left\lfloor \frac{n}{4d \ln n} \right\rfloor.$$

By Lemma 6.6, given $K \in \mathcal{K}^2_+$, we can choose m points, namely $y_1, \ldots, y_m$, on ∂K. And the Voronoi cells $\mathrm{Vor}(y_i)$ of these points dissect K into m parts. Let

$$Y_i = \mathrm{Vol_d}(\mathrm{Vor}(y_i) \cap K) - \mathrm{Vol_d}(\mathrm{Vor}(y_i) \cap \Pi_n),$$

$i = 1, \ldots, m$. So

$$Y = \sum_i Y_i = \mathrm{Vol_d}(K) - \mathrm{Vol_d}(\Pi_n). \tag{17}$$

Moreover, these Voronoi cells also dissect the boundary of K into m parts, and each contains a cap C_i with d-dimensional volume

$$\mathrm{Vol_d}(C_i) = \Theta(m^{-\frac{d+1}{d-1}}),$$

by Lemma 6.6. Now by Lemma 6.2 it is a boundary cap with $(d-1)$-dimensional volume

$$\mu(C_i \cap \partial K) = \Theta(m^{-1}) = \Theta\left(\frac{4d \ln n}{n} \right).$$

Denote by A_i $(i = 1, \ldots, m)$ the number of points generated by the Poisson point process of intensity n contained in $C_i \cap \partial K$, hence A_i is Poisson distributed with mean $\lambda = n\mu(C_i \cap \partial K) = \Theta(4d \ln n)$. Then

$$\mathbb{P}(A_i = 0) = e^{-\lambda} = O(n^{-4d}).$$

And by (15),

$$\mathbb{P}(A_i \geq 3\lambda) = \mathbb{P}(A_i \geq 12d \ln n) = O(n^{-4d}).$$

Now let A^m be the event that there is at least one point and at most $12d \ln n$ points in every A_i for $i = 1, \ldots, m$. Then

$$1 \geq \mathbb{P}(A^m) = \mathbb{P}(\cap_i \{1 \leq A_i \leq 12d \ln n\}) \geq 1 - \Omega(n^{-4d+1}). \tag{18}$$

The rest of the proof is organized as follows. We first prove the central limit theorem for $\mathrm{Vol_d}(\Pi_n)$ when we condition on A^m, then we show removing the condition doesn't affect the estimate much, as A^m holds almost surely. Let $\widetilde{\mathbb{P}}$ denote the conditional probability measure induced by the Poisson point process $X(n)$ on ∂K given A^m, i.e.

$$\widetilde{\mathbb{P}}(\mathrm{Vol_d}(\Pi_n) \leq x) = \mathbb{P}(\mathrm{Vol_d}(\Pi_n) \leq x | A^m).$$

Similarly, we define the corresponding conditional expectation and variance to be $\widetilde{\mathbb{E}}$ and $\widetilde{\mathrm{Var}}$, then

Lemma 5.2

$$\left| \widetilde{\mathbb{P}}\left(\frac{\mathrm{Vol_d}(\Pi_n) - \widetilde{\mathbb{E}}\mathrm{Vol_d}(\Pi_n)}{\sqrt{\widetilde{\mathrm{Var}}\mathrm{Vol_d}(\Pi_n)}} \leq x \right) - \Phi(x) \right| = O(n^{-\frac{d+1}{4(d-1)}} \ln^{\frac{d+2}{d-1}} n). \tag{19}$$

Proof Note that by (17), $\mathrm{Vol_d}(\Pi_n) - \widetilde{\mathbb{E}}\mathrm{Vol_d}(\Pi_n) = \widetilde{\mathbb{E}}Y - Y$, and $\widetilde{\mathrm{Var}}Y = \widetilde{\mathrm{Var}}\mathrm{Vol_d}(\Pi_n) = \Theta(n^{-\frac{d+3}{d-1}})$, by Theorem 5.5. Hence it suffices to show Y satisfies the Central Limit Theorem under $\widetilde{\mathbb{P}}$.

Given A^m, we define the dependency graph on random variables $Y_i, i = 1, \ldots, m$ as follows: we connect Y_i and Y_j if $\mathrm{Vor}(y_i) \cap C(y_j, c, m^{-\frac{2}{d-1}}) \neq \emptyset$ for some constant c which satisfies Lemma 6.8. To check dependency, we see that if $Y_i \nsim Y_j$, then $\mathrm{Vor}(y_i) \cap C(y_j, c, m^{-\frac{2}{d-1}}) = \emptyset$. Thus, for any point $P_1 \in \mathrm{Vor}(y_i) \cap \partial K$, $P_2 \in \mathrm{Vor}(y_j) \cap \partial K$, the line segment $[P_1, P_2]$ cannot be contained in the boundary of Π_n. Otherwise, it would be a contradiction to Lemma 6.8. Therefore, there is no edge of Π_n between vertices in $\mathrm{Vor}(y_i)$ and $\mathrm{Vor}(y_j)$, hence Y_i and Y_j are independent given A^m.

To apply Theorem 5.1 to Y, we are left to estimate parameters D and B.

By Lemma 6.7, $C(y_i, c, m^{-\frac{2}{d-1}})$ $(i = 1, \ldots, m)$ can intersect at most $O(1)$ many $\mathrm{Vor}(y_i)$'s. Hence $D = O(1)$.

By Lemma 6.8, for any point x_i in C_i, $i = 1, \ldots, m$,

$$\delta^H(K, \Pi_n) \leq \delta^H(K, [x_1, \ldots, x_m]) = O(m^{-\frac{2}{d-1}}).$$

So

$$\mathrm{Vor}(y_i) \backslash \Pi_n \subseteq C(y_i, h'), \tag{20}$$

where $h' = O(m^{-\frac{2}{d-1}})$. By Lemma 6.5 and (20),

$$Y_i \le \mathrm{Vol_d}(C(y_i, h')) = O(m^{-\frac{d+1}{d-1}}) = O\left(\left(\frac{4d\ln n}{n}\right)^{\frac{d+1}{d-1}}\right) := B.$$

Hence by the Baldi–Rinott Theorem, the rate of convergence in (19) is $n^{-\frac{d+1}{4(d-1)}} \times (\ln n)^{\frac{d+2}{d-1}}$, and we finish the proof. □

Now, we will remove the condition A^m. First observe an easy fact.

Proposition 5.3 *For any events A and B,*

$$|\mathbb{P}(B \mid A) - \mathbb{P}(B)| \le \mathbb{P}(A^c).$$

Hence we can deduce:

Lemma 5.4

$$|\widetilde{\mathbb{P}}(\mathrm{Vol_d}(\Pi_n) \le x) - \mathbb{P}(\mathrm{Vol_d}(\Pi_n) \le x)| = O(n^{-4d+1}), \tag{21}$$

$$|\widetilde{\mathbb{E}}\mathrm{Vol_d^k}(\Pi_n) - \mathbb{E}\mathrm{Vol_d^k}(\Pi_n)| = O(n^{-4d+1}), \tag{22}$$

$$|\widetilde{\mathrm{Var}}\mathrm{Vol_d}(\Pi_n)) - \mathrm{Var}\,\mathrm{Vol_d}(\Pi_n)| = O(n^{-4d+1}). \tag{23}$$

The proofs of these three equations follow more or less from Proposition 5.3 with $\mathbb{P}((A^m)^c) = O(n^{-4d+1})$, and can be found in [10]. As a result of Lemma 5.4, we can remove the condition A^m and obtain Theorem 1.5 as follows. For notational convenience, we denote $\mathrm{Vol_d}(\Pi_n)$ by X temporarily. For each x, let $\tilde{x}$ be such that

$$\mathbb{E}X + x\sqrt{\mathrm{Var}\,X} = \widetilde{\mathbb{E}}X + \tilde{x}\sqrt{\widetilde{\mathrm{Var}}X},$$

then

$$|x - \tilde{x}| = O(n^{-4d+1+\frac{d+3}{2(d-1)}}) + |x|O(n^{-4d+1+\frac{d+3}{d-1}}), \tag{24}$$

by (21) and Lemma 5.2. We have

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \le \mathbb{E}X + x\sqrt{\mathrm{Var}\,X}) = \widetilde{\mathbb{P}}(X \le \widetilde{\mathbb{E}}X + \tilde{x}\sqrt{\widetilde{\mathrm{Var}}}) + O(n^{-4d+1}) \\ &= \Phi(\tilde{x}) + O(n^{-\frac{d+1}{4(d-1)}}\ln^{\frac{d+2}{d-1}} n) + O(n^{-4d+1}). \end{aligned}$$

But $|\Phi(x) - \Phi(\tilde{x})| = O(n^{-1})$, since $|\Phi(x) - \Phi(\tilde{x})| \le |x - \tilde{x}| \le O(n^{-1})$ when $|x| \le n$ and by (24) $|\tilde{x}| \ge cn$ when $|x| \ge n$ which implies $|\Phi(x) - \Phi(\tilde{x})| \le \Phi(n) + \Phi(cn)$. So $|F_X(x) - \Phi(x)| = |\mathbb{P}(X \le \mathbb{E}X + x\sqrt{\mathrm{Var}\,X}) - \Phi(x)| = O(n^{-\frac{d+1}{4(d-1)}}\ln^{\frac{d+2}{d-1}} n)$. Hence finishes the proof of Theorem 1.5.

5.2 Approximating K_n by Π_n

As is pointed out in the introduction, Π_n approximates K_n quite well, as one might expect.

Theorem 5.5 *Let Π_n be the convex hull of points chosen on ∂K according to the Poisson point process* Pois(n). *Then*,

$$\mathbb{E}\mathrm{Vol_d}(\Pi_n) \approx \mathbb{E}\mathrm{Vol_d}(K_n) \approx 1 - c(K,d)n^{-\frac{2}{d-1}},$$

as $n \to \infty$, and

$$\mathrm{Var}\,\mathrm{Vol_d}(\Pi_n) = \Theta(\mathrm{Var}\,\mathrm{Vol_d}(K_n)) = \Theta(n^{-\frac{d+3}{d-1}}).$$

Proof Due to the conditioning property of Poisson point process, we have

$$\mathbb{E}\mathrm{Vol_d}(\Pi_n) = \sum_{|k-n|\le n^{7/8}} \mathbb{E}\mathrm{Vol_d}(K_k)e^{-n}\frac{n^k}{k!} + \sum_{|k-n|\ge n^{7/8}} \mathbb{E}\mathrm{Vol_d}(K_k)e^{-n}\frac{n^k}{k!}.$$

For Poisson distribution, the Chebyschev's inequality gives $\mathbb{P}(|k-n| \ge n^{7/8}) \le n^{-3/4}$. Hence the second summand is bounded above by $n^{-3/4}$ since $\mathbb{E}\mathrm{Vol_d}(K_k)$ is at most 1. By 2, $\mathbb{E}\mathrm{Vol_d}(K_k) = 1 - k^{-\frac{2}{d-1}} = 1 - (1+o(1))n^{-\frac{2}{d-1}}$, when $|k-n| \le n^{7/8}$.

For the variance, we can rewrite $\mathrm{Var}\,\mathrm{Vol_d}(\Pi_n)$ as follows:

$$\mathrm{Var}\,\mathrm{Vol_d}(\Pi_n) = \mathbb{E}_N\,\mathrm{Var}(\mathrm{Vol_d}(\Pi_n) \mid N) + \mathrm{Var}_N\,\mathbb{E}(\mathrm{Vol_d}(\Pi_n) \mid N).$$

By (15), the second term in the above equation becomes:

$$\begin{aligned}
&\mathrm{Var}\,\mathbb{E}(\mathrm{Vol_d}(\Pi_n) \mid N)\\
&= \mathbb{E}_N\mathbb{E}^2\mathrm{Vol_d}(K_N) - (\mathbb{E}_N\mathbb{E}\mathrm{Vol_d}(K_N))^2\\
&= \sum_{j=\frac{n}{2}}^{\infty}\sum_{k=\frac{n}{2}}^{\infty}(\mathbb{E}^2\mathrm{Vol_d}(K_k) - \mathbb{E}\mathrm{Vol_d}(K_k)\mathbb{E}\mathrm{Vol_d}(K_j))e^{-2n}\frac{n^{k+j}}{k!j!} + O\left(\left(\frac{e}{2}\right)^{-n/2}\right)\\
&= \sum_{j=\frac{n}{2}}^{\infty}\sum_{k=j}^{\infty}(\mathbb{E}\mathrm{Vol_d}(K_k) - \mathbb{E}\mathrm{Vol_d}(K_j))^2e^{-2n}\frac{n^{k+j}}{k!j!} + O\left(\left(\frac{e}{2}\right)^{-n/2}\right),
\end{aligned}$$

where the third equality is due to (15). By Lemma 6.9, $\mathbb{E}\mathrm{Vol_d}(K_{j+1}) - \mathbb{E}\mathrm{Vol_d}(K_j) = c(K,d)j^{-\frac{d+1}{d-1}}$ when $j \to \infty$, hence

$$\mathbb{E}\mathrm{Vol_d}(K_k) - \mathbb{E}\mathrm{Vol_d}(K_j) = \sum_{i=j}^{k-1}\mathbb{E}\mathrm{Vol_d}(K_{i+1}) - \mathbb{E}\mathrm{Vol_d}(K_i) \le c(K,d)(k-j)j^{-\frac{d+1}{d-1}},$$

and

$$\operatorname{Var}\mathbb{E}(\operatorname{Vol_d}(\Pi_n)\mid N)\le c(K,d)\sum_{j=\frac{n}{2}}^{\infty}\sum_{k=j}^{\infty}(k-j)^2 j^{-\frac{2d+2}{d-1}}e^{-2n}\frac{n^{k+j}}{k!j!}+O\left(\left(\frac{e}{2}\right)^{-n/2}\right)$$

$$\le cn^{-\frac{2d+2}{d-1}}\operatorname{Var}N+O\left(\left(\frac{e}{2}\right)^{-n/2}\right)=O(n^{-\frac{d+3}{d-1}}).$$

Now, $\operatorname{Var}\operatorname{Vol_d}(K_n)=\Theta(n^{-\frac{d+3}{d-1}})$, so by (15) and (16), we have

$$\mathbb{E}\operatorname{Var}\operatorname{Vol_d}(\Pi_n|N)=\mathbb{E}(\Theta(N^{-\frac{d+3}{d-1}}))$$

$$=O\left(\mathbb{P}\left(N\le\frac{n}{2}\right)\right)+\mathbb{E}(N^{-\frac{d+3}{d-1}}\chi_{\{\frac{n}{2}<N\le 3n\}})+O(\mathbb{P}(3n<N))$$

$$=\Theta(n^{-\frac{d+3}{d-1}}).\qquad\square$$

Acknowledgement The authors would like to thank Imre Bárány for many enlightening conversations and discussions on this subject.

Appendix 1 Geometric Toolkit

6.1 Boundary Approximation

We begin with some basic notions and notation. For $K\in\mathcal{K}^2_+$,at each point $x\in\partial K$ there is a unique paraboloid Q_x, given by a quadratic form b_x, osculating ∂K at x. We may describe Q_x and b_x by identifying the tangent hyperplane of ∂K at x with $\mathbb{R}^{d-1}$ and x with the origin. This is a well known fact, see e.g. [10]. In a neighborhood of x, we can represent ∂K as the graph of a $\mathcal{C}^2$, convex function $f:\mathbb{R}^{d-1}\to\mathbb{R}$, i.e. each point in ∂K near x can be written in the form $(y, f_x(y))$, where $y\in R^{d-1}$ the form $(y^1,\dots,y^{d-1})$. Thus, we may write

$$b_x(y)=\frac{1}{2}\sum_{1\le i,j\le d-1}\frac{\partial f_x}{\partial y^i\partial y^j}(0)y^i y^j\quad\text{and}$$

$$Q_x=\{(y,z)\mid z\ge b_x(y),\ y\in\mathbb{R}^{d-1},z\in\mathbb{R}\},$$

here $\frac{\partial f_x}{\partial y^i\partial y^j}(0)$ denote the second partial derivative of f_x at the origin with respect to y^i and y^j. The main thrust of the above is that these paraboloids approximate the boundary structure. The formulation given here is due to Reitzner, who provides a proof in [12].

Lemma 6.1 *Let $K\in\mathcal{K}^2_+$ and choose $\delta>0$ sufficiently small. Then there exists a $\lambda>0$, depending only on δ and K, such that for each point $x\in\partial K$ the following holds*: *If we identify the tangent hyperplane to ∂K at x with $\mathbb{R}^{d-1}$ and x with the origin, then we may define the $\lambda-$neighborhood U^λ of $x\in\partial K$ by* $\operatorname{proj}U^\lambda=B^{d-1}(0,\lambda)$.

U^λ can be represented by a convex function $f_x(y) \in \mathcal{C}^2$, for $y \in B^{d-1}(0,\lambda)$. Furthermore,

$$(1+\delta)^{-1} b_x(y) \le f_x(y) \le (1+\delta) b_x(y) \quad \textit{and} \tag{25}$$

$$\sqrt{1+|\nabla f_x(y)|^2} \le (1+\delta), \tag{26}$$

for $y \in B^{d-1}(0,\lambda)$, where b_x is defined as above and $\nabla f_x(y)$ stands for the gradient of $f_x(y)$.

This lemma proves that at each point $x \in \partial K$, the deviation of the boundary of the approximating paraboloid ∂Q_x from ∂K is uniformly bounded in a small neighborhood of x.

We use this lemma to show how one can relate ϵ-caps to ϵ-boundary caps. This relationship is used repeatedly throughout the paper as it allows us to work with volumes of different dimensions.

Lemma 6.2 *For a given $K \in \mathcal{K}^2_+$, there exists constants $\epsilon_0, c, c' > 0$ such that for all $0 < \epsilon < \epsilon_0$ we have that for any ϵ-cap C of K,*

$$c^{-1}\epsilon^{(d-1)/(d+1)} \le \mu(C \cap \partial K) \le c\epsilon^{(d-1)/(d+1)}$$

and for any ϵ-boundary cap C' of K,

$$c'^{-1}\epsilon^{(d+1)/(d-1)} \le \mathrm{Vol_d}(C') \le c'\epsilon^{(d+1)/(d-1)}.$$

Proof We shall prove the first statement. Fix some $\delta > 0$ for Lemma 6.1.

Consider in $\mathbb{R}^d$ the paraboloid given by the equation

$$z^d \ge (z^1)^2 + (z^2)^2 + \cdots + (z^{d-1})^2.$$

Intersecting this paraboloid with the halfspace defined by the equation $z^d \le 1$ gives an object which we shall call the standard cap, E. We form $(1+\delta)^{-1}E$ and $(1+\delta)E$ similarly by the equations $z^d \ge (1+\delta)^{-1}((z^1)^2 + (z^2)^2 + \cdots, (z^{d-1})^2)$ and $z^d \ge (1+\delta)((z^1)^2 + (z^2)^2 + \cdots, (z^{d-1})^2)$, using the same halfspace as before. We note the inclusions

$$(1+\delta)^{-1}E \supset E \supset (1+\delta)E.$$

Let $c_1 = \mathrm{Vol_d}((1+\delta)^{-1}E)$ and $c_2 = \mathrm{Vol_d}((1+\delta)E)$, and further set $c_3 = \mu(\mathrm{proj}((1+\delta)^{-1}E))$ and $c_4 = \mu(\mathrm{proj}((1+\delta)E))$ where here proj is the orthogonal projection onto the hyperplane spanned by the first $(d-1)$ coordinates.

Now, let C be our ϵ-cap. Let x be the unique point in ∂K whose tangent hyperplane is parallel to the hyperplane defining C. Assuming that Lemma 6.1 applies, we may equate the tangent hyperplane of ∂K at x with $\mathbb{R}^{d-1}$, and view $C \cap \partial K$ as being given by some convex function $f : \mathbb{R}^{d-1} \to \mathbb{R}$. Further, let Q_x be the unique paraboloid osculating ∂K at x. Let A be a linear transform that takes E to Q_x. We observe that Q_x is the paraboloid defined by the set $z^d \ge b_x(z^1, \ldots, z^{d-1})$ intersected with the halfspace $z^d \le h$, for some $h > 0$. We can define $(1+\delta)^{-1}Q_x$ (resp.

$(1+\delta)Q_x$) to be the set defined by the intersection of this same half space and the points given by $z^d \geq (1+\delta)^{-1} b_x(z^1, \ldots, z^{d-1})$ (resp. $z^d \geq (1+\delta) b_x(z^1, \ldots, z^{d-1})$). Observe that $A((1+\delta)^{-1}E) = (1+\delta)^{-1}Q_x$ and $A((1+\delta)E) = (1+\delta)Q_x$.

Appealing to Lemma 6.1, we see that

$$(1+\delta)^{-1}Q_x \supset C \supset (1+\delta)Q_x.$$

This gives

$$c_1 |\det A| \geq \epsilon \geq c_2 |\det A|. \tag{27}$$

Let $\tilde{f} : \mathbb{R}^{d-1} \to \partial K$ be the function induced by f, i.e. $\tilde{f}(y) = (y, f_x(y))$. Using the inclusion

$$\tilde{f}(\text{proj}((1+\delta)^{-1}Q_x)) \supset C \cap \partial K \supset \tilde{f}(\text{proj}((1+\delta)Q_x))$$

and the bound

$$(1+\delta) \geq \sqrt{1 + |\nabla f|^2} \geq 1$$

furnished by Lemma 6.1, if A' represents the restriction of A to the first $(d-1)$ coordinates, we obtain

$$c_3 |\det A'|(1+\delta) \geq \mu(C \cap \partial K) \geq c_4 |\det A'|. \tag{28}$$

A simple computation shows $|\det A| = 2^{(d-1)/2} \kappa^{-1/2} h^{(d+1)/2}$ and $|\det A'| = 2^{(d-1)/2} \kappa^{-1/2} h^{(d-1)/2}$, where κ is the Gauß–Kronecker curvature of ∂K at x. Using this and (27) gives upper and lower bounds on h, and this bound with (28) gives

$$c_5 \epsilon^{(d-1)/(d+1)} \geq \mu(C \cap \partial K) \geq c_6 \epsilon^{(d-1)/(d+1)},$$

where here c_5, c_6 are constants depending only on κ. As K is compact and κ is always positive we can assume we can change c_5 and c_6 to be independent of κ, and hence x.

Finally, we return to the issue of values of ϵ (hence h) for which Lemma 6.1 applies. We note that in general every quadratic form b_x can be given by

$$b_x(y) = \frac{1}{2} \sum_i k_i (y^i)^2,$$

where k_i are the principal curvatures. We observe that as the Gauß–Kronecker curvature is positive then there are positive constants k' and k'' depending only on K such that $0 < k' < k_i < k''$. This bounds the possible geometry of Q_x, and implies the existence of an ϵ_0 such that for $0 < \epsilon < \epsilon_0$, such that $\text{proj}((1+\delta)^{-1}Q_x) \subset B(0, \lambda)$ (λ as given in Lemma 6.1), allowing us to apply Lemma 6.1. This completes the proof of the first statement. The second statement is similar. Relaxing constants allows the statement as given. □

Remark 6.3 It is important to note that the above is *not* true for general convex bodies. In particular, any polytope P provides an example of a convex body with caps C such that the quantities $\text{Vol}_d(C)$ and $\mu(C \cap \partial P)$ are unrelated.

6.2 Caps and Cap Covers

Lemma 6.4 through 6.8 and their proofs below can be found in [10].

Lemma 6.4 *Given $K \in \mathcal{K}^2_+$, there exist constants d_1, d_2 such that for each cap $C(x,h)$ with $h \le h_0$, we have*

$$\partial K \cap B(x, d_1 h^{\frac{1}{2}}) \subset C(x,h) \subset B(x, d_2 h^{\frac{1}{2}}).$$

Lemma 6.5 *Given $K \in \mathcal{K}^2_+$, there exists a constant d_3 such that for each cap $C(x,h)$ with $h \le h_0$, we have*

$$\mathrm{Vol_d}(C(x,h)) \le d_3 h^{\frac{d+1}{2}}.$$

Lemma 6.6 (Cap covering) *Given $m \ge m_0$ and $K \in \mathcal{K}^2_+$, there are points $y_1, \ldots, y_m \in \partial K$, and caps $C_i = C(y_i, h_m)$ and $\overline{C}_i = C(y_i, (2d_2/d_1)^2 h_m)$ with*

$$\begin{aligned} &C_i \subset B(y_i, d_2 h_m^{1/2}) \subset \mathrm{Vor}(y_i), \\ &\mathrm{Vor}(y_i) \cap \partial K \subset B(y_i, 2d_2 h_m^{1/2}) \cap \partial K \subset \overline{C}_i \quad \textit{and} \\ &h_m = \Theta(m^{-\frac{2}{d-1}}). \end{aligned}$$

Here $\mathrm{Vor}(y_i)$ *is the Voronoi cell of y_i in K defined by*:

$$\mathrm{Vor}(y_i) = \{x \in K : \| x - y_i \| \le \| x - y_k \| \textit{ for all } k \ne i\},$$

and we have

$$\mathrm{Vol_d}(C_i) = \Theta(m^{-\frac{d+1}{d-1}}),$$

for all $i = 1, \ldots, m$.

Proof The proof follows from the fact that given m, for a suitable r_m, we can find balls $B(y_i, r_m)$, $i = 1, \ldots, m$ such that they form a maximal packing of ∂K, hence $B(y_i, 2r_m)$ form a covering of ∂K. Use Lemma 6.4, one can convert between the height of cap h_m and radius of the ball r_m. □

Lemma 6.7 *Let K, m be given, and $y_i, i = 1, \ldots, m$ be chosen as in Lemma* 6.6. *The number of Voronoi cells* $\mathrm{Vor}(y_j)$ *intersecting a cap $C(y_i, h)$ is $O((h^{\frac{1}{2}} m^{\frac{1}{d-1}} + 1)^{d+1})$, $i = 1, \ldots, m$.*

Lemma 6.8 *Let m, K and $y_i, C_i, i = 1, \ldots, m$ be chosen as in Lemma* 6.6. *Choose on the boundary within each cap C_i an arbitrary point x_i (i.e. $x_i \in C_i \cap \partial K$), then*

$$\delta^H(K, [x_1, \ldots, x_m]) = O(m^{-\frac{2}{d-1}}),$$

and there is a constant c such that for any $y \in \partial K$ with $y \notin C(y_i, cm^{-\frac{2}{d-1}})$, the line segment $[y, x_i]$ intersects the interior of the convex hull $[x_1, \ldots, x_m]$.

Lemma 6.9 *For large n,*

$$\mathbb{E}\mathrm{Vol_d}(K_{n+1}) - \mathbb{E}\mathrm{Vol_d}(K_n) = O(n^{-(d+1)/(d-1)}).$$

This lemma can be proved using techniques from integral geometry similar to that found in [11]. Alternatively, one case use the notion of ϵ-floating bodies to give an appropriate bound. We give a proof sketch below of a slightly weaker version below, and note that through techniques similar to that used to prove Theorem 1.2 and in [18], we can remove the logarithmic factor.

Sketch of the Proof Following the notation found in the concentration proof, let $\Omega' = \{t = (t_1, \ldots, t_n) \mid t_i \in \partial K\}$, and put $L = \{t_1, \ldots, t_n\}$.

Observe that we can write

$$\begin{aligned}
&\mathbb{E}\mathrm{Vol_d}(K_{n+1}) - \mathbb{E}\mathrm{Vol_d}(K_n) \\
&= \int_{\Omega'} \int_{\partial K} \mathrm{Vol_d}([t_1, \ldots, t_n, t_{n+1}]) - \mathrm{Vol_d}([t_1, \ldots, t_n]) dt_{n+1} dt \\
&= \int_{\Omega'} \int_{\partial K} \Delta_{t_{n+1}, L} dt_{n+1} dt.
\end{aligned}$$

Let A be the event that $F_\epsilon \subseteq [L]$. The integrand can be estimated by

$$\Delta_{t_{n+1}, L} \leq g(\epsilon)\chi_A + \chi_{\bar{A}}.$$

Here, we use $g(\epsilon)$ as an upperbound for $\Delta_{t_{n+1}, L}$ when $F_\epsilon \subseteq [L]$ and 1 otherwise. This bound is independent of t_{n+1}, so our integral is upper bounded by

$$\int_{\Omega'} g(\epsilon)\chi_A + \chi_{\bar{A}} dt \leq g(\epsilon) + P(F_\epsilon \nsubseteq [L]).$$

Setting $\epsilon = c \ln n / n$ so that it satisfies Lemma 4.2 we find that $g(\epsilon) = \Theta(\epsilon^{(d+1)/(d-1)}) = \Theta(n^{-(d+1)/(d-1)} poly(\ln n))$ and $P(F_\epsilon \nsubseteq [L]) = \exp(-c'\epsilon n) = n^{-cc'}$. Choosing c to be sufficiently large we find that

$$\mathbb{E}\mathrm{Vol_d}(K_{n+1}) - \mathbb{E}\mathrm{Vol_d}(K_n) = O(n^{-(d+1)/(d-1)} poly(\ln n)). \qquad \square$$

Appendix 2 Proof of Corollaries 1.3 and 1.4

Proof of Corollary 1.3 Let $\lambda_0 = \frac{\alpha}{4} n^{\frac{d-1}{3d+1} + \frac{2(d+1)\eta}{d-1}}$ be the upper bound for λ given in Theorem 1.2. So for $\lambda > \lambda_0$, by (1.2)

$$\begin{aligned}
\mathbb{P}(|Z - \mathbb{E}Z| \geq \sqrt{\lambda V_0}) &\leq \mathbb{P}(|Z - \mathbb{E}Z| \geq \sqrt{\lambda_0 V_0}) \\
&\leq 2\exp(-\lambda_0/4) + \exp(-cn^{\frac{d-1}{3d+1} - \eta}).
\end{aligned}$$

Combining (1.2) and the above, we get for any $\lambda > 0$,

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq \sqrt{\lambda V_0}) \leq 2\exp(-\lambda/4) + 2\exp(-\lambda_0/4) + \exp(-cn^{\frac{d-1}{3d+1} - \eta}). \quad (29)$$

We then compute the kth moment M_k of Z, beginning with the definition:

$$M_k = \int_0^\infty t^k d\mathbb{P}(|Z - \mathbb{E}Z| < t).$$

If we set $\gamma(t) = \mathbb{P}(|Z - \mathbb{E}Z| \geq t)$ then we can write

$$\begin{aligned} M_k &= \int_0^\infty t^k d\mathbb{P}(|Z - \mathbb{E}Z| < t) = -\int_0^\infty t^k d\gamma(t) \\ &= (-t^k\gamma(t))|_0^\infty + \int_0^\infty kt^{k-1}\gamma(t)dt = \int_0^1 kt^{k-1}\gamma(t)dt. \end{aligned}$$

Note that the limits of integration can be limited to $[0, 1]$ because we've assumed the volume of K is normalized to 1.

Setting $t = \sqrt{\lambda V_0}$ we get

$$\begin{aligned} &\int_0^1 kt^{k-1}\gamma(t)dt \\ &= \int_0^{1/V_0} k(\sqrt{\lambda V_0})^{k-1}\mathbb{P}(|Z - \mathbb{E}Z| \geq \sqrt{\lambda V_0})\frac{\sqrt{V_0}}{2\sqrt{\lambda}}d\lambda \\ &\leq \frac{k}{2}V_0^{k/2}\int_0^{1/V_0} \lambda^{\frac{k}{2}-1}2\exp(-\lambda/4) + 2\exp(-\lambda_0/4) + \exp(-cn^{\frac{d-1}{3d+1}-\eta})d\lambda \\ &\qquad\qquad \text{by (29).} \end{aligned}$$

We may now evaluate each term separately.

For the first term we observe that

$$\int_0^{1/V_0} 2\lambda^{\frac{k}{2}-1}\exp(-\lambda/4)d\lambda \leq \int_0^\infty 2\lambda^{\frac{k}{2}-1}\exp(-\lambda/4)d\lambda = c_k,$$

where c_k is a constant depending only on k.

Since

$$V_0 = \alpha n^{-\frac{d+3}{d-1}} \gg n^{-5},$$

we can compute the second term:

$$\begin{aligned} \int_0^{1/V_0} \lambda^{\frac{k}{2}-1}2\exp(-\lambda_0/4)d\lambda &\leq \frac{2}{k}2\exp(-\lambda_0/4)V_0^{-\frac{k}{2}} \\ &\leq \frac{2}{k}2\exp\left(-\frac{\alpha}{16}n^{\frac{d-1}{3d+1}+\frac{2(d+1)\eta}{d-1}}\right)n^{\frac{5k}{2}} = o(1). \end{aligned}$$

The last term can be computed similarly and gives $o(1)$ again. Hence,

$$M_k \leq (c_k + o(1))kV_0^{k/2} = O(V_0^{k/2}). \qquad \square$$

Proof of Corollary 1.4

$$\mathbb{P}\left(\left|\frac{Z_n}{\mathbb{E}Z_n} - 1\right| f(n) \geq \delta(n)\right) \leq \mathbb{P}\left(|Z_n - \mathbb{E}Z_n| \geq \mathbb{E}Z_n\sqrt{32n^{-\frac{d+3}{d-1}}\ln n}\right)$$

$$\leq \mathbb{P}(|Z_n - \mathbb{E}Z_n| \geq \sqrt{8\ln n V_0})$$
$$\leq 2\exp(-8\ln n/4) + \exp(-cn^{\frac{d-1}{3d+1}-\eta})$$
$$\leq 3\exp(-2\ln n) \leq 3n^{-2},$$

by Theorem 1.2. The second inequality above is due to the fact that $\mathbb{E}Z_n = 1 - c_K n^{-\frac{2}{d-1}} > 1/2$ when n is large. Since $\sum n^{-2}$ is convergent, by Borel–Cantelli, $|\frac{Z_n}{\mathbb{E}Z_n} - 1| f(n)$ converges to 0 almost surely, hence the corollary. □

Appendix 3 Proof of Lemma 3.1

We first prove the following claim. The notation follows that found in Section 3.1

Claim 8.1 *Let $x \in \partial K$. There is some $h(K) > 0$ such that for $h(K) > h > 0$ there exists a constant $c(r) > 0$ depending only on r and K such that*

$$\frac{1}{2}|\det A|^2 c(r) \leq \mathrm{Var}_Y(\mathrm{Vol_d}([Y, Av_1, \ldots, Av_d])) \leq 2|\det A|^2 c(r),$$

and Y is a random point chosen in $D_0'(x)$ according to the distribution on ∂K.

Proof To prove this claim, we compute. Recall that A is the linear map which takes E to the paraboloid Q_x. We shall denote by A' the map A restricted to $\mathbb{R}^{d-1}$. We shall denote by $f : T_x(\partial K) \approx \mathbb{R}^{d-1} \to \mathbb{R}$ the function whose graph defines ∂K locally, and $\tilde{f} : \mathbb{R}^{d-1} \to \partial K$ the function induced by f. Thus, we have:

$$\begin{aligned}
&\mathbb{E}_Y(\mathrm{Vol_d}([Y, Av_1, \ldots, Av_d])) \\
&= \frac{\int_{D_0} \mathrm{Vol_d}([\tilde{f}(Y), Av_1, \ldots, Av_d])\rho(\tilde{f}(Y))\sqrt{1 + f_{Y^1}^2 + \cdots + f_{Y^{d-1}}^2}\,dY}{\int_{A'(C_0)} \rho(f'(Y))\sqrt{1 + f_{Y^1}^2 + \cdots + f_{Y^{d-1}}^2}\,dY} \\
&= \left(|\det A'| \int_{C_0} \mathrm{Vol_d}([\tilde{f}(AX), Av_1, \ldots, Av_d])\rho(\tilde{f}(AX))\right. \\
&\quad \left. \times \sqrt{1 + f_{Y^1}^2 + \cdots + f_{Y^{d-1}}^2}(AX)dX\right) \Big/ \left(|\det A'| \int_{C_0} \rho(f'(AX)) \right. \\
&\quad \left. \times \sqrt{1 + f_{Y^1}^2 + \cdots + f_{Y^{d-1}}^2}(AX)dY\right). \qquad (30)
\end{aligned}$$

Observe that if we set $A^{-1} \circ \tilde{f}(AX) = f^*$ to be the pullback of $\tilde{f}$ under A then $\mathrm{Vol_d}([\tilde{f}(AX), Av_1, \ldots, Av_d]) = |\det A| \cdot \mathrm{Vol_d}([f^*(X), v_1, \ldots, v_d])$. Letting $b : \mathbb{R}^{d-1} \to \mathbb{R}$ denote the quadratic form defining E, $\tilde{b} : \mathbb{R}^{d-1} \to \partial E$ the induced function, we then use Lemma 6.1 to get the bound

$$2^{-1}b \leq f \circ A' \leq 2b, \qquad (31)$$

when h is sufficiently small. Thus, we get the bound

$$\begin{aligned}\mathrm{Vol_d}([2^{-1}b(X), v_1, \ldots, v_d]) &\geq \mathrm{Vol_d}([f^*(X), v_1, \ldots, v_d]) \\ &\geq \mathrm{Vol_d}([2b(X), v_1, \ldots, v_d]),\end{aligned}$$

which follows from the geometry. Now, since $v_1, \ldots, v_d$ form a $(d-1)$ simplex parallel to the plane $\mathbb{R}^{d-1}$ we can write $\mathrm{Vol_d}([b(X), v_1, \ldots, v_d]) = c_d(1 - b(X))$, where c_d is some positive constant depending only on dimension. We may write $b(X) = |X|^2$, and this allows us to see that

$$\begin{aligned}&\mathrm{Vol_d}([2^{-1}\tilde{b}(X), v_1, \ldots, v_d]) \\ &\quad = \mathrm{Vol_d}([\tilde{b}(X), v_1, \ldots, v_d])(1 - 2^{-1}|X|^2)/(1 - |X|^2) \\ &\quad = \mathrm{Vol_d}([\tilde{b}(X), v_1, \ldots, v_d])(1 - 2^{-1}|X|^2)(1 + |X|^2 + |X|^4 + \cdots) \\ &\quad = \mathrm{Vol_d}([\tilde{b}(X), v_1, \ldots, v_d])(1 + o_r(1)).\end{aligned}$$

Here, $o_r(1)$ indicates a function which goes to 0 as r goes to 0. Similarly, we have

$$\mathrm{Vol_d}([2\tilde{b}(X), v_1, \ldots, v_d]) = \mathrm{Vol_d}([\tilde{b}(X), v_1, \ldots, v_d])(1 + o_r(1)).$$

Thus, we may write

$$\begin{aligned}&\frac{\int_{C_0} \mathrm{Vol_d}([f^*(X), v_1, \ldots, v_d])\rho(\tilde{f}(AX))\sqrt{1 + f_{Y^1}^2 + \cdots + f_{Y^{d-1}}^2}(AX)dX}{\int_{C_0} \rho(\tilde{f}(AX))\sqrt{1 + f_{Y^1}^2 + \cdots + f_{Y^{d-1}}^2}(AX)dY} \\ &\quad \geq (1 + o_r(1)) \\ &\qquad \times \frac{\int_{C_0} \mathrm{Vol_d}([\tilde{b}(X), v_1, \ldots, v_d])\rho(\tilde{f}(AX))\sqrt{1 + f_{Y^1}^2 + \cdots + f_{Y^{d-1}}^2}(AX)dX}{\int_{C_0} \rho(\tilde{f}(AX))\sqrt{1 + f_{Y^1}^2 + \cdots + f_{Y^{d-1}}^2}(AX)dY}.\end{aligned} \tag{32}$$

Setting $F(X) = \rho(\tilde{f}(AX))\sqrt{1 + f_{Y^1}^2 + \cdots + f_{Y^{d-1}}^2}(AX)$ the above is thus

$$\geq (1 + o_r(1)) \cdot \frac{\min_{C_0} F(X)}{\max_{C_0} F(X)} \cdot \frac{\int_{C_0} \mathrm{Vol_d}([b(X), v_1, \ldots, v_d])dX}{\int_{C_0} dX}.$$

Now, if we can show that the term $\frac{\min_{C_0} F(X)}{\max_{C_0} F(X)} \geq (1 + o_{r,h}(1))$, only depending on r and h, then from our earlier observation we can conclude that (30) is bounded below by

$$|\det A| \cdot (1 + o_{r,h}(1)) \cdot \frac{\int_{C_0} \mathrm{Vol_d}([b(X), v_1, \ldots, v_d])dX}{\int_{C_0} dX}.$$

Note $o_{r,h}(1)$ denotes a function which goes to 0 as both r and h go to 0.

Invoking Lemma 6.1, we observe that we may make the term

$$\sqrt{1 + f_{Y^1}^2 + \cdots + f_{Y^{d-1}}^2}(AX)$$

sufficiently less than $(1+\delta)$, for any $\delta > 0$, by choosing r, h both sufficiently small (independent of f). Thus, we may write $\sqrt{1 + f_{Y^1}^2 + \cdots + f_{Y^{d-1}}^2}(AX) = (1 + o_{r,h}(1))$.

Next, we note that ρ is a uniformly continuous function on K. It is not too hard to see that the function $\min_{C_0} \rho(f'(AX))/\max_{C_0} \rho(f'(AX)) = (1 + o_{r,h}(1))$, where again the $o(1)$ function is independent of the basepoint. Using the fact that

$$\begin{aligned}&\min \rho(f'(AX))\sqrt{1 + f_{Y^1}^2 + \cdots + f_{Y^{d-1}}^2}(AX)\\ &\quad \geq (\min \rho(f'(AX)))(\min \sqrt{1 + f_{Y^1}^2 + \cdots + f_{Y^{d-1}}^2}(AX))\end{aligned}$$

(similarly for max) we thus find that

$$(1 + o_{r,h}(1)) \geq \frac{\min_{C_0} F(X)}{\max_{C_0} F(X)} \geq (1 + o_{r,h}(1)),$$

where the functions in question are independent of basepoint.

If we let

$$\phi_1(r) = \frac{\int_{C_0} \mathrm{Vol_d}([\tilde{b}(X), v_1, \ldots, v_d])dX}{\int_{C_0} dX}$$

then we can summarize our findings as, independent of basepoint,

$$\lim_{h\to 0} \frac{\mathbb{E}_Y(\mathrm{Vol_d}([Y, Av_1, \ldots, Av_d]))}{|\det A|\phi_1(r)} = (1 + o_r(1)). \tag{33}$$

By an identical argument, if we set $\phi_2(r) = \frac{\int_{C_0} \mathrm{Vol_d^2}([\tilde{b}(X), v_1, \ldots, v_d])dX}{\int_{C_0} dX}$ then we have

$$\lim_{h\to 0} \frac{\mathbb{E}_Y(\mathrm{Vol_d^2}([Y, Av_1, \ldots, Av_d]))}{|\det A|^2\phi_2(r)} = (1 + o_r(1)). \tag{34}$$

Using (33) and (34) we can compute:

$$\begin{aligned}&\lim_{h\to 0} \mathrm{Var}_Y([Y, Av_1, \ldots, Av_d])/|\det A|^2\\ &\quad = \lim_{h\to 0} \mathbb{E}_Y(\mathrm{Vol_d^2}([Y, Av_1, \ldots, Av_d]))/|\det A|^2\\ &\qquad - \lim_{h\to 0} \mathbb{E}_Y^2(\mathrm{Vol_d}([Y, Av_1, \ldots, Av_d]))/|\det A|^2\\ &\quad = \phi_2(r)(1 + o_r(1)) - \phi_1^2(r)(1 + o_r(1))^2 = (\phi_2(r) - \phi_1^2(r))(1 + o_r(1)).\end{aligned} \tag{35}$$

Thus, by letting r become sufficiently small so that the final $(1 + o_r(1)) > 0$ we note that (35) is positive, since this quantity $\phi_2(r) - \phi_1^2(r)$ is just the variance of

$\mathrm{Vol_d}([b(X), v_1, \ldots, v_d])$ where X is taken over C_0, thus always positive. This proves there exists $c_1 > 0$ such that for h sufficiently small,

$$\mathrm{Var}_Y([Y, Av_1, \ldots, Av_d]) \geq c_1 |\det A|^2.$$

By the same arguments we also get

$$\mathrm{Var}_Y([Y, Av_1, \ldots, Av_d]) \leq c_2 |\det A|^2.$$

So the claim is proved. □

With the preceding claim, we now prove Lemma 3.1. Instead of the convex hull of $[Y, Av_1, \ldots, Av_d]$ we shall study the convex hull $[Y, x_1, \ldots, x_d]$, where $x_i \in D_i'$, using the fact that the x_i are close to the Av_i when h is small. To do this, we'll need a second claim.

Claim 8.2 *There exists a $\delta > 0$ such that If for each i, $x_i \in B(v_i, d)$, then*

$$\mathrm{Vol_d}([2^{-1}b(X), x_1, \ldots, x_d]) = \mathrm{Vol_d}([b(X), x_1, \ldots, x_d])(1 + o_r(1))$$

and

$$\mathrm{Vol_d}([2b(X), x_1, \ldots, x_d]) = \mathrm{Vol_d}([b(X), x_1, \ldots, x_d])(1 + o_r(1)),$$

where the hidden functions depend only on r (i.e. they are not functions of the x_i).

Proof We simply note that there exists a $\delta > 0$ such that for any fixed choice of x_i,

$$\frac{\mathrm{Vol_d}([2^{-1}b(X), x_1, \ldots, x_d])}{\mathrm{Vol_d}([b(X), x_1, \ldots, x_d])} \to 1 \quad \text{as } X \to 0.$$

We also note that $X, x_1, \ldots, x_d$ lie in $C_0 \times B(v_1, \delta) \times \cdots \times B(v_d, \delta)$, a compact set. These two conditions guarantee that the maximum of the ratio, taken over all $x_1, \ldots, x_d$, converges to 1 as $X \to 0$. Thus, the ratio converges to 1 independently of the choice of $x_1, \ldots, x_d$, and hence the claimed result.

The statement for $\mathrm{Vol_d}([2b(X), x_1, \ldots, x_d])$ is analogous. □

With this claim, we can adapt Claim 8.1 to work for any $x_i \in B(v_i, \delta)$, by using the above claim in place of (32). With this we can show that for h sufficiently small we can choose r sufficiently small such that

$$\begin{aligned}
&\frac{1}{2}|\det A|^2 \,\mathrm{Var}_X(\mathrm{Vol_d}([b(X), x_1, \ldots, x_d])) \\
&\quad \leq \mathrm{Var}_Y(\mathrm{Vol_d}([Y, Ax_1, \ldots, Ax_d])) \\
&\quad \leq 2|\det A|^2 \,\mathrm{Var}_X(\mathrm{Vol_d}([b(X), x_1, \ldots, x_d])),
\end{aligned} \tag{36}$$

where here the quantity $\mathrm{Var}_X(\mathrm{Vol_d}([b(X), x_1, \ldots, x_d]))$ is the variance taken over C_0. But as $\mathrm{Var}_X(\mathrm{Vol_d}([b(X), v_1, \ldots, v_d]))$ is positive, continuity guarantees that

$$c' > \mathrm{Var}_X(\mathrm{Vol_d}([b(X), x_1, \ldots, x_d])) > c > 0$$

if the x_i are sufficiently close to the v_i, say $x_i \in B(v_i, \eta)$ for all i, for some $\eta > 0$. Then,

$$\frac{1}{2}|\det A|^2 c' \leq \mathrm{Var}_Y(\mathrm{Vol_d}([Y, Ax_1, \ldots, Ax_d])) \leq 2|\det A|^2 c, \tag{37}$$

if $x_i \in B(v_i, \eta)$ for all i.

Now, we need to verify that we can choose C_i sufficiently small such that points in D_i' always map into $B(v_i, \eta)$, which will complete the lemma. To do this, note that if we set $r' < \eta/2$, then we can choose $\epsilon > 0$ such that

$$\begin{aligned} U_i = \{(x, y) \in \mathbb{R}^d \mid x \in B(\mathrm{proj}v_i, \eta/2) \subset \mathbb{R}^{d-1} \text{ and} \\ (1+\epsilon)^{-1} b(x) \leq y \leq (1+\epsilon) b(x)\} \subset B(v_i, \eta) \end{aligned} \tag{38}$$

for each i. By Lemma 6.1 we can take h to be sufficiently small such that for all $x \in \partial K$

$$(1+\epsilon)^{-1} b_x(y) \leq f_x(y) \leq (1+\epsilon) b_x(y)$$

in all caps of height h. So if we thus choose C_i to be the $\eta/2$ ball about $\mathrm{proj}v_i$, then we note that $D_i' \subset A(U_i)$. Thus, any $y_i \in D_i'$ can be written as Ax_i for some $x_i \in U_i \subset B(v_i, \eta)$, and thus (37) holds. Hence, the lemma.

References

1. Alon, N., Spencer, J.: The Probabilistic Method. Wiley, New York (2000)
2. Azuma, K.: Weighted sums of certain dependent random variables. Tohoku Math. J. **19**, 357–367 (1967)
3. Baldi, P., Rinott, Y.: On normal approximations of distributions in terms of dependency graphs. Ann. Probab. **17**(4), 1646–1650 (1989)
4. Bárány, I.: Personal conversations, UCSD (2005)
5. Bárány, I.: Convex bodies, random polytopes, and approximation. In: Weil, W. (ed.) Stochastic Geometry. Springer (2005)
6. Bárány, I., Larman, D.: Convex bodies, economic cap coverings, random polytopes. Mathematika **35**(2), 274–291 (1988)
7. Bárány, I., Reitzner, M.: Central limit theorem for random polytopes in convex polytopes. Manuscript (2005)
8. Efron, B.: The convex hull of a random set of points. Biometrika **52**, 331–343 (1965)
9. Kim, J.H., Vu, V.H.: Concentration of multi-variate polynomials and its applications. Combinatorica **20**(3), 417–434 (2000)
10. Reitzner, M.: Central limit theorems for random polytopes. Probab. Theory Relat. Fields **133**, 483–507 (2005)
11. Reitzner, M.: Random polytopes and the Efron–Stein jacknife inequality. Ann. Probab. **31**, 2136–2166 (2003)
12. Reitzner, M.: Random points on the boundary of smooth convex bodies. Trans. Am. Math. Soc. **354**(6), 2243–2278 (2002)
13. Rényi, A., Sulanke, R.: Über die konvexe Hülle von n zufällig gewählten Punkten. Z. Wahrsch. Verw. Geb. **2**, 75–84 (1963)
14. Rinott, Y.: On normal approximation rates for certain sums of random variables. J. Comput. Appl. Math. **55**, 135–143 (1994)
15. Schneider, R.: Discrete aspects of stochastic geometry. In: Goodman, J., O'Rourke, J. (eds.) Handbook of Discrete and Computational Geometry, pp. 255–278. CRC Press, Boca Raton (2004)
16. Schütt, C., Werner, E.: Polytopes with vertices chosen randomly from the boundary of a convex body. In: Geometric Aspects of Functional Analysis 2001–2002. Lecture Notes in Mathematics, vol. 1807, pp. 241–422. Springer, New York (2003)

17. Sylvester, J.J.: Question 1491. Educational Times. London (April, 1864)
18. Vu, V.H.: Sharp concentration of random polytopes. Geom. Funct. Anal. **15**, 1284–1318 (2005)
19. Vu, V.H.: Central limit theorems for random polytopes in a smooth convex set. Adv. Math. **207**, 221–243 (2005)
20. Weil, W., Wieacker, J.: Stochastic geometry. In: Gruber, P., Wills, J. (eds.) Handbook of Convex Ceometry, vol. B, pp. 1391–1438. North-Holland, Amsterdam (1993)

An Optimal-Time Algorithm for Shortest Paths on a Convex Polytope in Three Dimensions

Yevgeny Schreiber · Micha Sharir

Abstract We present an optimal-time algorithm for computing (an implicit representation of) the shortest-path map from a fixed source s on the surface of a convex polytope P in three dimensions. Our algorithm runs in $O(n \log n)$ time and requires $O(n \log n)$ space, where n is the number of edges of P. The algorithm is based on the $O(n \log n)$ algorithm of Hershberger and Suri for shortest paths in the plane (Hershberger, J., Suri, S. in SIAM J. Comput. 28(6):2215–2256, 1999), and similarly follows the continuous Dijkstra paradigm, which propagates a "wavefront" from s along ∂P. This is effected by generalizing the concept of conforming subdivision of the free space introduced by Hershberger and Suri and by adapting it for the case of a convex polytope in $\mathbb{R}^3$, allowing the algorithm to accomplish the propagation in discrete steps, between the "transparent" edges of the subdivision. The algorithm constructs a dynamic version of Mount's data structure (Mount, D.M. in Discrete Comput. Geom. 2:153–174, 1987) that implicitly encodes the shortest paths from s to all other points of the surface. This structure allows us to answer single-source shortest-path queries, where the length of the path, as well as its combinatorial type, can be reported in $O(\log n)$ time; the actual path can be reported in additional $O(k)$ time, where k is the number of polytope edges crossed by the path.

Work on this paper was supported by NSF Grants CCR-00-98246 and CCF-05-14079, by a grant from the U.S.-Israeli Binational Science Foundation, by grant 155/05 from the Israel Science Fund, and by the Hermann Minkowski–MINERVA Center for Geometry at Tel Aviv University. The paper is based on the Ph.D. Thesis of the first author, supervised by the second author. A preliminary version has been presented in Proc. 22nd Annu. ACM Sympos. Comput. Geom., pp. 30–39, 2006.

Y. Schreiber (✉) · M. Sharir
School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel
e-mail: syevgeny@tau.ac.il

M. Sharir
Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA
e-mail: michas@tau.ac.il

The algorithm generalizes to the case of m source points to yield an implicit representation of the geodesic Voronoi diagram of m sites on the surface of P, in time $O((n+m)\log(n+m))$, so that the site closest to a query point can be reported in time $O(\log(n+m))$.

Keywords Continuous Dijkstra · Geodesics · Polytope surface · Shortest path · Shortest path map · Unfolding · Wavefront

1 Introduction

1.1 Background

The problem of determining the Euclidean shortest path on the surface of a convex polytope in $\mathbb{R}^3$ between two points, or, more generally, computing a compact representation of all such paths that emanate from a fixed source point s, is a classical problem in geometric optimization, first studied by Sharir and Schorr [36]. Their algorithm, whose running time is $O(n^3 \log n)$, constructs a planar layout of the *shortest path map*, and then the length and combinatorial type of the shortest path from s to any given query point q can be found in $O(\log n)$ time; the path itself can be reported in $O(k)$ additional time, where k is the number of edges of P that are traversed by the shortest path from s to q. Soon afterwards, Mount [27] gave an improved algorithm for convex polytopes with running time $O(n^2 \log n)$. Moreover, in [28], Mount has shown that the problem of storing shortest path information can be treated separately from the problem of computing it, presenting a data structure of $O(n \log n)$ space that supports $O(\log n)$-time shortest-path queries. However, the question whether this data structure can be constructed in subquadratic time, has been left open.

For a general, possibly nonconvex polyhedron P, O'Rourke et al. [31] gave an $O(n^5)$-time algorithm for the single source shortest path problem. Subsequently, Mitchell et al. [26] presented an $O(n^2 \log n)$ algorithm, extending the technique of [27]. All algorithms in [26, 27, 36] use the same general approach, called "continuous Dijkstra", first formalized in [26]. The technique keeps track of all the points on the surface whose shortest path distance to the source s has the same value t, and maintains this "wavefront" as t increases. The approach treats certain elements of ∂P (vertices, edges, or other elements) as nodes in a graph, and follows Dijkstra's algorithm to extract the unprocessed element currently closest to s and to propagate from it, in a continuous manner, shortest paths to other elements. The same general approach is also used in our algorithm.

Chen and Han [8] use a rather different approach (for a not necessarily convex polyhedral surface). Their algorithm builds a shortest path sequence tree, using an observation that they call "one angle one split" to bound the number of branches, maintaining only $O(n)$ nodes in the tree in $O(n^2)$ total running time. The algorithm of [8] also constructs a planar layout of the shortest path map (which is "dual" to the layout of [36]), which can be used similarly for answering shortest path queries in $O(\log n)$ time (or $O(k+\log n)$ time for path reporting). (Their algorithm is somewhat simpler for the case of a convex polytope P, relying on the property, established by Aronov and O'Rourke [6], that this layout of P does not overlap itself.) In [9], Chen

and Han follow the general idea of Mount [28] to solve the problem of storing shortest path information separately, for a general, possibly nonconvex polyhedral surface. They obtain a tradeoff between query time $O(d \log n / \log d)$ and space complexity $O(n \log n / \log d)$, where d is an adjustable parameter. Again, the question whether this data structure can be constructed in subquadratic time, has been left open.

The problem has been more or less "stuck" after Chen and Han's paper, and the quadratic-time barrier seemed very difficult to break. For this and other reasons, several works [2–4, 16, 17, 19, 24, 25, 38] presented approximate algorithms for the 3-dimensional shortest path problem. Nevertheless, the major problem of obtaining a subquadratic, or even near-linear, exact algorithm remained open. In 1999, Kapoor [21] announced such an algorithm for the shortest path problem on an arbitrary polyhedral surface P (see also a review of the algorithm in O'Rourke's column [29]). The algorithm follows the continuous Dijkstra paradigm, and claims to be able to compute a shortest path *between two given points* in $O(n \log^2 n)$ time (so it does not preprocess the surface for answering shortest path queries). However, as far as we know, the details of Kapoor's algorithm have not yet been published.

The Algorithm of Hershberger and Suri for Polygonal Domains A dramatic breakthrough on a loosely related problem took place in 1995,[1] when Hershberger and Suri [18] obtained an $O(n \log n)$-time algorithm for computing shortest paths *in the plane* in the presence of polygonal obstacles (where n is the number of obstacle vertices). The algorithm actually computes a shortest path map from a fixed source point to all other (non-obstacle) points of the plane, which can be used to answer single-source shortest path queries in $O(\log n)$ time.

Our algorithm uses (adapted variants of) many of the ingredients of [18], including the continuous Dijkstra method—in [18], the wavefront is propagated amid the obstacles, where each wave emanates from some obstacle vertex already covered by the wavefront; see Fig. 1(a).

The key new ingredient in [18] is a quad-tree-style subdivision of the plane, of size $O(n)$, on the vertices of the obstacles (temporarily ignoring the obstacle edges).

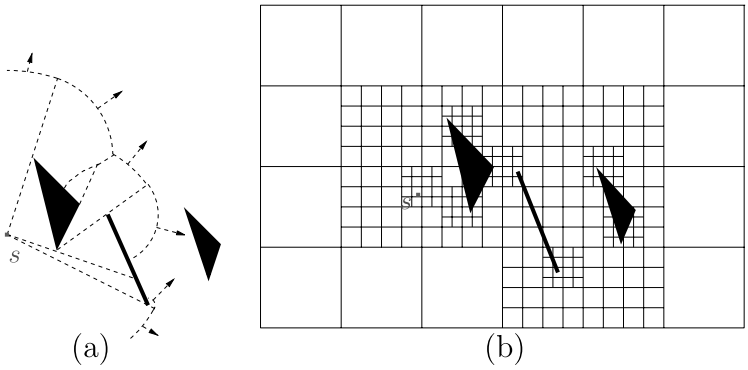


Fig. 1 The planar case: (**a**) The wavefront propagated from s, at some fixed time t. (**b**) The conforming subdivision of the free space

[1] A preliminary (symposium) version has appeared in 1993; the last version was published in 1999.

See Fig. 1(b) for an illustration. Each cell of this *conforming subdivision* is bounded by $O(1)$ axis-parallel straight line edges (called *transparent edges*), contains at most one obstacle vertex, and satisfies the following crucial "conforming" property: For any transparent edge e of the subdivision, there are only $O(1)$ cells within distance $2|e|$ of e. Then the obstacle edges are inserted into the subdivision, while maintaining both the linear size of the subdivision and its conforming property—except that now a transparent edge e has the property that there are $O(1)$ cells within *shortest path distance* $2|e|$ of e. These transparent edges form the elements on which the Dijkstra-style propagation is performed—at each step, the wavefront is ascertained to (completely) cover some transparent edge, and is then advanced into $O(1)$ nearby cells and edges. Since each cell is "simple," the wavefront propagation inside a cell can be implemented efficiently. The conforming nature of the subdivision guarantees the crucial property that each transparent edge e needs to be processed *only once*, in the sense that no path that reaches e after the simulation time at which it is processed can be a shortest path, so the Dijkstra style of propagation works correctly for the transparent edges.

1.2 An Overview of Our Algorithm

As in [18], we construct a conforming subdivision of ∂P to control the wavefront propagation. We first construct an oct-tree-like *3-dimensional* axis-parallel subdivision S_{3D}, only on the vertices of ∂P. Then we intersect S_{3D} with ∂P, to obtain a *conforming surface subdivision* S. (We use the term "facet" when referring to a triangle of ∂P, and we use the term "face" when referring to the square faces of the 3-dimensional cells of S_{3D}. Furthermore, each such face is subdivided into square "subfaces".) In our case, a transparent edge e may traverse many facets of P, but we still want to treat it as a single simple entity. To this end, we first replace each actual intersection ξ of a subface of S_{3D} with ∂P by the *shortest path* on ∂P that connects the endpoints of ξ and traverses the same facet sequence of ∂P as ξ, and make those paths our transparent edges. We associate with each such transparent edge e the *polytope edge sequence* that it crosses, which is stored in compact form and is used to unfold e to a straight segment. To compute the unfolding efficiently, we preprocess ∂P into a *surface unfolding data structure* that allows us to process any such unfolding query in $O(\log n)$ time. This is a nontrivial addition to the machinery of [18] (where the transparent edges are simply straight segments, which are trivial to represent and to manipulate).

However, in order to propagate the wavefront along the surface of P, we have to overcome another difficulty. On top of the main problem that a surface cell may intersect many (up to $\Theta(n)$) facets of P, it can in general be unfolded in more than one way, and such an unfolding may *overlap* itself (see [11]). To overcome this, we introduce a *Riemann structure* that efficiently represents the unfolded regions of the polytope surface that the algorithm processes. This representation subdivides each surface cell into $O(1)$ simple *building blocks* that have the property that a planar unfolding of such a block (a) is unique, and (b) is a simply connected polygon bounded by $O(1)$ straight line segments (and does not overlap itself). A global unfolding is a concatenation of unfolded images of a sequence, or more generally a tree, of certain

blocks. It may overlap itself, but we ignore these overlaps, treating them as different layers of a Riemann surface.

We maintain two *one-sided wavefronts* instead of one exact wavefront at each transparent edge e, so that, for any point $p \in e$, the true shortest path distance from s to p is the smaller of the two distances to p encoded in the two one-sided wavefronts. At each step of the wavefront propagation phase, the algorithm picks up a transparent edge e, constructs each of the one-sided wavefronts at e by *merging* the wavefronts that have already reached *e from a fixed side*, and propagates from e each of its two one-sided wavefronts to $O(1)$ nearby transparent edges f, following the general scheme of [18]. Each propagation that reaches f from e proceeds along a fixed sequence of building blocks that connect e to f. For a fixed edge e, there are only $O(1)$ successor transparent edges f and only $O(1)$ block sequences for any of those f's.

A key difference from [18] is that in our case shortest paths "fold" over ∂P, and need to be unfolded onto some plane (on which they look like straight segments). We cannot afford to perform all these unfoldings explicitly—this would by itself degrade the storage and running time to quadratic in the worst case. Instead we maintain partial unfolding transformations at the nodes of our structure, composing them on the fly (as rigid transformations of 3-space) to perform the actual unfoldings whenever needed.

During each propagation, we keep track of combinatorial changes that occur *within* the wavefront: At each of these events, we either split a wave into two waves when it hits a vertex, or eliminate a wave when it is "overtaken" by its two neighbors. Following a modified variant of the analysis of [18], we show that the algorithm encounters a total of only $O(n)$ "events," and processes each event in $O(\log n)$ time.

After the wavefront propagation phase, we perform further preprocessing to facilitate efficient processing of shortest path queries. This phase is rather different from the shortest path map construction in [18], since we do not provide, nor know how to construct, an explicit representation of the shortest path map on P in $o(n^2)$ time.[2] However, our implicit representation of all the shortest paths from the source suffices for answering any shortest path query in $O(\log n)$ time. The query "identifies" the path combinatorially. It can immediately produce the length of the path (assuming the real RAM model of computation), and the direction at which it leaves s to reach the query point. An explicit representation of the path takes $O(k)$ additional time to compute, where k is the number of polytope edges crossed by the path.

To aid readers familiar with [18], the structure of our paper closely follows that of [18], although each part that corresponds to a part of [18] is quite different in technical details. Section 2 provides some preliminary definitions and describes the construction of the conforming surface subdivision using an already constructed conforming 3D-subdivision S_{3D}, while the construction of S_{3D}, which is slightly more involved, is deferred to Sect. 6 (it is nevertheless very similar to its counterpart in [18], and we only describe the differences between the two procedures). The construction in Sect. 2 is new and involves many ingredients that cater to the spatial structure of convex polytopes. Section 3 also has no parallel in [18]—it presents the Riemann structure, which represents the unfolding of the polytope surface, as needed for the

[2] An explicit representation is tricky in any case, because the map, in its folded form, has quadratic complexity in the worst case.

implementation of the wavefront propagation phase. Section 4 describes the wavefront propagation phase itself. The data structures and the implementation details of the algorithm, as well as the final phase of the preprocessing for shortest path queries, are presented in Sect. 5. We close in Sect. 7 with a discussion, which includes the extension to the construction of *geodesic Voronoi diagrams* on ∂P, and with several open problems.

The full version of the paper [34] is even longer than this journal version—it builds upon the already long paper [18], and adds many new technical steps in full detail. This shorter journal version contains most of its ingredients, but omits certain steps, such as those sufficiently similar to their counterparts in [18].

2 A Conforming Surface Subdivision

A key ingredient of the algorithm is a special subdivision S of ∂P, which we construct in two steps. The first step, sketched in Sect. 6, builds a rectilinear oct-tree-like subdivision S_{3D} of $\mathbb{R}^3$ by taking into account only the vertices of P (see [34, Sect. 6] for details). In the present section, we only state the properties that S_{3D} should satisfy, assume that it is already available, and describe the second step, which constructs S from S_{3D}. We start with some preliminary definitions.

2.1 Preliminaries

Without loss of generality, we assume that s is a vertex of P, that all facets of P are triangles, and that no edge of P is axis-parallel. Our model of computation is the real RAM.

We borrow some definitions from [26, 35, 36]. A *geodesic path* π is a simple path along ∂P so that, for any two sufficiently close points $p, q \in \pi$, the portion of π between p and q is the unique shortest path that connects them on ∂P. Such a path π is always piecewise linear; its length is denoted as $|\pi|$. For any two points $a, b \in \partial P$, a *shortest geodesic path* between them is denoted by $\pi(a, b)$. Generally, $\pi(a, b)$ is unique, but there are degenerate placements of a and b for which there exist several geodesic shortest paths that connect them. For convenience, the word "geodesic" is omitted in the rest of the paper. For any two points $a, b \in \partial P$, at least one shortest path $\pi(a, b)$ exists [26]. We use the notation $\Pi(a, b)$ to denote the set of all shortest paths connecting a and b. The length of any path in $\Pi(a, b)$ is the shortest path distance between a and b, and is denoted as $d_S(a, b)$. We occasionally use $d_S(X, Y)$ to denote the shortest path distance between two compact sets of points $X, Y \subseteq \partial P$, which is the minimum $d_S(x, y)$, over all $x \in X$ and $y \in Y$. We use $d_{3D}(x, y)$ (resp., $d_\infty(x, y)$) to denote the Euclidean (resp., the L_∞) distance in $\mathbb{R}^3$ between x, y; when considering points x, y on a plane, we sometimes denote $d_{3D}(x, y)$ by $d(x, y)$.

If facets f and f' share a common edge χ, the *unfolding* of f' onto (the plane containing) f is the rigid transformation that maps f' into the plane containing f, effected by an appropriate rotation about the line through χ, so that f and the image of f' lie on opposite sides of that line. Let $\mathcal{F} = (f_0, f_1, \ldots, f_k)$ be a sequence of distinct facets such that f_{i-1} and f_i have a common edge χ_i, for $i = 1, \ldots, k$. We say that $\mathcal{F}$ is the *corresponding facet sequence* of the *edge sequence* $\mathcal{E} = (\chi_1, \chi_2, \ldots, \chi_k)$

(and that $\mathcal{E}$ is the corresponding edge sequence of $\mathcal{F}$). The unfolding transformation $U_{\mathcal{E}}$ is the transformation of 3-space that represents the rigid motion that maps f_0 to the plane of f_k, through a sequence of unfoldings at the edges $\chi_1, \chi_2, \ldots, \chi_k$. That is, for $i = 1, \ldots, k$, let φ_i be the rigid transformation of 3-space that unfolds f_{i-1} to the plane of f_i about χ_i. The unfolding $U_{\mathcal{E}}$ is then the composed transformation $\Phi_{\mathcal{E}} = \varphi_k \circ \varphi_{k-1} \circ \cdots \circ \varphi_1$. (The unfolding of an empty edge sequence is the identity transformation.) However, in what follows, we will also use $U_{\mathcal{E}}$ to denote the collection of *all partial unfoldings* $\Phi_{\mathcal{E}}^{(i)} = \varphi_k \circ \varphi_{k-1} \circ \ldots \circ \varphi_i$, for $i = 1, \ldots, k$. Thus $\Phi_{\mathcal{E}}^{(i)}$ is the unfolding of f_{i-1} onto the plane of f_k. The *domain* of $U_{\mathcal{E}}$ is then defined as the union of all points in $f_0, f_1, \ldots, f_k$, and the plane of the last facet f_k is denoted as the *destination plane* of $U_{\mathcal{E}}$. Since each rigid transformation in $\mathbb{R}^3$ can be represented as a 4×4 matrix [32] (see [34] for details), the entire sequence $\Phi_{\mathcal{E}} = \Phi_{\mathcal{E}}^{(1)}, \Phi_{\mathcal{E}}^{(2)}, \ldots, \Phi_{\mathcal{E}}^{(k)}$ can be computed in $O(k)$ time.

The unfolding $U_{\mathcal{E}}(\mathcal{F})$ of the facet sequence $\mathcal{F}$ is the union $\bigcup_{i=0}^{k} \Phi_{\mathcal{E}}^{(i+1)}(f_i)$ of the unfoldings of each of the facets $f_i \in \mathcal{F}$, in the destination plane of $U_{\mathcal{E}}$ (here the unfolding transformation for f_k is the identity).[3] The unfolding $U_{\mathcal{E}}(\pi)$ of a path $\pi \subset \partial P$ that traverses the edge sequence $\mathcal{E}$, is the path consisting of the unfolded images of all the points of π in the destination plane of $U_{\mathcal{E}}$.

The following properties of shortest paths are proved in [8, 26, 35, 36]: (i) The intersection of a shortest path π with any facet f of ∂P is a (possibly empty) line segment. (ii) If π traverses the edge sequence $\mathcal{E}$, then the unfolded image $U_{\mathcal{E}}(\pi)$ is a straight line segment. (iii) A shortest path π never crosses a vertex of P (but it may start or end at a vertex). (iv) Two shortest paths from the same source point s, so that none of them is an extension of the other, cannot intersect each other except at s and, if they have the same destination point, possibly at that point too.

The Elements of the Shortest Path Map We consider the problem of computing shortest paths from a fixed *source* point $s \in \partial P$ to all points of ∂P. A point $z \in \partial P$ is called a *ridge* point if there exist at least two distinct shortest paths from s to z. The *shortest path map* with respect to s, denoted SPM(s), is a subdivision of ∂P into at most n connected regions, called *peels*, whose interiors are vertex-free and contain neither ridge points nor points belonging to shortest paths from s to vertices of P, and such that for each such peel Φ, there is only one shortest path $\pi(s, p) \in \Pi(s, p)$ to any $p \in \Phi$, which also satisfies $\pi(s, p) \subset \Phi$.

There are two types of intrinsic vertices of SPM(s) (excluding intersections of peel boundaries with edges of P): ridge points that are incident to three or more peels, and vertices of P (including s). The boundaries of the peels form the *edges* of SPM(s). There are two types of edges (see Fig. 2): (i) shortest paths from s to a vertex of P, and (ii) *bisectors*, each being a maximal connected polygonal path of ridge points between two vertices of SPM(s) that does not contain any vertex of SPM(s).

It is proved in [36] that: (1) A shortest path from s to any point in ∂P cannot cross a bisector. (2) SPM(s) has only $O(n)$ vertices and (folded) edges, each of which is a union of $O(n)$ straight segments.

[3] Our definition of unfolding is asymmetric, in the sense that we could equally unfold into the plane of any of the other facets of $\mathcal{F}$. We sometimes ignore the exact choice of the destination plane, since the appropriate rigid transformation that moves between these planes is easy to compute.

Denote by $\mathcal{E}_i$ the *maximal polytope edge sequence* crossed by a shortest path from s to a vertex of a peel Φ_i inside Φ_i ($\mathcal{E}_i$ is unique, since Φ_i does not contain polytope vertices in its interior). Denote by s_i the unfolded source image $U_{\mathcal{E}_i}(s)$; for the sake of simplicity, we also denote by s_i the unfolded source image $U_{\mathcal{E}'_i}(s)$, where $\mathcal{E}'_i$ is some prefix of $\mathcal{E}_i$. A bisector between two adjacent peels Φ_i, Φ_j is denoted by $b(s_i, s_j)$. It is the locus of points q equidistant from s_i and s_j (on some common plane), so that there are at least two shortest paths in $\Pi(s, q)$—one, completely contained in Φ_i, traverses a prefix of the polytope edge sequence $\mathcal{E}_i$, and the other, completely contained in Φ_j, traverses a prefix of the polytope edge sequence $\mathcal{E}_j$. Note that for two maximal polytope edge sequences $\mathcal{E}_i, \mathcal{E}_j$, the bisector $b(s_i, s_j)$ between the source images $s_i = U_{\mathcal{E}_i}(s)$ and $s_j = U_{\mathcal{E}_j}(s)$ satisfies both the following properties: $U_{\mathcal{E}_i}(b(s_i, s_j)) \subset U_{\mathcal{E}_i}(\mathcal{F}_i)$, and $U_{\mathcal{E}_j}(b(s_i, s_j)) \subset U_{\mathcal{E}_j}(\mathcal{F}_j)$, where $\mathcal{F}_i, \mathcal{F}_j$ are the respective corresponding facet sequences of $\mathcal{E}_i, \mathcal{E}_j$.

2.2 The 3-Dimensional Subdivision and Its Properties

We begin by introducing the subdivision S_{3D} of $\mathbb{R}^3$, whose construction is sketched in Sect. 6. The subdivision is composed of 3D-cells, each of which is an axis-parallel cube, either whole, or *perforated* by a single axis-parallel cube-shaped hole;[4] see Fig. 3. The boundary face of each 3D-cell is divided into either 16×16 or 64×64 square subfaces with axis-parallel sides.

Let $l(h)$ denote the edge length of a square subface h.

The crucial property of S_{3D} is the *well-covering* of its subfaces. Specifically, a subface h of S_{3D} is said to be *well-covered* if the following three conditions hold:

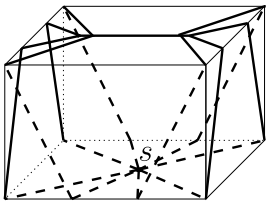


Fig. 2 Peels are bounded by *thick lines* (*dashed and solid*). The bisectors (the set of all the ridge points) are the *thick solid lines*, while the *dashed solid lines* are the shortest paths from s to the vertices of P

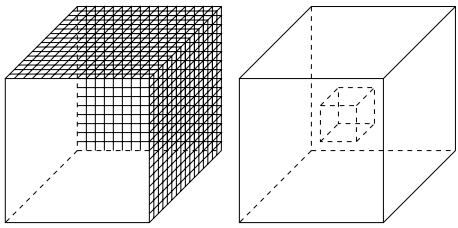


Fig. 3 Two types of a 3D-cell: a whole cube (where the subdivision of three of its faces is shown), and a perforated cube (it is not shown here that each of its inner and outer faces is subdivided into subfaces)

[4]The 3D-subdivision S_{3D} is similar to a (compressed) oct-tree in that all its faces are axis-parallel and their sizes grow by factors of 4. However, the cells of S_{3D} may be nonconvex and the union of the surfaces of the 3D-subdivision itself may be disconnected.

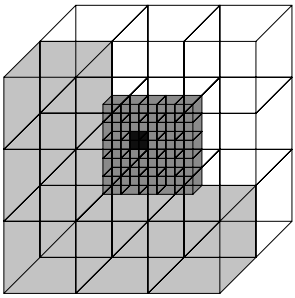


Fig. 4 The well-covering region of the *darkly shaded* face h contains, in this example, a total of 39 3D-cells (nine transparent large cells on the back, five *lightly shaded* large cells on the front, and 25 small cells, also on the front). Each face of the boundary of each 3D-cell in this figure is further subdivided into subfaces (not shown). The well-covering region of each of the subfaces of h coincides with $R(h)$

(W1) There exists a set of $O(1)$ cells $C(h) \subseteq S_{3D}$ such that h lies in the interior of their union $R(h) = \bigcup_{c \in C(h)} c$. The region $R(h)$ is called the *well-covering region* of h (see Fig. 4).
(W2) The total complexity of the subdivisions of the boundaries of all the cells in $C(h)$ is $O(1)$.
(W3) If g is a subface on $\partial R(h)$, then $d_{3D}(h, g) \geq 16 \max\{l(h), l(g)\}$.

A subface h is *strongly well-covered* if the stronger condition (W3′) holds:[5]

(W3′) For any subface g so that h and g are portions of nonadjacent (undivided) faces of the subdivision, $d_{3D}(h, g) \geq 16 \max\{l(h), l(g)\}$.

Let V denote the set of vertices of the polytope (including the source vertex s). A 3D-subdivision S_{3D} is called a *(strongly) conforming 3D-subdivision* for V if the following three conditions hold.

(C1) Each cell of S_{3D} contains at most one point of V in its closure.
(C2) Each subface of S_{3D} is (strongly) well-covered.
(C3) The well-covering region of every subface of S_{3D} contains at most one vertex of V.

S_{3D} also has the following *minimum vertex clearance property*:

(MVC) For any point $v \in V$ and for any subface h, $d_{3D}(v, h) \geq 4l(h)$.

As mentioned, the algorithm for computing a strongly conforming 3D-subdivision of V is sketched in Sect. 6. We state the main result shown there.[6]

Theorem 2.1 (Conforming 3D-subdivision Theorem) *Every set of n points in $\mathbb{R}^3$ admits a strongly conforming 3D-subdivision S_{3D} of $O(n)$ size that also satisfies the*

[5]The wavefront propagation algorithm described in Sects. 4 and 5 requires the subfaces of S_{3D} only to be well-covered, but not necessarily strongly well-covered. The stronger condition (W3′) of subfaces of S_{3D} is needed only in the construction of the surface subdivision S.

[6]Note that we do not assume that the points of V are in convex position.

minimum vertex clearance property. In addition, each input point is contained in the interior of a distinct whole cube cell. Such a 3D-subdivision can be constructed in $O(n \log n)$ *time.*

2.3 Computing the Surface Subdivision

Transparent Edges We intersect the subfaces of S_{3D} with ∂P. Each maximal connected portion ξ of the intersection of a subface h of S_{3D} with ∂P induces *a surface-subdivision (transparent) edge* e of S with the same pair of endpoints. (We textitasize here that $e \neq \xi$. The precise construction of e is detailed below.) A single subface h can therefore induce up to four transparent edges (since P is convex and h is a square, and the construction of S_{3D} ensures that none of its edges is incident to a polytope edge; see Fig. 5). If ξ is a closed cycle fully contained in the interior of h, we break it at its x-rightmost and x-leftmost points (or y-rightmost and y-leftmost points, if h is perpendicular to the x-axis). These two points are regarded as two new endpoints of transparent edges. These endpoints, as well as the endpoints of the open connected intersection portions ξ, are referred to as *transparent endpoints*.

Let $\xi(a, b)$ be a maximal connected portion of the intersection of a subface h of S_{3D} with ∂P, bounded by two transparent endpoints a, b. Let $\mathcal{E} = \mathcal{E}_{a,b}$ denote the sequence of polytope edges that $\xi(a, b)$ crosses from a to b, and let $\mathcal{F} = \mathcal{F}_{a,b}$ denote the facet sequence corresponding to $\mathcal{E}$. We define the *transparent edge* $e_{a,b}$ as the shortest path from a to b *within* the union of $\mathcal{F}$ (a priori, $U_{\mathcal{E}}(e_{a,b})$ is not necessarily a straight segment, but we will shortly show that it is); see Fig. 6. We say that $e_{a,b}$ *originates from the cut* $\xi(a, b)$. Obviously, its length $|e_{a,b}|$ is equal to $|U_{\mathcal{E}}(e_{a,b})| \leq |\xi(a, b)|$. (This initial collection of transparent edges may contain crossing pairs, and

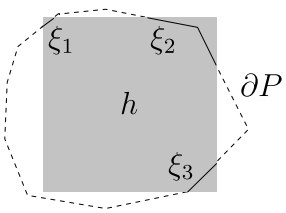


Fig. 5 A subface h and three maximal connected portions ξ_1, ξ_2, ξ_3 that constitute the intersection $h \cap \partial P$

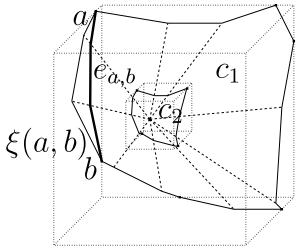


Fig. 6 The cuts of the boundaries of the 3D-cells c_1, c_2 with ∂P are denoted by *thin solid lines*, and the *dashed lines* denote polytope edges. The transparent edge $e_{a,b}$ that originates from the cut $\xi(a, b)$ is *bold*. (To simplify the illustration, this figure ignores the fact that the faces of S_{3D} are actually subdivided into smaller subfaces)

each initial transparent edge will be split into sub-edges at the points where other edges cross it—see below.)

Lemma 2.2 *No polytope vertex can be incident to transparent edges. That is, for each transparent edge $e_{a,b}$, the unfolded path $U_{\mathcal{E}}(e_{a,b})$ is a straight segment.*

Proof By (MVC), for any subface h of S_{3D} and for any $v \in V$, we have $d_{3D}(h, v) \geq 4l(h)$. Let $e_{a,b}$ be a transparent edge originating from $\xi(a, b) \subset h \cap \partial P$. Then $|e_{a,b}| \leq |\xi(a, b)|$, by definition of transparent edges, and $|\xi(a, b)| \leq 4l(h)$, since $\xi(a, b) \subseteq h$ is convex, and h is a square of side length $l(h)$. Therefore $d_{3D}(a, v) \geq |e_{a,b}|$, which shows that $e_{a,b}$ cannot reach any vertex v of P. □

Lemma 2.3 *A transparent endpoint is incident to at least two and at most $O(1)$ transparent edges.*

Proof Easy, and omitted; it follows from the structure of S_{3D}. □

Lemma 2.4 *Each transparent edge that originates from some face ϕ of S_{3D}, meets at most $O(1)$ other transparent edges that originate from faces of S_{3D} adjacent to ϕ (or from ϕ itself), and does not cross any other transparent edges (which originate from faces of S_{3D} not adjacent to ϕ).*

Proof Let $e_{a,b}$ be a transparent edge originating from the cut $\xi(a, b)$, and let $e_{c,d}$ be a transparent edge originating from the cut $\xi(c, d)$. Let h, g be the subfaces of S_{3D} that contain $\xi(a, b)$ and $\xi(c, d)$, respectively. Since $a, b \in h$, we have $d_{3D}(e_{a,b}, h) < \frac{1}{2}|e_{a,b}| \leq \frac{1}{2}|\xi(a, b)| \leq 2l(h)$. Similarly, $d_{3D}(e_{c,d}, g) \leq 2l(g)$. Recall that S_{3D} is a strongly conforming 3D-subdivision. Therefore, if h, g are incident to non-adjacent faces of S_{3D}, then, by (W3′), $d_{3D}(h, g) \geq 16 \max\{l(h), l(g)\}$, hence $e_{a,b}$ does not intersect $e_{c,d}$. Since are only $O(1)$ faces of S_{3D} that are adjacent to the face of h, and each of them contains $O(1)$ subfaces g, there are at most $O(1)$ possible choices of g for each h. □

Splitting Intersecting Transparent Edges Crossing transparent edges are illustrated in Fig. 7. We first show how to compute the intersection points; then, each intersection point is regarded as a new transparent endpoint, splitting each of the two intersecting edges into sub-edges.

Lemma 2.5 *A maximal contiguous facet subsequence that is traversed by a pair of intersecting transparent edges e, e' contains either none or only one intersection point of $e \cap e'$. In the latter case, it contains an endpoint of e or e' (see Fig. 8).*

Proof Consider some maximal common facet subsequence $\tilde{\mathcal{F}} = (f_0, \ldots, f_k)$ that is traversed by e and e', so that the union R of the facets in $\tilde{\mathcal{F}}$ contains an intersection point of $e \cap e'$. Since $\tilde{\mathcal{F}}$ is maximal, no edge of ∂R is crossed by both e and e'; in particular, $\tilde{\mathcal{F}}$ cannot be a single triangle, so $k \geq 1$. Since e and e' are shortest paths

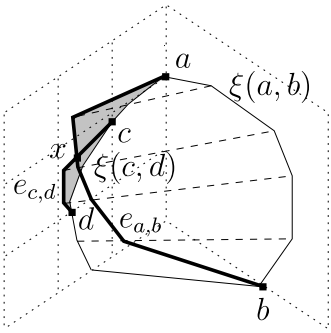


Fig. 7 Subfaces are bounded by *dotted lines*, polytope edges are *dashed*, the cuts of $\partial P \cap S_{3D}$ are *thin solid lines*, and the two transparent edges $e_{a,b}, e_{c,d}$ are drawn as *thick solid lines*. The edges $e_{a,b}, e_{c,d}$ intersect each other at the point $x \in \partial P$; the *shaded region* of ∂P (including the point x on its boundary) lies in this illustration beyond the plane that contains the cut $\xi(c, d)$

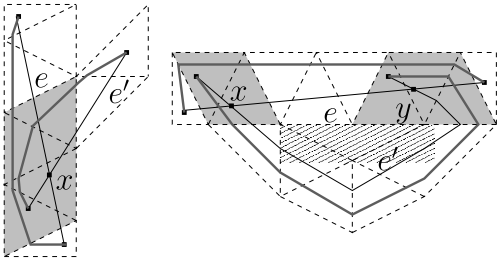


Fig. 8 Two examples of intersecting transparent edges e, e' (*thin solid lines*); the corresponding original cuts (*thick solid lines*) never intersect each other. The maximal contiguous facet subsequences that are traversed by both e, e' and contain an intersection point of $e \cap e'$ are *shaded*. In the second example, the "hole" of ∂P between the facet sequence traversed by e and the facet sequence traversed by e' is hatched

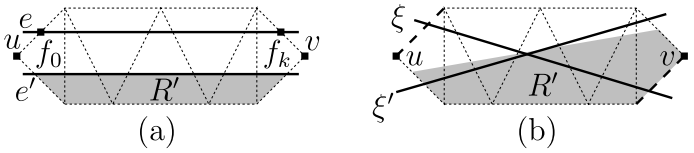


Fig. 9 (**a**) e' divides R into two regions, one of which, R' (*shaded*), contains neither u nor v. (**b**) If R' contains v but not u, ξ' (crossing the same edge sequence as e') intersects ξ (which must cross the bold dashed edges, since R is maximal)

within R, they cannot cross each other (within R) more than once, which proves the first part of the lemma.

To prove the second claim, assume the contrary — that is, R does not contain any endpoint of e and of e'. Denote by u (resp., v) the vertex of f_0 (resp., f_k) that is not incident to f_1 (resp., f_{k-1}). We claim that e' divides R into two regions, one of which contains both u and v, and the other, which we denote by R', contains neither u nor v. Indeed, if each of the two subregions contained exactly one point from $\{u, v\}$ then, by maximality of $\tilde{\mathcal{F}}$, e and e' would have to traverse facet sequences that "cross" each other, which would have forced the corresponding original cuts ξ, ξ' also to cross each other, contrary to the construction; see Fig. 9. The transparent edge e intersects

∂R in exactly two points that are not incident to R'. Since e intersects e' in R, e must intersect $\partial R' \cap e'$ in two points—a contradiction. □

By Lemma 2.4, each transparent edge e has at most $O(1)$ candidate edges that can intersect it (at most four times, as follows from Lemma 2.5). For each such candidate edge e', we can find each of the four possible intersection points, using Lemma 2.5, as follows. First, we check for each of the extreme facets in the facet sequence traversed by e, whether it is also traversed by e', and vice versa (if all the four tests are negative, then e and e' do not intersect each other). We describe in the proof of Lemma 2.11 below how to perform these tests efficiently. For each positive test—when a facet f that is extreme in the facet sequence traversed by one of e, e', is present in the facet sequence traversed by the other—we unfold both e, e' to the plane of f, and find the (image in the plane of f of the) intersection point of $e \cap e'$ that is closest to f (among the two possible intersection points).

Surface Cells After splitting the intersecting transparent edges, the resulting transparent edges are pairwise openly disjoint and subdivide ∂P into connected (albeit not necessarily simply connected) regions bounded by cycles of transparent edges, as follows from Lemma 2.3. These regions, which we call *surface cells*, form a planar (or, rather, spherical) map S on ∂P, which is referred to as the *surface subdivision* of P. Each surface cell is bounded by a set of cycles of transparent edges that are induced by some 3D-cell c_{3D}, and possibly also by a set of other 3D-cells adjacent to c_{3D} whose originally induced transparent edges split the edges originally induced by c_{3D}.

Corollary 2.6 *Each 3D-cell induces at most $O(1)$ (split) transparent edges.*

Proof Follows immediately from the property that the boundary of each 3D-cell consists of only $O(1)$ subfaces, from the fact that each subface induces up to four transparent edges, and from Lemmas 2.4 and 2.5. □

Corollary 2.7 *For each surface cell c, all transparent edges on ∂c are induced by $O(1)$ 3D-cells.*

Proof Follows immediately from Lemma 2.4. □

Corollary 2.8 *Each surface cell is bounded by $O(1)$ transparent edges.*

Proof Follows immediately from Corollaries 2.6 and 2.7. □

Well-Covering We require that *all transparent edges be well-covered* in the surface subdivision S (compare to the well-covering property of the subfaces of S_{3D}), in the following modified sense.

($W1_S$) For each transparent edge e of S, there exists a set $C(e)$ of $O(1)$ cells of S such that e lies in the interior of their union $R(e) = \bigcup_{c \in C(e)} c$, which is called the *well-covering region* of e.

($W2_S$) The total number of transparent edges in all the cells in $C(e)$ is $O(1)$.
($W3_S$) Let e_1 and e_2 be two transparent edges of S such that e_2 lies on the boundary of the well-covering region $R(e_1)$. Then $d_S(e_1, e_2) \geq 2\max\{|e_1|, |e_2|\}$.

As the next theorem shows, our surface subdivision S is a *conforming surface subdivision* for P, in the sense that the following three properties hold.

($C1_S$) Each cell of S is a region on ∂P that contains at most one vertex of P in its closure.
($C2_S$) Each edge of S is well-covered.
($C3_S$) The well-covering region of every edge of S contains at most one vertex of P.

Theorem 2.9 (Conforming Surface-Subdivision Theorem) *Each convex polytope P with n vertices admits a conforming surface subdivision S into $O(n)$ transparent edges and surface cells, constructed as described above.*

Proof The properties ($C1_S$), ($C3_S$) follow from the properties (C1), (C3) of S_{3D}, respectively, and from the fact that each cycle $\mathcal{C}$ of transparent edges that forms a connected component of the boundary of some cell of S traverses the same polytope edge sequence as the original intersections of S_{3D} with ∂P that induce $\mathcal{C}$.

To show well-covering of edges of S (property ($C2_S$)), consider an original transparent edge $e_{a,b}$ (before the splitting of intersecting edges). The endpoints a, b are incident to some subface h that is well-covered in S_{3D}, by a region $R(h)$ consisting of $O(1)$ 3D-cells. We define the well-covering region $R(e)$ of every edge e, obtained from $e_{a,b}$ by splitting, as the connected component containing e, of the union of the surface cells that originate from the 3D-cells of $R(h)$. There are clearly $O(1)$ surface cells in $R(e)$, since each 3D-cell of S_{3D} induces at most $O(1)$ (transparent edges that bound at most $O(1)$) surface cells. $R(e)$ is not empty and it contains e in its interior, since all the surface cells that are incident to e originate from 3D-cells that are incident to h and therefore are in $R(h)$. For each transparent edge e' originating from a subface g that lies on the boundary of (or outside) $R(h)$, $d_S(h, g) \geq d_{3D}(h, g) \geq 16\max\{l(h), l(g)\}$. The length of e satisfies $|e| \leq |e_{a,b}| \leq |\xi(a, b)| \leq 4l(h)$, and, similarly, $|e'| \leq 4l(g)$. Therefore, for each $p \in e$ we have $d_{3D}(p, h) \leq 2l(h)$, and for each $q \in e'$ we have $d_{3D}(q, g) \leq 2l(g)$. Hence, for each $p \in e, q \in e'$, we have $d_S(p, q) \geq d_{3D}(p, q) \geq (16-4)\max\{l(h), l(g)\}$, and therefore $d_S(e, e') \geq 2\max\{|e|, |e'|\}$. □

We next simplify S by deleting (all the transparent edges of) each group of surface cells whose union completely covers exactly one hole of a single surface cell c and contains no vertices of P, thereby eliminating the hole and making it part of c; see Fig. 10. (This optimization clearly does not violate any of the properties of S proved above.) After the optimization, each hole of a surface cell of S must contain a vertex.

The following lemma sharpens a simple property of S that is used later in Sect. 3.

Lemma 2.10 *A transparent edge e intersects any polytope edge in at most one point.*

Proof A polytope edge χ can intersect e at most once, since e is a shortest path (within the union of a facet sequence); since we assume that no edge of P is axis-parallel, $e \cap \chi$ cannot be a nontrivial segment. □

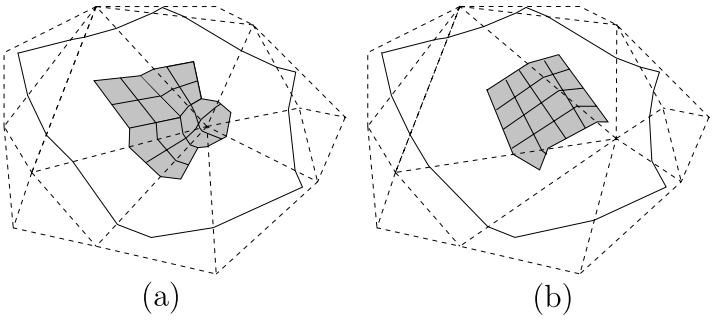


Fig. 10 Simplifying the subdivision (dashed edges denote polytope edges, and solid edges denote transparent edges). (**a**) None of the cells is discarded, since, although the *shaded cells* are completely contained inside a single hole of another cell, one of them contains a vertex of P. (**b**) All the *shaded cells* are discarded, and become part of the containing cell

2.4 The Surface Unfolding Data Structure

In this subsection we present the *surface unfolding data structure*, which we define and use to efficiently construct the surface subdivision. This data structure is also used in Sect. 3 to construct more complex data structures for wavefront propagation and in Sect. 5 by the wavefront propagation algorithm.

Sort the vertices of P in ascending z-order, and sweep a horizontal plane ζ upwards through P. At each height z of ζ, the cross section $P(z) = \zeta \cap P$ is a convex polygon, whose vertices are intersections of some polytope edges with ζ. The cross-section remains combinatorially unchanged, and each of its edges retains a fixed orientation, as long as ζ does not pass through a vertex of P. When ζ crosses a vertex v, the polytope edges incident to v and pointing downwards are deleted (as vertices) from $P(z)$, and those that leave v upwards are added to $P(z)$.

We can represent $P(z)$ by the circular sequence of its vertices, namely the circular sequence of the corresponding polytope edges. We use a linear, rather than a circular, sequence, starting with the x-rightmost vertex of $P(z)$ and proceeding counterclockwise (when viewed from above) along $\partial P(z)$. (It is easy to see that the rightmost vertex of $P(z)$ does not change as long as we do not sweep through a vertex of P.) We use a persistent search tree T_z (with path-copying, as in [20], for reasons detailed below) to represent the cross section. Since the total number of combinatorial changes in $P(z)$ is $O(n)$, the total storage required by T_z is $O(n \log n)$, and it can be constructed in $O(n \log n)$ time.

We can use T_z to perform the following type of query: Given a horizontal subface $h = [a, b] \times [c, d] \times \{z_1\}$ of S_{3D}, compute efficiently the convex polygon $P \cap h$, and represent its boundary in compact form (without computing $P \cap h$ explicitly). We access the value $T_z(z_1)$ of T_z at $z = z_1$ (which represents $P(z_1)$), and compute the intersection points of each of the four edges of h with P. It is easily seen that this can be done in a total of $O(\log n)$ time. We obtain at most eight intersection points, which partition $\partial P(z_1)$ into at most eight portions, and every other portion in the resulting sequence is contained in h. Since these are contiguous portions of $\partial P(z_1)$, each of them can be represented as the disjoint union of $O(\log n)$ subtrees of $T_z(z_1)$, where

the endpoints of the portions (the intersection points of ∂h with $\partial P(z_1)$) do not appear in the subtrees, but can be computed explicitly in additional $O(1)$ time. Hence, we can compute, in $O(\log n)$ time, the polytope edge sequence of the intersection $P \cap h$, and represent it as the disjoint concatenation of $O(\log n)$ canonical sequences, each formed by the edges stored in some subtree of T_z.

We can also use T_z for another (simpler) type of query: Given a facet f of ∂P and some $z = z_1$, locate the endpoints of $f \cap P(z_1)$ (which must be stored at two consecutive leaves in the cyclic order of leaves of the corresponding version of T_z), or report that $f \cap P(z_1) = \emptyset$. As noted above, the *slopes* of the edges of $P(z)$ do not change when z varies, as long as $P(z)$ does not change combinatorially. Moreover, these slopes increase monotonically, as we traverse $P(z_1)$ in counterclockwise direction from its x-leftmost vertex v_L to its x-rightmost vertex v_R, and then again from v_R to v_L. This allows us to locate f in the sequence of edges of $P(z_1)$, in $O(\log n)$ time, by a binary search in the sequence of their slopes. To make binary search possible in $O(\log n)$ time (as well as to enable a somewhat more involved search over T_z that we use in the proof of Lemma 3.12), we store at each node of T_z a pair of pointers to the rightmost and leftmost leaves of its subtree. These extra pointers can be easily maintained during the insertions to and deletions from T_z; it is also easy to see that updating these pointers is coherent with the path-copying method.

However, the most important part of the structure is as follows. With each node ν of T_z, we precompute and store the unfolding U_ν of the sequence $\mathcal{E}_\nu$ of polytope edges stored at the leaves of the subtree of ν, exploiting the following obvious observation. Denote by $\mathcal{F}_\nu$ the corresponding facet sequence of $\mathcal{E}_\nu$. If ν_1, ν_2 are the left and the right children of ν, respectively, then the last facet in $\mathcal{F}_{\nu_1}$ coincides with the first facet of $\mathcal{F}_{\nu_2}$. Hence $U_\nu = U_{\nu_2} \circ U_{\nu_1}$, from which the bottom-up construction of all the unfoldings U_ν is straightforward. Each node stores exactly one rigid transformation, and each combinatorial change in $P(z)$ requires $O(\log n)$ transformation updates, along the path from the new leaf (or from the deleted leaf) to the root. (The rotations that keep the tree balanced do not affect the asymptotic time complexity; maintaining the unfolding information while rebalancing the tree can be performed in a manner similar to that used in another related data structure, described in Sect. 5.1, with full, and fairly routine, details given in [34].) Hence the total number of transformations stored in T_z is $O(n \log n)$ (for all z, including the nodes added to the persistent tree with each path-copying), and they can all be constructed in $O(n \log n)$ time.

Let $\mathcal{F} = (f_0, f_1, \ldots, f_k)$ denote the corresponding facet sequence of the sequence of edges stored at the leaves of T_z at some fixed z. We next show how to use the tree T_z to perform another type of query: Compute the unfolded image $U(q)$ of some point $q \in f_i \in \mathcal{F}$ in the (destination) plane of some other facet $f_j \in \mathcal{F}$ (which is not necessarily the last facet of $\mathcal{F}$), and return the (implicit representation of) the corresponding edge sequence $\mathcal{E}_{ij}$ between f_i and f_j. If $i = j$, then $\mathcal{E}_{ij} = \emptyset$ and $U(q) = q$. Otherwise, we search for f_i and f_j in T_z (in $O(\log n)$ time, as described above). Denote by U_i (resp., U_j) the unfolding transformation that maps the points of f_i (resp., f_j) into the plane of f_k. Then $U(q) = U_j^{-1} U_i(q)$.

We describe next the computation of U_i, and U_j is computed analogously. If f_i equals f_k, then U_i is the identity transformation. Otherwise, denote by ν_i the leaf of T_z that stores the polytope edge $f_i \cap f_{i+1}$, and denote by r the root of T_z. We traverse, bottom up, the path $\mathcal{P}$ from ν_i to r, and compose the transformations stored

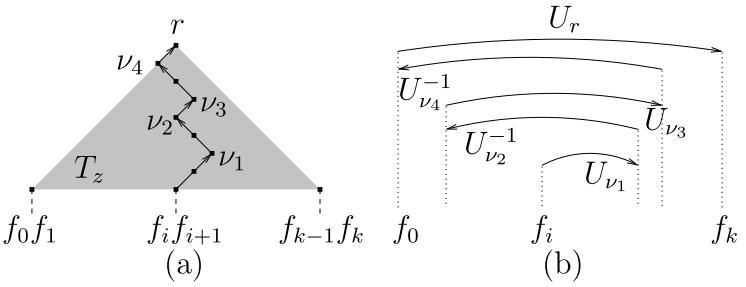


Fig. 11 Constructing U_i by traversing the path from the polytope edge succeeding the facet f_i to the root r of T_z. (**a**) The nodes ν_1, ν_3 are the left turns, and the nodes ν_2, ν_4 are the right turns in this example. (**b**) Composing the corresponding transformations stored at $\nu_1, \ldots, \nu_4$ and at r

at the nodes of $\mathcal{P}$, initializing U_i as the identity transformation and proceeding as follows. We define a node ν of $\mathcal{P}$ to be a *left turn* (resp., *right turn*) if we reach ν from its left (resp., right) child and proceed to its parent ν' so that ν is the right (resp., left) child of ν'. When we reach a left (resp., right) turn ν that stores U_ν, we update $U_i := U_\nu U_i$ (resp., $U_i := U_\nu^{-1} U_i$). If we reach r from its right child, we do nothing; otherwise we update $U_i := U_r U_i$, where U_r is the transformation stored at r. See Fig. 11 for an illustration. Thus, U_i (and U_j) can be computed in $O(\log n)$ time, and so $U(q) = U_j^{-1} U_i(q)$ can be computed in $O(\log n)$ time.

We construct, in a completely symmetric fashion, two additional persistent search trees T_x and T_y, by sweeping P with planes orthogonal to the x-axis and to the y-axis, respectively.

Hence we can compute, in $O(\log n)$ time, the image of any point $q \in \partial P$ in any unfolding formed by a contiguous sequence of polytope edges crossed by an axis-parallel plane that intersects the facet of q. The surface unfolding data structure that answers these queries requires $O(n \log n)$ space and $O(n \log n)$ preprocessing time.

Lemma 2.11 *Given the* 3*D-subdivision* S_{3D}, *the conforming surface subdivision* S *can be constructed in* $O(n \log n)$ *time and space.*

Proof First, we construct the surface unfolding data structure (the enhanced persistent trees T_x, T_y, and T_z) in $O(n \log n)$ time, as described above. Then, for each subface h of S_{3D}, we use the data structure to find $P \cap h$ in $O(\log n)$ time. If $P \cap h$ is a single component, we split it at its rightmost and leftmost points into two portions as described in the beginning of Sect. 2.3—it takes $O(\log n)$ time to locate the split points using a binary search.

To split the intersecting transparent edges, we check each pair of edges (e, e') that might intersect, as follows. First, we find, in the surface unfolding data structure, the edge sequences $\mathcal{E}$ and $\mathcal{E}'$ traversed by e and e', respectively (by locating the cross sections $P \cap h$, $P \cap h'$, where h, h' are the respective subfaces of S_{3D} that induce e, e'). Denote by $\mathcal{F} = (f_0, \ldots, f_k)$ (resp., $\mathcal{F}' = (f'_0, \ldots, f'_{k'})$) the corresponding facet sequence of $\mathcal{E}$ (resp., $\mathcal{E}'$). We search for f_0 in $\mathcal{F}'$, using the unfolding data structure. If it is found, that is, both e and e' intersect f_0, we unfold both edges to the plane of f_0 and check whether they intersect each other within f_0. We search in the same manner

for f_k in $\mathcal{F}'$, and for f'_0 and $f'_{k'}$ in $\mathcal{F}$. This yields up to four possible intersections between e and e' (if all searches fail, e does not cross e'), by Lemma 2.5. Each of these steps takes $O(\log n)$ time. As follows from Lemma 2.4, there are only $O(n)$ candidate pairs of transparent edges, which can be found in a total of $O(n)$ time; hence the whole process of splitting transparent edges takes $O(n \log n)$ time.

Once the transparent edges are split, we combine their pieces to form the boundary cycles of the cells of the surface subdivision. This can easily be done in time $O(n)$. The optimization that deletes each group of surface cells whose union completely covers exactly one hole of a single surface cell and contains no vertices of P also takes $O(n)$ time (using, e.g., DFS on the adjacency graph of the surface cells), since, during the computation of the cell boundaries, we have all the needed information to find the transparent edges to be deleted. □

3 Surface Unfoldings and Shortest Paths

In this section we show how to unfold the surface cells of S and how to represent these unfoldings for the wavefront propagation algorithm (described in Sects. 4 and 5) as *Riemann structures*. Informally, this representation consists of unfolded "flaps," which we call *building blocks*, all lying in a common plane of unfolding. We glue them together locally without overlapping, but they may globally have some overlaps, which however are ignored, since we consider the corresponding flaps to lie at different "layers" of the unfolding.

3.1 Building Blocks and Contact Intervals

Maximal Connecting Common Subsequences Let e and e' be two transparent edges, and let $\mathcal{E} = (\chi_1, \chi_2, \ldots, \chi_k)$ and $\mathcal{E}' = (\chi'_1, \chi'_2, \ldots, \chi'_{k'})$ be the respective polytope edge sequences that they cross. We say that a common (contiguous) subsequence $\tilde{\mathcal{E}}$ of $\mathcal{E}$ and $\mathcal{E}'$ is *connecting* if none of its edges $\tilde{\chi}$ is intersected by a transparent edge between $\tilde{\chi} \cap e$ and $\tilde{\chi} \cap e'$; see Fig. 12(a). We define $G(e, e')$ to be the collection of all *maximal* connecting common subsequences of $\mathcal{E}$ and $\mathcal{E}'$.

Let e and $\mathcal{E}$ be as above, and let v be a vertex of P. Denote by $\mathcal{E}' = (\chi'_1, \chi'_2, \ldots, \chi'_{k'})$ the cyclic sequence of polytope edges that are incident to v, in their counterclockwise order about v. We regard $\mathcal{E}'$ as an infinite cyclic sequence, and we define $G(e, v)$ to be the collection of *maximal* connecting common subsequences of $\mathcal{E}$ and $\mathcal{E}'$, similarly to the definition of $G(e, e')$. See Fig. 12(b).

In either case, the elements of such a collection $G(x, y)$ do not share any polytope edge. We say that a subsequence in $G(x, y)$ *connects* x and y.

The Building Blocks Let c be a cell of the surface subdivision S. Denote by $E(c)$ the set of all the transparent edges on ∂c. Denote by $V(c)$ the set of (zero or one) vertices of P inside c (recall the properties of S). Define $G(c)$ to be the union of all collections $G(x, y)$ so that x, y are distinct elements of $E(c) \cup V(c)$. Fix such a pair of distinct elements $x, y \in E(c) \cup V(c)$. Let $\mathcal{E}_{x,y} = (e_0, e_1, \ldots, e_k) \in G(x, y)$ be a maximal subsequence that connects x and y, and let $\mathcal{F} = (f_0, f_1, \ldots, f_k)$ be its corresponding facet sequence. Define the *shortened facet sequence* of $\mathcal{E}_{x,y}$ to be

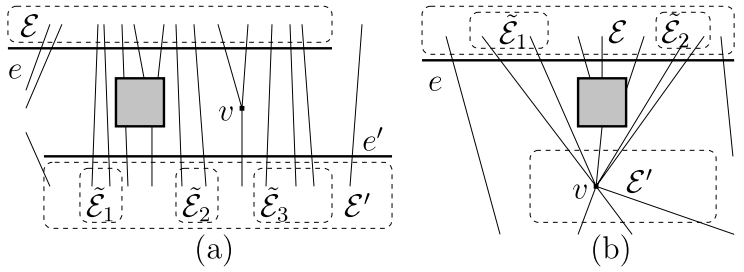


Fig. 12 Maximal connecting common subsequences of polytope edges (drawn as thin solid lines) in (**a**) $G(e, e')$, and (**b**) $G(e, v)$. The transparent edges are drawn thick, and the interiors of the transparent boundary edge cycles that separate $\tilde{\mathcal{E}}_1$ and $\tilde{\mathcal{E}}_2$ are *shaded*

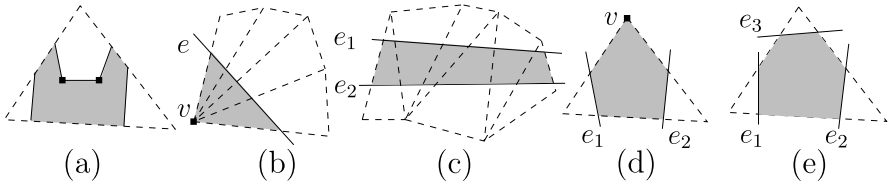


Fig. 13 Building blocks (*shaded*): (**a**), (**b**), (**c**) of types I, II and III, respectively, and (**d**), (**e**) of type IV

$\mathcal{F} \setminus \{f_0, f_k\}$ (so that the extreme edges e_0, e_k of $\mathcal{E}_{x,y}$ are on the boundary of its union), and note that the shortened sequence can be empty (when $k = 1$). We define the following four types of *building blocks* of c.

Type I: Let f be a facet of ∂P. Any connected component of the intersection region $c \cap f$ that meets the interior of f and has an endpoint of some transparent edge of ∂c in its closure is a *building block of type I* of c. See Fig. 13(a) for an illustration.

Type II: Let v be the unique vertex in $V(c)$ (assuming it exists), e a transparent edge in ∂c, and $\mathcal{E}_{e,v} \in G(e, v)$ a maximal subsequence connecting e and v. Then the region B, between e and v in the *shortened* facet sequence of $\mathcal{E}_{e,v}$, if nonempty, is a *building block of type II* of c; see Fig. 13(b).

Type III: Let e, e' be two distinct transparent edges in ∂c, and let $\mathcal{E}_{e,e'} \in G(c)$ be a maximal connecting subsequence between e and e'. The region B between e and e' in the *shortened* facet sequence of $\mathcal{E}_{e,e'}$, if nonempty, is a *building block of type III* of c; see Fig. 13(c).

Type IV: Let f be a facet of ∂P. Any connected component of the region $c \cap f$ that meets the interior of f, does not contain endpoints of any transparent edge, and whose boundary contains a portion of each of the *three* edges of f, is a *building block of type IV* of c. See Fig. 13(d), (e).

We associate with each building block one or two edge sequences along which it can be unfolded. For blocks B contained in a single facet, we associate with B the empty sequence. For other blocks B (which must be of type II or III), the maximal connecting edge sequence $\mathcal{E} = (\chi_1, \ldots, \chi_k)$ that defines B contains at least two polytope edges. Then we associate with B the two *shortened* (possibly empty) sequences

$\mathcal{E}_1 = (\chi_2, \ldots, \chi_{k-1})$, $\mathcal{E}_2 = (\chi_{k-1}, \ldots, \chi_2)$. Note that neither $\mathcal{E}_1$ nor $\mathcal{E}_2$ is cyclic, and that the unfolded images $U_{\mathcal{E}_1}(B)$, $U_{\mathcal{E}_2}(B)$ are congruent.

We say that two distinct points $p, q \in \partial P$ *overlap* in the unfolding $U_{\mathcal{E}}$ of some edge sequence $\mathcal{E}$, if $U_{\mathcal{E}}(p) = U_{\mathcal{E}}(q)$. We say that two sets of surface points $X, Y \subset \partial P$ *overlap* in $U_{\mathcal{E}}$, if there are at least two points $x \in X$ and $y \in Y$ so that $U_{\mathcal{E}}(x) = U_{\mathcal{E}}(y)$. The following lemma states an important property of building blocks (which easily follows from their definition).

Lemma 3.1 *Let c be a surface cell of S, and let B be a building block of c. Let $\mathcal{E}$ be an edge sequence associated with B. Then no two points $p, q \in B$ overlap in $U_{\mathcal{E}}$.*

Proof Easy, and omitted. □

Lemma 3.2 *Let B be a building block of type IV of a surface cell c, and let f be the facet that contains B. Then either* (*a*) *B is a convex pentagon, bounded by portions of the three edges of f, a vertex of f, and portions of two transparent edges* (*see Fig.* 13(*d*)), *or* (*b*) *B is a convex hexagon, whose boundary alternates between portions of the edges of f and portions of transparent edges* (*see Fig.* 13(*e*)). *In the latter case, B contains no vertices of P* (*i.e., of f*).

Proof Easy, and omitted. □

Corollary 3.3 *Let B be a building block of type II, III, or IV, and let $\mathcal{E}$ be an edge sequence associated with B. Then $U_{\mathcal{E}}(B)$ is convex.*

Proof If B is of type II, then $U_{\mathcal{E}}(B)$ is a triangle, by construction. If B is of type IV, then by Lemma 3.2, $U_{\mathcal{E}}(B) = B$ is a convex pentagon or hexagon. If B is of type III, then $U_{\mathcal{E}}(B)$ is a convex quadrilateral, by construction. □

Corollary 3.4 *There are no holes in building blocks.*

Proof Immediate for blocks of type II, III, IV, and follows for blocks of type I from the optimization procedure described after the proof of Theorem 2.9. □

Lemma 3.5 *Any surface cell c has only $O(1)$ building blocks.*

Proof There are $O(1)$ transparent edges in c (by construction of S), and therefore $O(1)$ transparent endpoints, and each endpoint x can be incident to at most one building block of c of type I (or to at most two such blocks, if our general position assumption is not strong enough—in that case x may be incident to an edge, but not to a vertex, of P).

There are $O(1)$ transparent edges and at most one vertex of P in c, by construction of S. Therefore there are at most $O(1)$ pairs (e', v) in c so that e' is a transparent edge and v is a vertex of P. Since there are at most $O(1)$ transparent edge cycles in ∂c that intersect polytope edges delimited by v and crossed by e', and since each such cycle can split the connecting sequence of polytope edges between e' and v at most

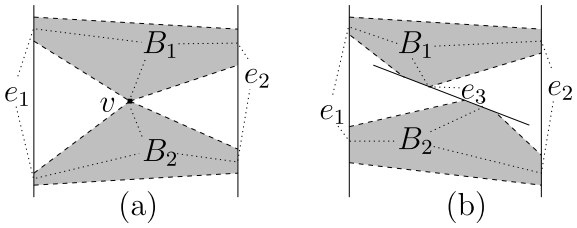


Fig. 14 The triple, of (**a**) two transparent edges and a vertex of P, or (**b**) three transparent edges, contributes to two building blocks B_1, B_2. The corresponding graphs $K_{3,2}$ are illustrated by dotted lines. If the triple contributed to three building blocks, we would have obtained an impossible plane drawing of $K_{3,3}$

once, there are at most $O(1)$ maximal connecting common subsequences in $G(e', v)$. Hence, there are $O(1)$ building blocks of type II of c.

Similarly, there are $O(1)$ pairs of transparent edges (e', e'') in c. There are at most $O(1)$ other transparent edges and at most one vertex of P in c that can lie between e' and e'', resulting in at most $O(1)$ maximal connecting common subsequences in $G(e', e'')$. Hence, there are $O(1)$ building blocks of type III of c.

By Lemma 3.2, the boundary of a building block B of type IV contains either two transparent edge segments and a polytope vertex or three transparent edge segments. In either case, we say that this *triple* of elements (either two transparent edges and a vertex of P, or three transparent edges) *contributes* to B. We claim that one triple can contribute to at most two building blocks of type IV (see Fig. 14). Indeed, if a triple, say, (e_1, e_2, e_3), contributed to three type IV blocks B_1, B_2, B_3, we could construct from this configuration a plane drawing of the graph $K_{3,3}$ (as is implied in Fig. 14), which is impossible. There are $O(1)$ transparent edges and at most one vertex of P in c, by construction of S; therefore there are at most $O(1)$ triples that contribute to at most $O(1)$ building blocks of type IV of c. □

Lemma 3.6 *The interiors of the building blocks of a surface cell c are pairwise disjoint.*

Proof The polytope edges subdivide c into pairwise disjoint components (each contained in a single facet of P). Each building block of type I or IV contains (and coincides with) exactly one such component, by definition. Each building block of type II or III contains one or more such components, and each component is fully contained in the block. Hence it suffices to show that no two distinct blocks can share a component; the proof of this claim is easy, and omitted. □

Let B be a building block of a surface cell c. A *contact interval* of B is a maximal straight segment of ∂B that is incident to one polytope edge $\chi \subset \partial B$ and is not intersected by transparent edges, except at its endpoints. See Fig. 13 for an illustration (contact intervals are drawn as dashed segments on the boundary of the respective building blocks). Our propagation algorithm considers portions of shortest paths that traverse a surface cell c from one transparent edge bounding c to another such edge. Such a path, if not contained in a single building block, traverses a sequence of such blocks, and crosses from one such block to the next through a common contact interval.

Lemma 3.7 *Let c be a surface cell, and let B be one of its building blocks. Then B has at most $O(1)$ contact intervals. If B is of type II or III, then it has exactly two contact intervals, and if B is of type IV, it has exactly three contact intervals.*

Proof If B is of type I, then B is a (simply connected) polygon contained in a single facet f, so that every segment of ∂B is either a transparent edge segment or a segment of a polytope edge bounding f (transparent edges cannot overlap polytope edges, by Lemma 2.10). Every transparent edge of c can generate at most one boundary segment of B, since it intersects ∂f at most twice. There are $O(1)$ transparent edges, and at most one vertex of P in c, by construction of S. Since each contact interval of B is bounded either by two transparent edges or by a transparent edge and a vertex of P, it follows that B has at most $O(1)$ contact intervals.

If B is of type II, III, or IV, the claim is immediate. □

Corollary 3.8 *Let $I_1 \neq I_2$ be two contact intervals of any pair of building blocks. Then either I_1 and I_2 are disjoint, or their intersection is a common endpoint.*

Proof By definition. □

Lemma 3.9 *Let c be a surface cell. Then each point of c that is not incident to a contact interval of any building block of c, is contained in (exactly) one building block of c.*

Proof Fix a point $p \in c$, and denote by f the facet that contains p. Denote by Q the connected component of $c \cap f$ that contains p. If Q contains in its closure at least one endpoint of some transparent edge of ∂c, then p is in a building block of type I, by definition.

Otherwise, Q must be a convex polygon, bounded by portions of transparent edges and by portions of edges of f; the boundary edges alternate between transparent edges and polytope edges, with the possible exception of a single pair of consecutive polytope edges that meet at the unique vertex v of f that lies in c. Thus only the following cases are possible: (1) Q is a triangle bounded by the two edges χ_1, χ_2 of f that meet at v and by a transparent edge e. See Fig. 15(a). The subsequence (χ_1, χ_2) connects e and v, hence p is in a building block of type II (f clearly lies in the shortened facet sequence). (2) Q is a quadrilateral bounded by the two edges χ_1, χ_2 of f and by two transparent edges e_1, e_2. See Fig. 15(b). Then (χ_1, χ_2) connects

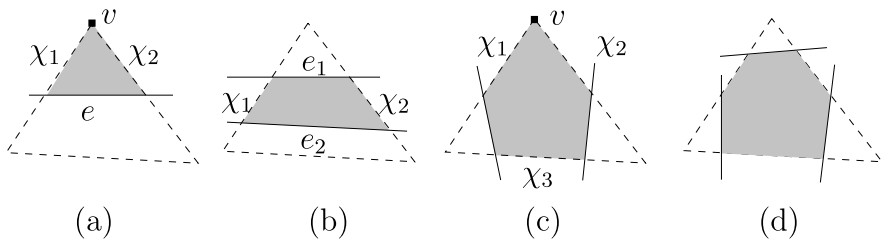


Fig. 15 If Q (shaded) does not contain a transparent endpoint, it must be either a portion of a building block of (**a**) type II or (**b**) type III, or (**c**), (**d**) a building block of type IV

e_1 and e_2, hence p is in a building block of type III (again, f lies in the shortened facet sequence). (3) Q is a pentagon bounded by the two edges χ_1, χ_2 of f incident to v, by two transparent edges, and by the third edge χ_3 of f. See Fig. 15(c). Then p lies in a building block of type IV. (4) Q is a hexagon bounded by all three edges of f and by three transparent edges. See Fig. 15(d). Again, by definition, p lies in a building block of type IV. This (and the disjointness of building blocks established in Lemma 3.6) completes the proof of the lemma. □

The following two auxiliary lemmas are used in the proof of Lemma 3.12, which gives an efficient algorithm for computing (the boundaries of) all the building blocks of a single surface cell.

Lemma 3.10 *Let c be a surface cell. We can compute the boundaries of all the building blocks of c of type I in $O(\log n)$ total time.*

Proof We compute the boundary of each such block by a straightforward iterative process that starts at a transparent endpoint a lying in some facet f of P, and traces the block boundary from a along an alternating sequence of transparent edges and edges of f (with the possible exception of traversing, once, two consecutive edges of f through a common vertex), until we get back to a.

Since, by Corollary 3.4, there are no holes inside building blocks, after each boundary tracing step we compute one building block of type I of c. Hence, by Lemma 3.5, there are $O(1)$ iterations. In each iteration we process $O(1)$ segments of the current building block boundary. Processing each segment takes $O(\log n)$ time, since it involves unfolding $O(1)$ transparent edges in $O(\log n)$ time, using the surface unfolding data structure. (Although we work in a single facet f, each transparent edge that we process is represented relative to its destination plane, which might be incident to another facet of P. Thus we need to unfold it to obtain its portion within f.) □

Lemma 3.11 *We can compute the boundaries of all the building blocks that are incident to vertices of P in total $O(n \log n)$ time.*

Proof Let c be a surface cell that contains some (unique) vertex v of P in its interior. Denote by $\mathcal{F}_v$ the cyclic sequence of facets that are incident to v. Compute all the building blocks of type I of c in $O(\log n)$ time, applying the algorithm of Lemma 3.10. Denote by $\mathcal{H}$ the set of facets in $\mathcal{F}_v$ that contain building blocks of c of type I *that are incident to* v. Denote by $\mathcal{Y}$ the set of maximal contiguous subsequences that constitute $\mathcal{F}_v \setminus \mathcal{H}$. To compute $\mathcal{Y}$, we locate each facet of $\mathcal{H}$ in $\mathcal{F}_v$, and then extract the contiguous portions of $\mathcal{F}_v$ between those facets. To traverse $\mathcal{F}_v$ around each vertex v of P takes a total of $O(n)$ time (since we traverse each facet of P exactly three times).

We process $\mathcal{Y}$ iteratively. Each step picks a nonempty sequence $\mathcal{F} \in \mathcal{Y}$ and traverses it, until a building block of type II or IV is found and extracted from $\mathcal{F}$.

Let $\mathcal{F}$ be a sequence in $\mathcal{Y}$. Since there are no cyclic transparent edges, by construction, it easily follows that $\mathcal{H} \cap \mathcal{F}_v \neq \emptyset$, and therefore $\mathcal{F}$ is not cyclic. Denote the facets of $\mathcal{F}$ by $f_1, \ldots, f_k$, with $k \geq 1$. Denote by $(\chi_1, \ldots, \chi_{k-1})$ the corresponding

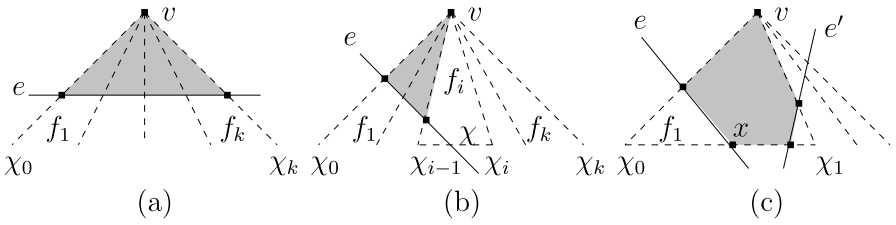


Fig. 16 Extracting from $\mathcal{F}$ building blocks (drawn *shaded*) of type II (cases (**a**), (**b**)) or IV (case (**c**))

polytope edge sequence of $\mathcal{F}$ (if $k = 1$, it is an empty sequence). If $k > 1$, denote by χ_0 the edge of f_1 that is incident to v and does not bound f_2, and denote by χ_k the edge of f_k that is incident to v and does not bound f_{k-1}. Otherwise ($k = 1$), denote by χ_0, χ_1 the polytope edges of f_1 that are incident to v. Among all the $O(1)$ transparent edges of ∂c, find the transparent edge e that intersects χ_0 closest to v (by unfolding all these edges and finding their intersections with χ_0). We traverse $\mathcal{F}$ either until it ends, or until we find a facet $f_i \in \mathcal{F}$ so that e intersects χ_{i-1} but does not intersect χ_i (that is, e intersects the polytope edge $\chi \subset \partial f_i$ that is opposite to v). Note that $\mathcal{F}$ cannot be interrupted by a hole in c, since the endpoints of the transparent edges of such a hole lie in blocks of type I, which belong to $\mathcal{H}$.

In the former case (see Fig. 16(a)), mark the region of ∂P between e, χ_0, and χ_k as a building block of type II, delete $\mathcal{F}$ from $\mathcal{Y}$, and terminate this iteration of the loop. In the latter case, there are two possible cases. If $i > 1$ (see Fig. 16(b)), mark the region of ∂P between e, χ_0, and χ_{i-1} as a building block of type II, delete $f_1, f_2, \ldots, f_{i-1}$ from $\mathcal{F}$, and terminate this iteration of the loop. Otherwise ($f_i = f_1$), denote by x the intersection point $e \cap \chi$, and denote by χ' the portion of χ whose endpoint is incident to χ_1. Among all transparent edges of ∂c, find the transparent edge e' that intersects χ' closest to x (such an edge must exist, or else c would contain two vertices of P). The edge e' must intersect χ_1, since otherwise f_i would contain a building block of type I incident to v, and thus would belong to $\mathcal{H}$. See Fig. 16(c) for an illustration. Mark the region bounded by $\chi_0, \chi_1, \chi, e, e'$ as a building block of type IV, and delete f_1 from $\mathcal{F}$.

At each iteration we compute a single building block of c, hence there are only $O(1)$ iterations. We traverse the facet sequence around v twice (once to compute $\mathcal{Y}$, and once during the extraction of building blocks), which takes $O(n)$ total time for all vertices of P. At each iteration we perform $O(1)$ unfoldings (as well as other constant-time operations), hence the total time of the procedure for all the cells of S is $O(n \log n)$. □

Lemma 3.12 *We can compute* (*the boundaries of*) *all the building blocks of all the surface cells of S in total $O(n \log n)$ time.*

Proof Let c be a surface cell. Compute the boundaries of all the (unfoldings of the) building blocks of c of types I and II, and the building blocks of type IV that contain the single vertex v of P in c, applying the algorithms of Lemmas 3.10 and 3.11. Denote the set of all these building blocks by $\mathcal{H}$. (Note that $\mathcal{H}$ cannot be empty, because ∂c contains at least two transparent edges, which have at least two endpoints

that are contained in at least one building block of type I.) Construct the list L of the contact intervals of all the building blocks in $\mathcal{H}$. For each contact interval I that appears in L twice, remove both instances of I from L. If L becomes (or was initially) empty, then $\mathcal{H}$ contains all the building blocks of c. Otherwise, each interval in L is delimited by two transparent edges, since all building blocks that contain v are in $\mathcal{H}$. Each contact interval in L bounds two building blocks of c, one of which is in $\mathcal{H}$ (it is either of type I or contains a vertex of P in its closure), and the other is not in $\mathcal{H}$ and is either of type III or a convex hexagon of type IV. The union of all building blocks of c that are not in $\mathcal{H}$ consists of several connected components. Since there are no blocks of $\mathcal{H}$ among the blocks in a component, neither transparent edges nor polytope edges terminate inside it; therefore such a component is not punctured (by boundary cycles of transparent edges or by a vertex of P), and its boundary alternates between contact intervals in L and portions of transparent edges. For each contact interval I in L, denote by $\mathit{limits}(I)$ the pair of transparent edges that delimit it.

Denote by $\mathcal{Y}$ the partition of contact intervals in L into cyclic sequences, so that each sequence bounds a different component, and so that each pair of consecutive intervals in the same sequence are separated by a single transparent edge. By construction, each contact interval in $\mathcal{Y}$ appears in a unique cycle. Since there are only $O(1)$ building blocks of c, we can compute the sequences of $\mathcal{Y}$ in constant time. Let $Y = (I_1, I_2, \ldots, I_k)$ be a cyclic sequence in $\mathcal{Y}$ (with $I_{zk+l} = I_l$, for any $l = 1, \ldots, k$ and any $z \in \mathbb{Z}$). Then, for every pair of consecutive intervals $I_j, I_{j+1} \in Y$, $\mathit{limits}(I_j) \cap \mathit{limits}(I_{j+1})$ is nonempty, and consists of one or two transparent edges (two if the cyclic sequence at hand is a doubleton). Obviously, any cyclic sequence in $\mathcal{Y}$ contains two or more contact intervals. As argued above, the portion of ∂P bounded by these contact intervals and by their connecting transparent edges is a portion of c which consists of only building blocks of types III and IV. In particular, it does not contain in its interior any vertex of P, nor any transparent edge.

We process $\mathcal{Y}$ iteratively. Each step picks a sequence $Y \in \mathcal{Y}$, and, if necessary, splits it into subsequences, each time extracting a single building block of type III or IV, as follows.

If Y contains exactly two contact intervals, they must bound a single building block of type III, which we can easily compute, and then discard Y. Otherwise, let I_{j-1}, I_j, I_{j+1} be three consecutive contact intervals in Y, and denote by $\chi_{j-1}, \chi_j, \chi_{j+1}$ the (distinct) polytope edges that contain I_{j-1}, I_j and I_{j+1}, respectively. Define the common bounding edge $e_j = \mathit{limits}(I_j) \cap \mathit{limits}(I_{j+1})$ (there is only one such edge, since $|Y| > 2$), and denote by $\mathcal{E}_j$ the polytope edge sequence intersected by e_j. Similarly, define $\mathcal{E}_{j-1}$ as the polytope edge sequence traversed by the transparent edge $e_{j-1} = \mathit{limits}(I_{j-1}) \cap \mathit{limits}(I_j)$. Without loss of generality, assume that both $\mathcal{E}_{j-1}$ and $\mathcal{E}_j$ are directed from χ_j, to χ_{j-1} and to χ_{j+1}, respectively. See Fig. 17.

We claim that $\bar{\mathcal{E}} = \mathcal{E}_{j-1} \cap \mathcal{E}_j$ is a contiguous subsequence of both sequences. Indeed, assume to the contrary that $\bar{\mathcal{E}}$ contains at least two subsequences $\bar{\mathcal{E}}_1, \bar{\mathcal{E}}_2$, and there is an edge $\bar{\chi}$ between them that belongs to only one of the sequences $\mathcal{E}_{j-1}, \mathcal{E}_j$. Then the region R of ∂P between the last edge of $\bar{\mathcal{E}}_1$, the first edge of $\bar{\mathcal{E}}_2$, e_{j-1} and e_j is contained in the region bounded by the contact intervals of Y and by their connecting transparent edges, and $\bar{\chi}$ must have an endpoint in R, contradicting the fact that this region does not contain any vertex of P. We can therefore use a binary

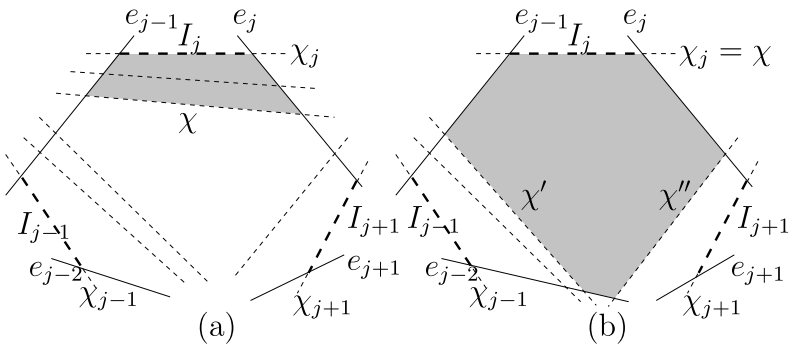


Fig. 17 There are two possible cases: (**a**) There is more than one edge in $\bar{\mathcal{E}}$, hence a building block of type III (whose unfolded image is shown *shaded*) can be extracted. (**b**) $|\bar{\mathcal{E}}| = 1$ (that is, $\chi_j = \chi$), therefore there must be a building block of type IV (whose image is shown shaded) that can be extracted

search to find the last polytope edge χ in $\bar{\mathcal{E}}$, by traversing the unfolding data structure tree T that contains $\mathcal{E}_{j-1}$ from the root r to the leaf that stores χ. To facilitate this search, we first search for ξ_j, which is the first edge of $\bar{\mathcal{E}}$. We then trace the search path $\mathcal{P}$ bottom-up. For each node μ on the path for which the path continues via its left child, we go to the right child ν, and test whether the edges stored at its leftmost leaf and rightmost leaf belong to the portion of $\mathcal{E}_j$ between χ_j and χ_{j+1}; for the sake of simplicity, we refer to this portion as $\mathcal{E}_j$. (As we will shortly argue, each of these tests can be performed in $O(1)$ time.) If both edges belong to $\mathcal{E}_j$, we continue up $\mathcal{P}$. If neither of them is in $\mathcal{E}_j$, then χ is stored at the rightmost leaf of the left child of μ. If only one of them (namely, the one at the leftmost leaf) is in $\mathcal{E}_j$, we go to ν, and start tracing a path from ν to the leaf that stores χ. At each step, we go to the left (resp., right) child if its rightmost leaf stores an edge that belongs (resp., does not belong) to $\mathcal{E}_j$.

To test, in $O(1)$ time, whether an edge $\chi^\star$ of P belongs to $\mathcal{E}_j$, we first recall that, by construction, all the edges of $\mathcal{E}_j$ intersect the original subface h_j of S_{3D} from which e_j originates, and so they appear as a contiguous subsequence of the sequence of edges of P stored at the surface unfolding data structure at the appropriate x-, y-, or z-coordinate of h_j. Moreover, the slopes of the segments that connect them in the corresponding cross-section of P (which are the cross-sections of the connecting facets) are sorted in increasing order.

We thus test whether $\chi^\star$ intersects h_j. We then test whether the slope of the cross-section of the facet that precedes $\chi^\star$ lies within the range of slopes of the facets between the edges χ_j and χ_{j+1}. Clearly, $\chi^\star$ belongs to $\mathcal{E}_j$ if and only if both tests are positive. Since each of these tests takes $O(1)$ time, the claim follows. Hence, we can construct $\bar{\mathcal{E}}$ in $O(\log n)$ time.

If $\chi \neq \chi_j$, then we find the unfoldings $U_{\bar{\mathcal{E}}}(e_j)$ and $U_{\bar{\mathcal{E}}}(e_{j-1})$ and compute a new contact interval I'_j that is the portion of χ bounded by e_j and e_{j-1}. See Fig. 17(a). The quadrilateral bounded by $U_{\bar{\mathcal{E}}}(e_j)$, $U_{\bar{\mathcal{E}}}(e_{j-1})$, $U_{\bar{\mathcal{E}}}(I'_j)$ and $U_{\bar{\mathcal{E}}}(I_j)$ is the unfolded image of a building block of type III. Delete I_j from Y and replace it by I'_j.

Otherwise, $\chi = \chi_j$. See Fig. 17(b). Denote by χ' (resp., χ'') the second edge in $\mathcal{E}_{j-1}$ (resp., $\mathcal{E}_j$); clearly, $\chi' \neq \chi''$. Since all blocks that contain either a vertex of P

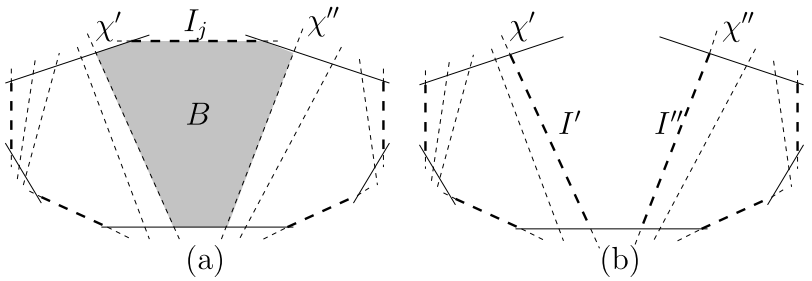


Fig. 18 (**a**) Before the extraction of B, Y contains five (*bold dashed*) contact intervals. (**b**) After the extraction of B, Y has been split into two new (cyclic) sequences Y', Y'' containing the respective contact intervals I', I''. I_j is no longer contained in any sequence in $\mathcal{Y}$

or a transparent edge endpoint are in $\mathcal{H}$, the edges χ_j, χ', χ'' bound a single facet, and there is a transparent edge that intersects both χ', χ'' (otherwise the block of type IV that we are extracting would be bounded by at least four polytope edges—a contradiction). Denote by e the transparent edge that intersects both χ', χ'' nearest to χ_j or, rather, nearest to e_{j-1} and to e_j, respectively (in Fig. 17(b) we have $e = e_{j-2}$). The region bounded by χ_j, χ', χ'' and e_{j-1}, e_j, e is a hexagonal building block of type IV. Compute its two contact intervals that are contained in χ' and χ'', and insert them into Y instead of I_j. If χ' contains I_{j-1} and χ'' contains I_{j+1}, Y is exhausted, and we terminate its processing. If χ' contains I_{j-1} and χ'' does not contain I_{j+1}, we remove I_j and I_{j-1} from Y and replace them by the portion of χ'' between e and e_j. Symmetric actions are taken when χ'' contains I_{j+1} and χ' does not contain I_{j-1}. Finally, if χ' does not contain I_{j-1}, nor does χ'' contain I_{j+1}, we split Y into two new cyclic subsequences, as shown in Fig. 18, and insert them into $\mathcal{Y}$ instead of Y.

In each iteration we compute the boundary of a single building block of type III or IV, hence there are $O(1)$ iterations; each performs $O(1)$ unfoldings, $O(1)$ binary searches, and $O(1)$ operations on constant-length lists, hence the time bound follows. □

3.2 Block Trees and Riemann Structures

In this section we combine the building blocks of a single surface cell into more complex structures.

Let e be a transparent edge on the boundary of some surface cell c, and let B be a building block of c so that e appears on its boundary. The *block tree* $T_B(e)$ is a rooted tree whose nodes are building blocks of c that is defined recursively as follows. The root of $T_B(e)$ is B. Let B' be a node in $T_B(e)$. Then its children are the blocks B'' that satisfy the three following conditions.

(1) B' and B'' are adjacent through a common contact interval;
(2) B'' does not appear as a node on the path in $T_B(e)$ from the root to B', except possibly as the root itself (that is, we allow $B'' = B$ if the rest of the conditions are satisfied);

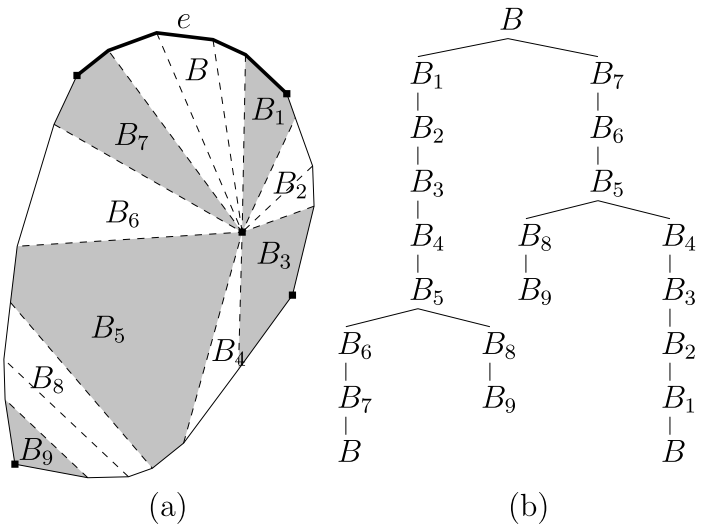


Fig. 19 (**a**) A surface cell c containing a single vertex of P and bounded by four transparent edges (*solid lines*) is partitioned in this example into ten building blocks (whose shadings alternate): B_1, B_3, B_7, B_9 are of type I, B, B_2, B_4, B_6 are of type II, B_8 of type III and B_5 of type IV. Adjacent building blocks are separated by contact intervals (*dashed lines*; other polytope edges are also drawn *dashed*). (**b**) The tree $T_B(e)$ of building blocks of c, where e is the (thick) transparent edge that bounds the building block B

(3) if $B'' = B$, then (a) it is of type II or III (that is, if a root is a building block of type I or IV, it cannot appear as another node of the tree), and (b) it is a leaf of the tree.

Note that a block may appear more than once in $T_B(e)$, but no more than once on each path from the root to a leaf, except possibly for the root B, which may also appear at leaves of $T_B(e)$ if it is of type II or III. However, B cannot appear in any other internal node of $T_B(e)$—see Fig. 19.

Remark Here is a motivation for the somewhat peculiar way of defining $T_B(e)$ (reflected in properties (2) and (3)). Since each building block is either contained in a single facet (and a single facet is never traversed by a shortest path in more than one connected segment), or has exactly two contact intervals (and a single contact interval is never crossed by a shortest path more than once), a shortest path $\pi(s, q)$ to a point q in a building block B may traverse B *through its contact intervals* in no more than two connected segments. Moreover, B may be traversed (through its contact intervals) in *two* such segments only if the following conditions hold: (i) $\pi(s, q)$ must enter B through a point p on a transparent edge on ∂c, (ii) B consists of components of at least two facets, and p and q are contained in two distinct facets, relatively "far" from each other in B, and (iii) $\pi(p, q)$ exits B through one contact interval and then re-enters B through another (before reaching q). See Fig. 20 for an illustration. This shows that the initial block B through which a shortest path from s enters a cell c may be traversed a second time, but only if it is of type II or III. After the second time, the path must exit c right away, or end inside B.

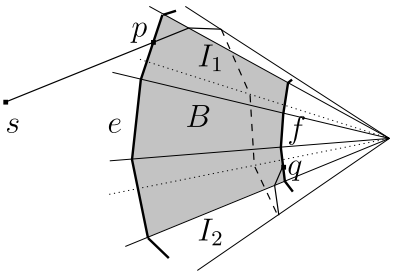


Fig. 20 The shortest path $\pi(s, q)$ enters the (*shaded*) building block B through the transparent edge e at the point p, leaves B through the contact interval I_1, and then reenters B through the contact interval I_2

We denote by $\mathcal{T}(e)$ the set of all block trees $T_B(e)$ of e (constructed from the building blocks of both cells containing e on their boundaries). Note that each block tree in $\mathcal{T}(e)$ contains only building blocks of one cell. We call $\mathcal{T}(e)$ the *Riemann surface structure of* e; it will be used in Sect. 5 for wavefront propagation block-by-block from e in all directions (this is why we include in it block trees of both surface cells that share e on their boundaries). This structure is indeed similar to standard Riemann surfaces (see, e.g., [39]); its main purpose is to handle effectively (i) the possibility of *overlap* between distinct portions of ∂P when unfolded onto some plane, and (ii) the possibility that shortest paths may traverse a cell c in "homotopically inequivalent" ways (e.g., by going around a vertex or a hole of c in two different ways—see below).

Remark Concerning (i), note that without the Riemann structure, unfolding an arbitrary portion of ∂P may result in a self-overlapping planar region (making it difficult to apply the propagation algorithm)—see [11] for a discussion of this topic. However, there exist schemes of cutting a polytope along lines other than its edges that produce a non-overlapping unfolding—see [1, 6, 8, 36]. It is plausible to conjecture that in the special case of surface cells of S, the unfolding of such a cell does not overlap itself, since S is induced by intersecting ∂P with $S_{3\mathrm{D}}$ (which is contained in an arrangement of three sets of parallel planes); however, related results [5, 30] do not suffice in our case, and we have not succeeded to prove this conjecture, which we leave for further research.

A *block sequence* $\mathcal{B} = (B_1, B_2, \ldots, B_k)$ is a sequence of building blocks of a surface cell c, so that for every pair of consecutive blocks $B_i, B_{i+1} \in \mathcal{B}$, we have $B_i \neq B_{i+1}$, and their boundaries share a common contact interval. We define $\mathcal{E}_\mathcal{B}$, the *edge sequence associated with* $\mathcal{B}$, to be the concatenation $\mathcal{E}_1||(\chi_1)||\mathcal{E}_2||(\chi_2)||\cdots||(\chi_{k-1})||\mathcal{E}_k$, where, for each i, χ_i is the polytope edge containing the contact interval that connects B_i with B_{i+1}, and $\mathcal{E}_i$ is the edge sequence associated with B_i that can be extended into $(\chi_{i-1})||\mathcal{E}_i||(\chi_i)$ (recall that there may be two oppositely oriented edge sequences associated with each B_i). Note that, given a sequence $\mathcal{B}$ of at least two blocks, $\mathcal{E}_\mathcal{B}$ is unique.

For each block tree $T_B(e)$ in $\mathcal{T}(e)$, each path in $T_B(e)$ defines a block sequence consisting of the blocks stored at its nodes. Conversely, every block sequence of c that consists of *distinct* blocks, with the possible exception of coincidence between its first and last blocks (where this block is of type II or III), appears as the sequence of blocks stored along some path of some block tree in $\mathcal{T}(e)$. We extend these important properties further in the following lemmas.

Lemma 3.13 *Let e, c and B be as above; then $T_B(e)$ has at most $O(1)$ nodes.*

Proof The construction of $T_B(e)$ is completed, when no path in $T_B(e)$ can be extended without violating conditions (1–3). In particular, each path of $T_B(e)$ consists of distinct blocks (except possibly for its leaf). Each building block of c contains at most $O(1)$ contact intervals and $O(1)$ transparent edge segments in its boundary, hence the degree of every node in $T_B(e)$ is $O(1)$. There are $O(1)$ building blocks of c, by Lemma 3.5, and this completes the proof of the lemma. □

Note that Lemma 3.13 implies that each building block is stored in at most $O(1)$ nodes of $T_B(e)$.

Lemma 3.14 *Let e, c and B be as above. Then each building block of c is stored in at least one node of $T_B(e)$.*

Proof Easy, and omitted. □

The following two lemmas summarize the discussion and justify the use of block trees. (Lemma 3.15 establishes rigorously the informal argument given right after the block tree definition.)

Lemma 3.15 *Let B be a building block of a surface cell c, and let $\mathcal{E}$ be an edge sequence associated with B. Let p, q be two points in c, so that there exists a shortest path $\pi(p, q)$ that is contained in c and crosses ∂B in at least two different points. Then $U_{\mathcal{E}}(\pi(p, q) \cap B)$ consists of either one or two disjoint straight segments, and the latter case is only possible if p, q lie in B.*

Proof Since $\pi(p, q)$ is a shortest path, every connected portion of $U_{\mathcal{E}}(\pi(p, q) \cap B)$ is a straight segment.

Suppose first that $p, q \in B$, and assume to the contrary that $U_{\mathcal{E}}(\pi(p, q) \cap B)$ consists of three or more distinct segments (the assumption in the lemma excludes the case of a single segment). Then at least one of these segments is bounded by two points $x, y \in \partial B$ and is incident to neither p nor q. Neither x nor y is incident to a transparent edge, since $\pi(p, q) \subset c$. Hence x, y are incident to two different respective contact intervals I_x, I_y on ∂B. The segment of $U_{\mathcal{E}}(\pi(p, q) \cap B)$ that is incident to p is also delimited by a point of intersection with a contact interval, by similar arguments. Denote this contact interval by I_p, and define I_q similarly. Obviously, the contact intervals I_x, I_y, I_p, I_q are all distinct. Since only building blocks of type I might have four contact intervals on their boundary (by Lemma 3.7), B must be of type I. But then B is contained in a single facet f, and $\pi(p, q)$ must be a straight segment contained in f, and thus cannot cross ∂f at all.

Suppose next that at least one of the points p, q, say p, is outside B. Assume that $U_{\mathcal{E}}(\pi(p, q) \cap B)$ consists of two or more distinct segments. Then at least one of these segments is bounded by two points x, y of ∂B (and is not incident to p). By the same arguments as above, x and y are incident to two different respective contact intervals I_x and I_y. The other segment of $U_{\mathcal{E}}(\pi(p, q) \cap B)$ is delimited by at least one point of intersection with some contact interval I_z, by similar arguments. Obviously,

the three contact intervals I_x, I_y, I_z are all distinct. In this case, B is either of type I or of type IV. In the former case, arguing as above, $\pi(p,q) \cap B$ is a single straight segment. In the latter case, B may have three contact intervals, but no straight line can meet all of them. Once again we reach a contradiction, which completes the proof of the lemma. □

Lemma 3.16 *Let e be a transparent edge bounding a surface cell c, and let B be a building block of c so that e appears on its boundary. Then, for each pair of points p, q, so that $p \in e \cap \partial B$ and $q \in c$, if the shortest path $\pi(p,q)$ is contained in c, then $\pi(p,q)$ is contained in the union of building blocks that form a single path in $T_B(e)$ (which starts from the root).*

Proof Let $p \in e \cap \partial B$ and $q \in c$ be two points as above, and denote by B' the building block that contains q. Denote by $\mathcal{B}$ the building block sequence crossed by $\pi(p,q)$. No building block appears in $\mathcal{B}$ more than once, except possibly B if $B = B'$ (by Lemma 3.15). Hence, the elements of B form a path in $T_B(e)$ from the root node (which stores B) to a node that stores B', as asserted. □

Corollary 3.17 *Let e be a transparent edge bounding a surface cell c, and let q be a point in c, such that the shortest path $\pi(s,q)$ intersects e, and the portion $\tilde{\pi}(s,q)$ of $\pi(s,q)$ between e and q is contained in c. Then $\tilde{\pi}(s,q)$ is contained in the union of building blocks that define a single path in some tree of $\mathcal{T}(e)$.*

Proof Follows from Lemma 3.16. □

Lemma 3.18 (*a*) *Let e be a transparent edge*; *then there are only $O(1)$ different paths from a root to a leaf in all trees in $\mathcal{T}(e)$.* (*b*) *It takes $O(n \log n)$ total time to construct the Riemann structures $\mathcal{T}(e)$ of all transparent edges e.*

Proof Let $T_B(e)$ be a block tree in $\mathcal{T}(e)$. There are $O(1)$ different paths from the root node to a leaf of $T_B(e)$ (see the proof of Lemma 3.13). There are two surface cells that bound e, and there are $O(1)$ building blocks of each surface cell, by Lemma 3.5. By Lemma 3.12, we can compute all the boundaries of all the building blocks in overall $O(n \log n)$ time. Hence the claim follows. □

For the surface cell c that contains s, we similarly define the set of block trees $\mathcal{T}(s)$, so that the root B of each block tree $T_B(s) \in \mathcal{T}(s)$ contains s on its boundary (recall that s is also regarded as a vertex of P). It is easy to see that Corollary 3.17 applies also to the Riemann structure $\mathcal{T}(s)$, in the sense that if q is a point in c, such that the shortest path $\pi(s,q)$ is contained in c, then $\pi(s,q)$ is contained in the union of building blocks that define a single path in some tree of $\mathcal{T}(s)$. It is also easy to see that Lemma 3.18 applies to $\mathcal{T}(s)$ as well.

3.3 Homotopy Classes

In this subsection we introduce certain topological constructs that will be used in the analysis of the shortest path algorithm in Sects. 4 and 5.

Let R be a region of ∂P. We say that R is *punctured* if either R is not simply connected, so its boundary consists of more than one cycle, or R contains a vertex of P in its interior; in the latter case, we remove any such vertex from R, and regard it as a new artificial singleton hole of R. We call these vertices of P and/or the holes of R *the islands* of R. Let X, Y be two disjoint connected sets of points in such a punctured region R, let $x_1, x_2 \in X$ and $y_1, y_2 \in Y$, and let $\pi(x_1, y_1), \pi(x_2, y_2)$ be two geodesic paths that connect x_1 to y_1 and x_2 to y_2, respectively, inside R. We say that $\pi(x_1, y_1)$ and $\pi(x_2, y_2)$ are *homotopic in R with respect to X and Y*, if one path can be continuously deformed into the other within R, while their corresponding endpoints remain in X and Y, respectively. (In particular, none of the deformed paths pass through a vertex of P.) When R is punctured, the geodesic paths that connect, within R, points in X to points in Y, may fall into several different *homotopy classes*, depending on the way in which these paths navigate around the islands of R. If R is not punctured, all the geodesic paths that connect, within R, points in X to points in Y, fall into a single homotopy class. In the analysis of the algorithm in Sects. 4 and 5, we only encounter homotopy classes of *simple geodesic subpaths* from one transparent edge e to another transparent edge f, inside a region R that is either a well-covering region of one of these edges or a single surface cell that contains both edges on its boundary. (We call these paths *subpaths*, since the full paths to f start from s.)

Since the algorithm only considers *shortest* paths, we can make the following useful observation. Consider the latter case (where the region R is a single surface cell c), and let $\mathcal{B}$ be a path in some block tree $T_B(e)$ within c that connects e to f. Then all the shortest paths that reach f from e via the building blocks in $\mathcal{B}$ belong to the same homotopy class. Similarly, in the former case (where R is a well-covering region consisting of $O(1)$ surface cells), all the shortest paths that connect e to f via a fixed sequence of building blocks, which itself is necessarily the concatenation of $O(1)$ sequences along paths in separate block trees (joined at points where the paths cross transparent edges between cells), belong to the same homotopy class.

4 The Shortest Path Algorithm

This section describes the wavefront propagation phase of the shortest path algorithm. Since this is the core of the algorithm, we present it here in detail, although its high-level description is very similar to the algorithm of [18]. Most of the problem-specific implementation details of the algorithm (which are quite different from those in [18]), as well as the final phase of the preprocessing for shortest path queries, are presented in Sect. 5.

The algorithm simulates a unit-speed *(true) wavefront* W expanding from s, and spreading along the surface of P. At *simulation time* t, W consists of points whose shortest path distance to s along ∂P is t. The true wavefront is a set of closed cycles; each cycle is a sequence of (folded) circular arcs (of equal radii), called *waves*. Each wave w_i of W at time t (denoted also as $w_i(t)$) is the locus of endpoints of a collection $\Pi_i(t)$ of shortest paths of length t from s that satisfy the following condition: There is a fixed polytope edge sequence $\mathcal{E}_i$ crossed by some path $\pi \in \Pi_i(t)$, so that the polytope edge sequence crossed by any other $\pi' \in \Pi_i(t)$ is a prefix of $\mathcal{E}_i$. The wave

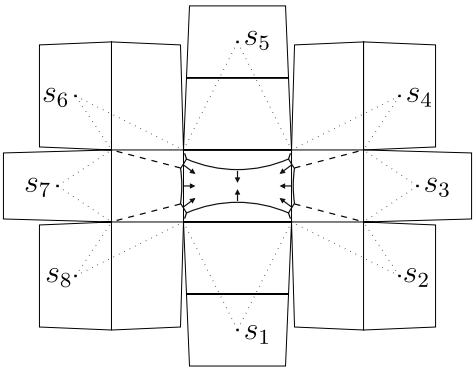


Fig. 21 The true wavefront W at some fixed time t, generated by eight source images $s_1, \ldots, s_8$. The surface of the box P (see the 3D illustration in Fig. 22) is unfolded in this illustration onto the plane of the last facet that W reaches; note that some facets of P are unfolded in more than one way (in particular, the facet that contains s is unfolded into eight distinct locations). The *dashed lines* are the bisectors between the current waves of W, and the *dotted lines* are the shortest paths to the vertices of P that are already reached by W

w_i is centered, in the destination plane of $U_{\mathcal{E}_i}$, at the source image $s_i = U_{\mathcal{E}_i}(s)$, called the *generator* of w_i. When w_i reaches, at some time t during the simulation, a point $p \in \partial P$, so that no other wave has reached p prior to time t, we say that s_i *claims* p, and put $claimer(p) := s_i$. We say that $\mathcal{E}_i$ is the *maximal polytope edge sequence of* s_i *at time* t. For each $p \in w_i(t)$ there exists a unique shortest path $\pi(s, p) \in \Pi_i(t)$ that intersects all the edges in the corresponding prefix of $\mathcal{E}_i$, and we denote it as $\pi(s_i, p)$. See Fig. 21.

The wave w_i has at most two neighbors w_{i-1}, w_{i+1} in W, each of which shares a single common point with w_i (if $w_{i-1} = w_{i+1}$, it shares two common points with w_i). As t increases and W expands accordingly (as well as the edge sequences $\mathcal{E}_i$ of its waves), each of the meeting points of w_i with its adjacent waves traces a *bisector*, which is the locus of points equidistant from the generators of the two corresponding waves; see Fig. 22. The bisector of the two consecutive generators s_i, s_{i+1} in W is denoted by $b(s_i, s_{i+1})$, and its unfolded image is a straight line.

During the simulation, the combinatorial structure of W changes at certain *critical events*, which may also change the topology of W. There are two kinds of critical events:
(i) *Vertex event*, where W reaches either a vertex of P or some other boundary vertex (an endpoint of a transparent edge) of the Riemann structure through which W is propagated. As will be described in Sect. 5, the wave in W that reaches a vertex event splits into two new waves after the event—see Fig. 23. These are the only events when a new wave is added to W. Our algorithm detects and processes all vertex events.[7]
(ii) *Bisector event*, when an existing wave is eliminated by other waves—the bisectors of all the involved generators meet at the event point. Our algorithm detects and

[7] A split at a vertex of P is a "real" split, because the two new waves continue past v along two different edge sequences. A split at a transparent endpoint is an artificial split, used to facilitate the propagation procedure; see Sect. 5 for details.

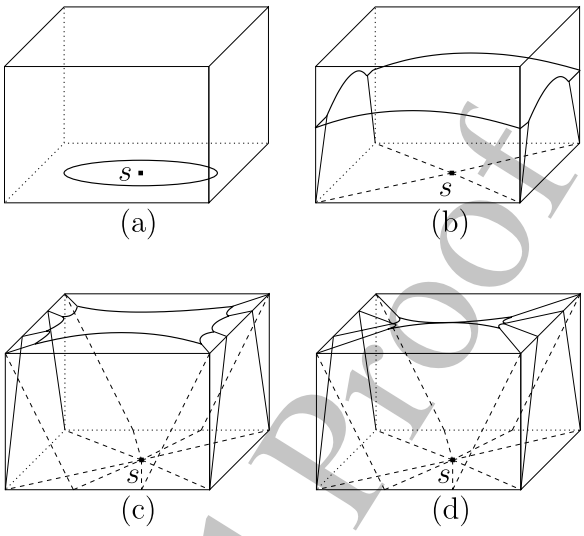


Fig. 22 W at different times t: (**a**) Before any critical event, it consists of a single wave. (**b**), (**c**) After the first four (resp., eight) vertex events W consists of four (resp., eight) (folded) waves. (**d**) After two additional critical events, which are bisector events, two waves are eliminated. Before the rest of the waves are eliminated, and immediately after (**d**), W disconnects into two distinct cycles

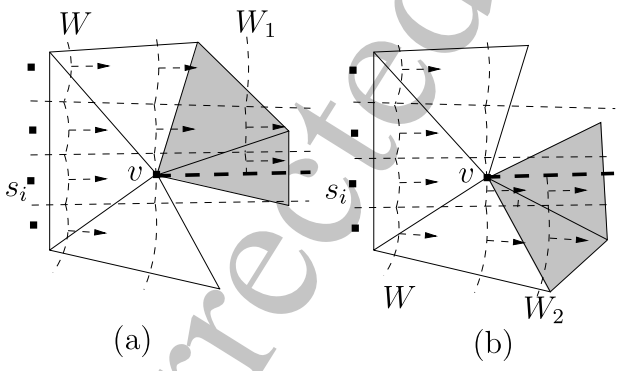


Fig. 23 Splitting the wavefront W at v (the triangles incident to v are unfoldings of its adjacent facets; note that the sum of all the facet angles at v is less than 2π). The *thick dashed line* coincides with the ray from s_i through v; it replaces the true bisector between the two new wavefronts W_1, W_2, which will later be calculated by the merging process. Each of W_1, W_2 is propagated separately after the event at v (through a different unfolding of the facet sequence around v—see, e.g., the shaded facets, each of which has a different image in (**a**) and (**b**))

processes only some of the bisector events, while others are not explicitly detected (recall that we only compute an implicit representation of SPM(s)). See Sect. 4.3 for further details.

4.1 The Propagation Algorithm

One-Sided Wavefronts The wavefront propagates between transparent edges across the cells of the conforming surface subdivision S. Propagating the exact wavefront explicitly appears to be inefficient (for reasons explained below), so at each transparent edge e we content ourselves with computing two *one-sided wavefronts*, passing

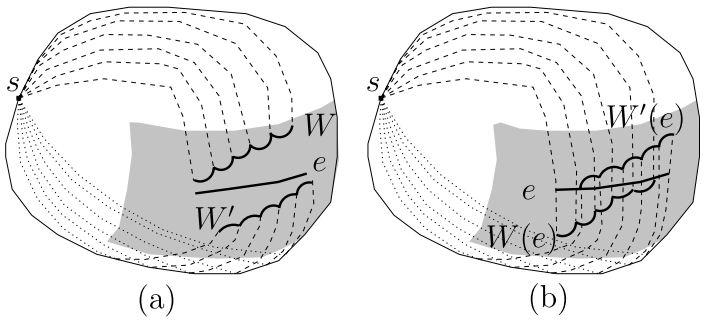


Fig. 24 (**a**) Two wavefronts W, W' are approaching e from two opposite directions, within $R(e)$ (*shaded*). (**b**) Two one-sided wavefronts $W(e)$, $W'(e)$, computed at the simulation time when e is completely covered by W, W', are propagated further within $R(e)$. However, some of the waves in $W(e)$, $W'(e)$ obviously do not belong to the true wavefront, since there is another wave in the opposite one-sided wavefront that claims the same points of e (before they do)

through e in opposite directions; together, these one-sided wavefronts carry all the information needed to compute the exact wavefront at e (but they also carry some superfluous information). Each spurious wave is the locus of endpoints of *geodesic* paths that traverse the same maximal edge sequence, but they need not be shortest paths. Still, our description of bisectors, maximal polytope edge sequences, and critical events that were defined for the true wavefront, also applies to the wavefront propagated by our algorithm.

In more detail, a one-sided wavefront $W(e)$ associated with a transparent edge e (and a specific side of e, which we ignore in this notation), is a sequence of waves $(w_1, \ldots, w_k)$ generated by the respective source images $s_1, \ldots, s_k$ (all unfolded to a common plane that is the same plane in which we compute the unfolded image of e), so that: (1) There exists a pairwise openly disjoint decomposition of e into k nonempty intervals $e_1, \ldots, e_k$, appearing in this order along e, and (2) For each $i = 1, \ldots, k$, for any point $p \in e_i$, the source image that claims p, among the generators of waves that reach p from the fixed side of e, is s_i. The algorithm maintains the following crucial *true distance* invariant (see Fig. 24 for an illustration):

(TD) For any transparent edge e and any point $p \in e$, the true distance $d_S(s, p)$ is the minimum of the two distances to p from the two source images that claim it in the two respective one-sided wavefronts for the opposite sides of e.

Remark For a fixed side of e, the corresponding one-sided wavefront $W(e)$ (implicitly) records the times at which the wavefront reaches the points of e from that side; note that $W(e)$ does not represent a fixed time t—each point on e is reached by the corresponding wave at a different time.

The Propagation Step The core of the algorithm is a method for computing a one-sided wavefront at an edge e based on the one-sided wavefronts of nearby edges. The set of these edges, denoted $input(e)$, is the set of transparent edges that bound $R(e)$, the well-covering region of e (cf. Sect. 2.3). To compute a one-sided wavefront at e, we propagate the one-sided wavefronts from each $f \in input(e)$ that has already been

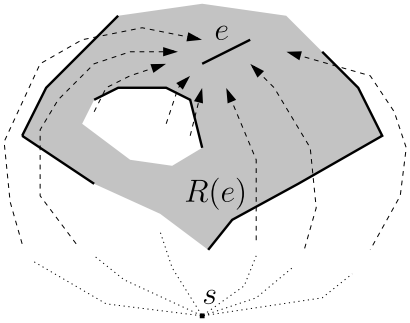


Fig. 25 The boundary of $R(e)$ (*shaded*) consists of two separate cycles. The transparent edge e and all the edges f in $input(e)$ that have been covered by the wavefront before time $covertime(e)$ are drawn as thick lines. The wavefronts $W(f, e)$ that contribute to the one-sided wavefronts at e have been propagated to e before time $covertime(e)$; wavefronts from other edges of $input(e)$ do not reach e either because of visibility constraints or because they are not ascertained to be completely covered at time $covertime(e)$ (in either case they do not include shortest paths from s to any point on e)

processed by the algorithm, to e inside $R(e)$, and then merge the results, separately on each side of e, to get the two one-sided wavefronts that reach e from each of its sides. See Fig. 25 for an illustration. The algorithm propagates the wavefronts inside $O(1)$ unfolded images of (portions of) $R(e)$, using the Riemann structure defined in Sect. 3.2. The wavefronts are propagated only to points that can be connected to the appropriate generator by straight lines inside the appropriate unfolded portion of $R(e)$ (these points are "visible" from the generator); that is, the shortest paths within this unfolded image, traversed by the wavefront as it expands from the unfolded image of $f \in input(e)$ to the image of e, must not bend (cf. Sect. 2.1 and Sect. 3). Because the image of the appropriate portion of $R(e)$ is not necessarily convex, its reflex corners may block portions of wavefronts from some edges of $input(e)$ from reaching e. The paths corresponding to blocked portions of wavefronts that exit $R(e)$ may then re-enter it through other edges of $input(e)$. For any point $p \in e$, the shortest path from s to p passes through some $f \in input(e)$ (unless $s \in R(e)$), so constraining the source wavefronts to reach e directly from an edge in $input(e)$, without leaving $R(e)$, does not lose any essential information.

We denote by $output(e)$ the set of direct "successor" edges to which the one-sided wavefronts of e should be propagated; specifically, $output(e) = \{f \mid e \in input(f)\}$.

Lemma 4.1 *For any transparent edge e, $output(e)$ consists of a constant number of edges.*

Proof Since $|R(f)| = O(1)$ for all f, and each $R(f)$ is a connected set of cells of S, no edge e can belong to $input(f)$ for more than $O(1)$ edges f (there are only $O(1)$ possible connected sets of $O(1)$ cells that contain e on the boundary of their union), and $|input(f)| = O(1)$, by construction. □

Remark As a wavefront is propagated from an edge $f \in input(e)$ to e, it may cross other intermediate transparent edges g (see Fig. 26). Such an edge g will be processed at an interleaving step, when wavefronts from edges $h \in input(g)$ are propagated to g

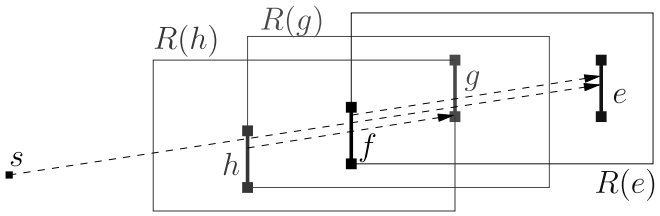


Fig. 26 Interleaving of the well-covering regions. The wavefront propagation from $h \subset \partial R(g)$ to g passes through f, and the propagation from $f \subset \partial R(e)$ to e passes through g

(and some of the propagated waves may reach g by crossing f first). This "leap-frog" behavior of the algorithm causes some overlap between propagations, but it affects neither the correctness nor the asymptotic efficiency of the algorithm.

The Simulation Clock The simulation of the wavefront propagation is loosely synchronized with the real "propagation clock" (which measures the distance from s). The main purpose of the synchronization is to ensure that the only waves that are propagated from a transparent edge e to edges in $output(e)$ are those that have reached e no later than $|e|$ simulation time units after e has been completely covered. This, and the well-covering property of e (which guarantees that at this time none of these waves has yet reached any $f \in output(e)$), allow us to propagate further all the shortest paths that cross e by "processing" e only once, thereby making the algorithm adhere to the continuous Dijkstra paradigm, and consequently be efficient.

For a transparent edge e, we define the *control distance from s to e*, denoted by $\tilde{d}_S(s,e)$, as follows. If $s \in R(e)$, and e contains at least one point p that is visible from s within at least one unfolded image $U(R(e))$, for some unfolding U, then e is called *directly reachable* (from s), and $\tilde{d}_S(s,e)$ is defined to be the distance from $U(s)$ to $U(p)$ within $U(R(e))$. The point $p \in e$ can be chosen freely, unless $U(s)$ and $U(e)$ are collinear within $U(R(e))$—then p must be taken as the endpoint of e whose unfolded image is closer to $U(s)$. Otherwise ($s \notin R(e)$ or e is completely hidden from s in every unfolded image of $R(e)$), we define $\tilde{d}_S(s,e) = \min\{d_S(s,a), d_S(s,b)\}$, where a, b are the endpoints of e, and $d_S(s,a), d_S(s,b)$ refer to their *exact* values. Thus, $\tilde{d}_S(s,e)$ is a rough estimate of the real distance $d_S(s,e)$, since $d_S(s,e) \le \tilde{d}_S(s,e) < d_S(s,e) + |e|$. The distances $d_S(s,a), d_S(s,b)$ are computed exactly by the algorithm, by computing the distances to a, b within each of the one-sided wavefronts from s to e, and by using the invariant (TD). We compute both one-sided wavefronts for e at the first time we can ascertain that e has been completely covered by wavefronts from either the edges in $input(e)$, or directly from s if e is directly reachable. This time is $\tilde{d}_S(s,e) + |e|$, a conservative yet "safe" upper bound of the real time $\max\{d_S(s,q) \mid q \in e\}$ at which e is completely run over by the true (not one-sided) wavefront.

The continuous Dijkstra propagation mechanism computes $\tilde{d}_S(s,e) + |e|$ on the fly for each edge e, using a variable $covertime(e)$. Initially, for every directly reachable e, we calculate $\tilde{d}_S(s,e)$, by propagating the wavefront from s within the surface cell which contains s, as described in Sect. 5, and put $covertime(e) := \tilde{d}_S(s,e) + |e|$. For all other edges e, we initialize $covertime(e) := +\infty$.

The simulation maintains a time parameter t, called the *simulation clock*, which the algorithm strictly increases in discrete steps during execution, and processes each edge e when t reaches the value $covertime(e)$. A high-level description of the simulation is as follows:

PROPAGATION ALGORITHM

Initialize $covertime(e)$, for all transparent edges e, as described above. Store with each directly reachable e the wavefronts that are propagated to e from s (without crossing edges in $input(e)$).

while there are still *unprocessed* transparent edges **do**

1. Select the unprocessed edge e with minimum $covertime(e)$, and set $t := covertime(e)$.
2. **Merge:** Compute the one-sided wavefronts for both sides of e, by *merging* together, separately on each side of e, the wavefronts that reach e from that side, either from all the already processed edges $f \in input(e)$ (these wavefronts are propagated to e in Step 3 below), or directly from s (those wavefronts are stored at e in the initialization step). Compute $d_S(s, v)$ exactly for each endpoint v of e (the minimum of at most two distances to v provided by the two one-sided wavefronts at e).
3. **Propagate:** For each edge $g \in output(e)$, compute the time $t_{e,g}$ at which one of the one-sided wavefronts from e first reaches an endpoint of g, by *propagating* the relevant one-sided wavefront from e to g. Set $covertime(g) := \min\{covertime(g), t_{e,g} + |g|\}$. Store with g the resulting wavefront propagated from e, to prepare for the later merging step at g.

endwhile

The following lemma establishes the correctness of the algorithm. That is, it shows that *covertime*() is correctly maintained and that the edges required for processing e have already been processed by the time e is processed. The description of Step 2 appears in Sect. 4.2 as the wavefront *merging* procedure; the computation of $t_{e,g}$ in Step 3 is a byproduct of the propagation algorithm as described below and detailed in Sect. 5. For the proof of the lemma we assume, for now, that the invariant (TD) is correctly maintained—this crucial invariant will be proved later in Lemma 4.5.

Lemma 4.2 *During the propagation, the following invariants hold for each transparent edge e:*

(a) *The final value of $covertime(e)$ (the time when e is processed) is $\tilde{d}_S(s, e) + |e|$; for directly reachable edges, it is at most $\tilde{d}_S(s, e) + |e|$. The variable $covertime(e)$ is set to this value by the algorithm before or at the time when the simulation clock t reaches this value.*

(b) *The value of $covertime(e)$ is updated only a constant number of times before it is set to $\tilde{d}_S(s, e) + |e|$.*

(c) *If there exists a path π from s that belongs to a one-sided wavefront at e, so that a prefix of π belongs to a one-sided wavefront at an edge $f \in input(e)$, then* $\tilde{d}_S(s, f) + |f| < \tilde{d}_S(s, e) + |e|$.

Proof (a) For directly reachable edges, this holds by definition of the control distance; for other edges e, we prove by induction on the (discrete steps of the) simulation clock, as follows. The shortest path π' to one of the endpoints of e (which reaches e at the time $|\pi'| = t_e = \tilde{d}_S(s, e)$) crosses some $f \in input(e)$ at an earlier time t_f, where $d_S(s, f) \le t_f < \tilde{d}_S(s, f) + |f|$; we may assume that f is the last such edge of $input(e)$. Note that we must have $t_e \ge t_f + d_S(e, f)$. By (W3$_S$), $d_S(e, f) \ge 2|f|$, and so $t_e \ge d_S(s, f) + 2|f|$. Since $\tilde{d}_S(s, f) < d_S(s, f) + |f|$, we have

$$|\pi'| = t_e \ge d_S(s, f) + 2|f| > \tilde{d}_S(s, f) + |f|. \tag{1}$$

By induction and by this inequality, f has already been processed before the simulation clock reaches t_e, and so $covertime(e)$ is set, in Step 3, to $t_{f,e} + |e| = t_e + |e| = \tilde{d}_S(s, e) + |e|$ (unless it has already been set to this value earlier), at time no later than $t_e = \tilde{d}_S(s, e)$ (and therefore no later than $\tilde{d}_S(s, e) + |e|$, as claimed). By (TD), the variable $covertime(e)$ cannot be set later (or earlier) to any *smaller* value; it follows that e is processed at simulation time $\tilde{d}_S(s, e) + |e|$.

(b) The value of $covertime(e)$ is updated only when we process an edge f such that $e \in output(f)$ (i.e., $f \in input(e)$), which consists of $O(1)$ edges, by construction.

(c) Any path π that is part of a one-sided wavefront at e must satisfy $d_S(s, e) \le |\pi| < \tilde{d}_S(s, e) + |e|$ (π cannot reach e earlier by definition, and if π reaches e later, then, by (a), e would have been already processed and π would not have contributed to any of the one-sided wavefronts at e). Since π passes through a transparent edge $f \in input(e)$, we can show that $|\pi| > \tilde{d}_S(s, f) + |f|$, by applying arguments similar to those used to derive (1) in (a). Hence we can conclude that $\tilde{d}_S(s, f) + |f| < \tilde{d}_S(s, e) + |e|$. □

Remark The synchronization mechanism above assures that if a wave w reaches a transparent edge e later than the time at which e has been ascertained to be completely covered by the wavefront, then w will not contribute to either of the two one-sided wavefronts at e. In fact, this important property yields an implicit interaction between all the wavefronts that reach e, allowing a wave to be propagated further only if it is not too "late"; that is, only if it reaches points on e no later than $2|e|$ simulation time units after a wave from another wavefront.[8]

Topologically Constrained Wavefronts Let f, e be two transparent edges so that $f \in input(e)$, and let H be a homotopy class of simple geodesic paths connecting f to e within $R(e)$ (recall that there might be multiple homotopy classes of that kind; see Sect. 3.3). We denote by $W_H(f, e)$ the unique maximal (contiguous) portion of the one-sided wavefront $W(f)$ that reaches e by traversing only the subpaths from f

[8]For a detailed discussion of why we use the bound $2|e|$ rather than just $|e|$ see the description of the simulation time maintenance in Sect. 5.3.1.

to e that belong to H. In Sect. 5 we regard $W_H(f,e)$ as a "kinetic" structure, consisting of a continuum of "snapshots," each recording the wavefront at some time t. In contrast, in the current section we only consider the (static) resulting wavefront that reaches e, where each point q on (an appropriate portion of) e is claimed by some wave of $W_H(f,e)$, at some time t_q. (Note that this static version is *not* a snapshot at a fixed time of the kinetic version.) We say that $W_H(f,e)$ is a *topologically constrained wavefront* (by H). To simplify notation, we omit H whenever possible, and simply denote the wavefront, somewhat ambiguously, as $W(f,e)$.

A topologically constrained wavefront $W_H(f,e)$ is bounded by a pair of extreme bisectors of an "artificial" nature, defined in one of the two following ways. We say that a vertex of P in $R(e)$ or a transparent endpoint $x \in \partial R(e)$ is a *constraint of* H if x lies on the boundary of R_H, which is the locus of all points traversed by all (geodesic) paths in H (see Fig. 27). It is easy to see that R_H is bounded by e, f, and by a pair of "chains," each of which connects f with e, and the unfolded image of which (along the polytope edge sequence corresponding to H) is a concave polygonal path that bends only at the constraints of H (this structure is sometimes called an *hourglass*; see [14] for a similar analysis).

Let s' be an extreme generator in $W_H(f,e)$, and let π be a simple geodesic path (in H) from s' that reaches f and touches ∂R_H; see the path π_1 in Fig. 27. It is easy to see that if such a path π exists, then it must be an extreme path among all paths encoded in $W_H(f,e)$, since any other path in $W_H(f,e)$ cannot intersect π (see Lemma 4.3 below); we therefore regard π as an extreme artificial bisector of $W_H(f,e)$. Another kind of an extreme artificial bisector arises when, during the propagation of (the kinetic version of) $W_H(f,e)$, an extreme generator s' is eliminated in a bisector event x, as described below, and the neighbor s'' of s' becomes extreme; then the path π from s'' through the location of x becomes extreme in $W_H(f,e)$—see the path π_2 in Fig. 27 for an example.[9]

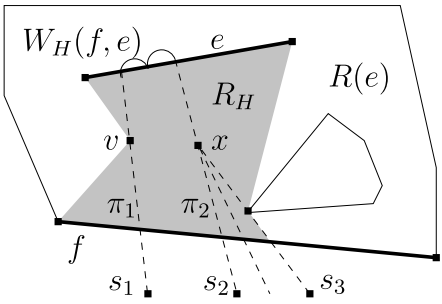


Fig. 27 The "hourglass" region R_H that is traversed by all paths in H is *shaded*. The extreme artificial bisectors of the topologically constrained wavefront $W_H(f,e)$ are the paths π_1 (from the extreme generator s_1 through the vertex v of P, which is one of the constraints of H) and π_2 (from the generator s_2, which became extreme when its neighbor s_3 was eliminated at a bisector event x, through the location of x)

[9]Even though π is geodesic, it is not a shortest path to any point beyond x; it is only a convenient (though conservative) way of bounding $W_H(f,e)$ without losing any essential information.

4.2 Merging Wavefronts

Consider the computation of the one-sided wavefront $W(e)$ at a transparent edge e that will be propagated further (through e) to, say, the left of e. The *contributing wavefronts* to this computation are all wavefronts $W(f, e)$, for $f \in input(e)$, that contain waves that reach e from the right (not later than at time $covertime(e)$). If e is directly reachable from s, and a wavefront $W(s, e)$ has been propagated from s to the right side of e, then $W(s, e)$ is also contributing to the computation of $W(e)$. The contributing wavefronts for the computation of the opposite one-sided wavefront at e are defined symmetrically.

To simplify notation, in the rest of the paper we assume each transparent edge e to be oriented, in an arbitrary direction (unless otherwise specified). For the special case $s \in R(e)$, we also treat the direct wavefront $W(s, e)$ from s to e as if s were another transparent edge f in $input(e)$.

We call the set of all points of e claimed by a contributing wavefront $W(f, e)$ the *claimed portion* or the *claim of* $W(f, e)$. The following lemma implies that this set is a (possibly empty) *connected* subinterval of e.

Lemma 4.3 *Let e be a transparent edge, and let $W(f, e)$ and $W(g, e)$ be two (topologically constrained) contributors to the one-sided wavefront $W(e)$ that reaches e from the right, say. Let x and x' be points on e claimed by $W(f, e)$, and let y be a point on e claimed by $W(g, e)$. Then y cannot lie between x and x'.*

Proof Suppose to the contrary that y does lie between x and x'. Consider a modified environment in which the paths that reach e from the left are "blocked" at e by a thin high obstacle, erected on ∂P at e. This modification does not influence the wavefronts $W(f, e)$ and $W(g, e)$, since no wave reaches e more than once. The simple geodesic paths $\pi(s, x), \pi(s, x')$, and $\pi(s, y)$ in the modified environment connect x and x' to f, and y to g, inside $R(e)$, and lie on the right side of e locally near x, x', and y; see Fig. 28(a). By (TD), the paths $\pi(s, x), \pi(s, x')$, and $\pi(s, y)$ are *shortest paths* from s to these points in the modified environment, and therefore do not cross each other. Since $W(f, e)$, $W(g, e)$ are topologically constrained by different homotopies (within $R(e)$), no path traversed by $W(g, e)$ can reach e and be fully contained in the portion Q of ∂P delimited by f, e, and by the portions of $\pi(s, x), \pi(s, x')$ between f

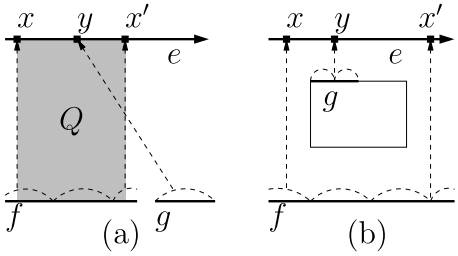


Fig. 28 (**a**) $W(g, e)$ cannot claim the point y, for otherwise the shortest path $\pi(s, y)$ (which crosses the transparent edge g) would have to cross one of the paths $\pi(s, x), \pi(s, x')$, which is impossible for shortest paths. The region Q delimited by f, e, and the portions of $\pi(s, x), \pi(s, x')$ between f and e is shaded. (**b**) If $W(f, e)$ is not topologically constrained, $W(g, e)$ may claim an in-between point y on e

and e. Therefore, the portion of the shortest path $\pi(s, y)$ between g and e must enter the region Q through one of the paths $\pi(s, x), \pi(s, x')$, which is a contradiction. □

Remark Lemma 4.3 may fail if $W(f, e)$ is *not* a topologically constrained wavefront; see Fig. 28(b) for an example. Moreover, if $W(g, e)$ reaches e from *the other side* of e then it is possible for $W(g, e)$ to claim portions of $\overline{xx'}$ without claiming x and x'. It is this fact that makes the explicit merging of the two one-sided wavefronts expensive.

We now proceed to describe the *merging process*, applied to the contributing wavefronts that reach a transparent edge e from a fixed side; the process results in the construction of the corresponding one-sided wavefront at e. Most of the low-level details of the process are embedded in the procedures supported by the data structure described in Sect. 5.1; for now, before proceeding with Lemma 4.4, we briefly review the basic operations, and assert their time complexity bounds. Each contributing wavefront W is maintained as a list of generators in a balanced tree data structure; we may therefore assume that each of the operations of constructing a single bisector, finding its intersection point with e, measuring the distance to a point on e from a single generator, and concatenating the lists representing two wavefront portions into a single list, takes $O(\log n)$ time. This will be further explained and verified in Sect. 5.

Lemma 4.4 *For each transparent edge e and for each $f \in input(e)$, we can compute the claim of each of the wavefront portions $W(f, e)$ that contribute to the one-sided wavefront $W(e)$ that reaches e from the right, say, in $O((1 + k)\log n)$ total time, where k is the total number of generators in all wavefronts $W(f, e)$ that are absent from $W(e)$.*

Proof For each contributing wavefront $W(f, e)$, we show how to determine its claim in the presence of only one other contributing wavefront $W(g, e)$. The (connected) intersection of these claimed portions, taken over all other $O(1)$ contributors $W(g, e)$, is the part of e claimed by $W(f, e)$ in $W(e)$. This results in the algorithm asserted in the lemma.

Orient e from one endpoint a to the other endpoint b. We refer to a (resp., b) as the *left* (resp., *right*) endpoint of e. We determine whether the claim of $W(f, e)$ is *to the left or to the right* of that of $W(g, e)$, as follows. If both $W(f, e)$ and $W(g, e)$ claim a, then, in $O(\log n)$ time, we check which of them reaches it earlier (we only need to check the distances from a to the first and the last generator in each of the two wavefronts, since we assume that $W(f, e)$, $W(g, e)$ only contain waves that reach e). Otherwise, one of $W(f, e)$, $W(g, e)$ reaches a point $p \in e$ (not necessarily a) that is left of any point reached by the other; by Lemma 4.3, the claim that contains p, by "winning" wavefront, is to the left of the claim of the other wavefront. To find p, we intersect the first and the last (artificial) bisectors of each of $W(f, e)$, $W(g, e)$ with e; p is the intersection closest to a.

A basic operation performed here and later in the merging process is to determine the order of two points x, y along e. Using the surface unfolding data structure of

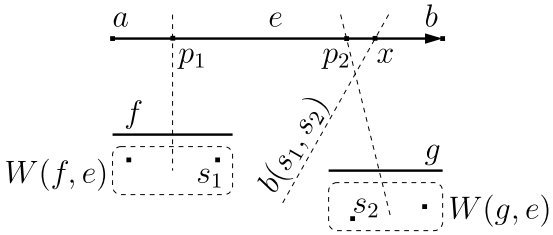


Fig. 29 The source image s_2 is eliminated from $W(e)$, because its contribution to $W(e)$ must be to the left of p_2 and to the right of x, and therefore does not exist along e

Sect. 2.4, we can compute the polytope edge sequence $\mathcal{E}_e$ crossed by e, in $O(\log n)$ time, and compare $U_{\mathcal{E}_e}(x)$ with $U_{\mathcal{E}_e}(y)$.

Without loss of generality, assume that the claim of $W(f,e)$ is left of that of $W(g,e)$. Note that in this definition we also allow for the case where $W(g,e)$ is completely annihilated by $W(f,e)$.

Let s_1 denote the generator in $W(f,e)$ that claims the rightmost point on e among all points claimed by $W(f,e)$; by assumption, s_1 is an extreme generator of $W(f,e)$. Let p_1 be the left endpoint of the claim of s_1 on $U_{\mathcal{E}_e}(e)$ (as determined by $W(f,e)$ alone; it is the intersection of $U_{\mathcal{E}_e}(e)$ and the left bisector of s_1). Similarly, let s_2 denote the generator in $W(g,e)$ claiming the leftmost point on e (among all points claimed by $W(g,e)$), and let p_2 be the right endpoint of the claim of s_2 on $U_{\mathcal{E}_e}(e)$ (as determined by $W(g,e)$ alone). We compute the (unfolded) bisector of s_1 and s_2, and find its intersection point x with $U_{\mathcal{E}_e}(e)$; see Fig. 29. If x is to the left of p_1 or x does not exist and the entire e is to the right of $b(s_1,s_2)$, then we delete s_1 from $W(f,e)$, reset s_1 to be the next generator in $W(f,e)$, and recompute p_1. If x is to the right of p_2 or x does not exist and the entire e is to the left of $b(s_1,s_2)$, then we update $W(g,e)$, s_2 and p_2 symmetrically. In either case, we recompute x and repeat this test. If p_1 is to the left of p_2 and x lies between them, then x is the right endpoint of the claim of $W(f,e)$ in the presence of $W(g,e)$ and the left endpoint of the claim of $W(g,e)$ in the presence of $W(f,e)$.

Consider next the time complexity of this process. Merging each of the $O(1)$ pairs $W(f,e)$, $W(g,e)$ of wavefronts involves $O(1+k)$ operations, where k is the number of generators that are deleted from the wavefronts during that merge, and where each operation either computes a single bisector, or finds its intersection point with e, or measures the distance to a point on e from a single generator, or deletes an extreme wave from a wavefront, or concatenates two wavefront portions into a single list. As stated above, each of these operations can be implemented in $O(\log n)$ time. Summing over all $O(1)$ pairs $W(f,e)$, $W(g,e)$, the bound follows. □

The following lemma proves the correctness of the process, with the assumption that the propagation procedure, whose details are not provided yet, is correct.

Lemma 4.5 (i) *Any generator deleted during the construction of a one-sided wavefront at the transparent edge e does not contribute to the true wavefront at e.* (ii) *Assuming that the propagation algorithm deletes a wave from the wavefront not earlier than the time when the wave becomes dominated by its neighbors, every generator that contributes to the true wavefront at e belongs to one of the* (*merged*) *one-sided wavefronts at e.*

Proof The first part is obvious—each point in the claim of each deleted generator s_i along e is reached earlier either by its neighbor generator in the same contributing wavefront or by a generator of a competing wavefront. It is possible that these generators are further dominated by other generators in the true wavefront, but in either case s_i cannot claim any portion of e in the true wavefront. The second part follows by induction on the order in which transparent edges are being processed, based on the following two facts: (i) Any wave that contributes to the true wavefront at e must arrive either directly from s inside $R(e)$, or through some edge $f \in input(e)$. (ii) The one-sided wavefronts at each edge $f \in input(e)$ that have been covered before e is processed, have already been computed (by Lemma 4.2). Hence each generator s_i that contributes to the true wavefront at e contributes to the true wavefront at some such edge f, and the induction hypothesis implies that s_i belongs to the appropriate one-sided wavefront at f. Since, by the assumption that is established in the next section, the propagation algorithm from f to e deletes from the wavefront only the waves that become dominated by other waves, s_i participates in the merging process at e, and, by the first part of the lemma, cannot be fully eliminated in that process. □

4.3 The Bisector Events

When we propagate a one-sided wavefront $W(e)$ to the edges of $output(e)$, as will be described in detail in Sect. 5.2, and when we merge the wavefronts that reach the same transparent edge, as described in Sect. 4.2, *bisector events* may occur, as defined above. We distinguish between the following two kinds of bisector events.

(i) *Bisector events of the first kind* are detected when we simulate the advance of the wavefront $W(e)$ from a transparent edge e to another edge g to compute the wavefront $W(e, g)$, where $g \in output(e)$. In any such event, two non-adjacent generators s_{i-1}, s_{i+1} become adjacent due to the elimination of the intermediate wave generated by s_i (as we show in Lemma 5.6, this is the only kind of events that occur when waves from the *same* topologically constrained wavefront collide with each other); see Fig. 30(a) for an illustration. This event is the starting point of $b(s_{i-1}, s_{i+1})$, which reaches g in $W(e, g)$ if both waves survive the trip.

A bisector event, at which the *first* generator s_1 in the propagated wavefront is eliminated, is treated somewhat differently; see Fig. 30(b), (c) for an illustration. In

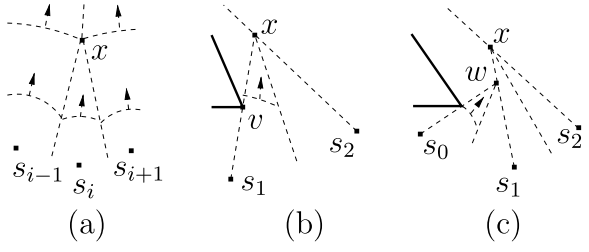


Fig. 30 When a bisector event (of the first kind) takes place at x: (**a**) The wave of s_i is eliminated, and the new bisector $b(s_{i-1}, s_{i+1})$ is computed. (**b**), (**c**) The wave of s_1 is eliminated, and the ray from s_2 through x becomes the leftmost (artificial) bisector of W, instead of the former leftmost bisector, which is the ray from s_1 through either (**b**) a transparent edge endpoint v (a visibility constraint), or (**c**) the location w of an earlier bisector event, where s_0, the previous leftmost generator of W, has been eliminated

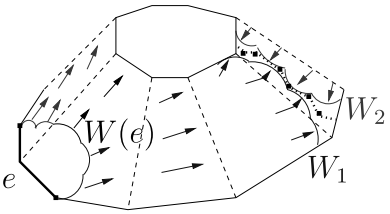


Fig. 31 $W(e)$, propagated from e, is split inside $R(e)$ when it reaches the inner (*top*) boundary cycle. Then the two new topologically constrained wavefronts partially collide into each other, creating a sequence of bisectors (*dotted lines*, bounded by *thick points* where bisector events of the second kind occur), eliminating a sequence of waves in each wavefront

this case s_1 is deleted from the wavefront W and the next generator s_2 becomes the first in W. The ray from s_2 through the event location becomes the first (that is, extreme), artificial bisector of W, meaning that W needs to be maintained only on the s_2-side of this bisector (which is a conservative bound). Indeed, any point $p \in \partial P$ for which the path $\pi(s_2, p)$ crosses $b(s_1, s_2)$ into the region of ∂P that is claimed by s_1 (among all generators in W), can be reached by a shorter path from s_1. The case when the *last* generator of W is eliminated is treated symmetrically.

(ii) *Bisector events of the second kind* occur when waves from *different* topologically constrained wavefronts collide with each other. Our algorithm does not explicitly detect these events; however, they are all (implicitly) considered at the query processing time, as described in Sect. 5.4, and some of them undergo additional (albeit still implicit) processing, as briefly described next.

If a generator s_i contributes to one of the input wavefronts $W(e, g)$ but not to the merged one-sided wavefront $W(g)$ at g, then s_i is involved in at least one bisector event (of the second kind) on the way from e to g, and there must exist some generator s_j in another (topologically constrained) wavefront $W(f, g)$ that also reaches g, which eliminates the wave of s_i. This event is implicitly recognized by the algorithm when s_i is deleted from $W(e, g)$ during the merging process at g.

Another kind of such an event occurs when a one-sided wavefront $W(e)$ is split during its propagation inside $R(e)$ (either at of a vertex of P or at a hole of $R(e)$ that may contain one or more vertices of P), and the two portions of the split wavefront partially collide into each other during their further propagation inside $R(e)$, as distinct topologically constrained wavefronts, before they reach $\partial R(e)$—see Fig. 31. The algorithm implicitly processes some of these events, by realizing that these waves attempt to exit the current block tree, by re-entering an already visited building block. The algorithm then simply discards these waves from further processing; see Sect. 5.3.1.

Tentatively False and True Bisector Events Consider the time $t = covertime(e)$. There may be waves that have reached e before time t (although not earlier than time $t - 2|e|$), and some of these waves could have participated in bisector events of the first kind "beyond" e that could have taken place before time t. As described in Sect. 5, the algorithm detects these (currently considered as) "false" bisector events when the wavefronts from the edges in $input(e)$ are propagated to e, but the generators that are eliminated in these events are not deleted from their corresponding

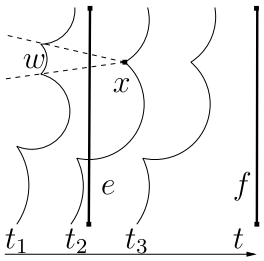


Fig. 32 The bisector event at x occurs at time t_2. It is first detected when the wavefront is propagated toward the transparent edge e, which has not been fully covered yet. Since x is beyond e, the event is currently considered false (and the eliminated wave w is not deleted from the wavefront, so that it shows up on $W(e)$). When e is ascertained (at time $t_3 = covertime(e)$) to be fully covered, the one-sided wavefront $W(e)$ is computed, and then propagated toward the transparent edge f, starting from some time $t < covertime(e)$ (e.g., t_2). Since w is part of $W(e)$, the bisector event at x is detected again, and this time it is considered to be true

contributing wavefronts before time t. This is done to ensure that the invariant (TD) is satisfied. However, such a bisector event is detected again, and considered to be true, when the wavefront is propagated further, after processing e. This latter propagation from e can be considered to start at the time when the first among such events occurs, which might happen earlier than $covertime(e)$; see Fig. 32. Further details are given in Sect. 5, where we also show that the number of all "true" and "false" processed events is only $O(n)$.

Remark Note that a detected "true" event does not necessarily appear as a vertex of SPM(s), since it involves only waves from a single one-sided wavefront, and its location x can actually be claimed by a wave from another wavefront. To find the true claimer of x (or any other query point), we make use of the fact that x belongs to only $O(1)$ well-covering regions, each of which is traversed by only $O(1)$ wavefronts; knowing the claimer of x in each of these wavefronts gives us the "global" claimer of x—see Sect. 5.4.

5 Implementation Details

5.1 The Data Structures

A one-sided wavefront is an ordered list of generators (source images). Our algorithm performs the following three types of operations on these lists (the first two types are similar to those in [18]):

1. *List operations*: CONCATENATE, SPLIT, and DELETE.[10] Each operation is applied to the list of generators that represents the wavefront at any particular simulation time.

[10]Note that the algorithm does not use INSERT operations; a new wave is created only during a SPLIT operation, and generating it is part of the SPLIT. Similarly, the omitted CREATE operation is performed only once, when the first singleton wavefront at s is created.

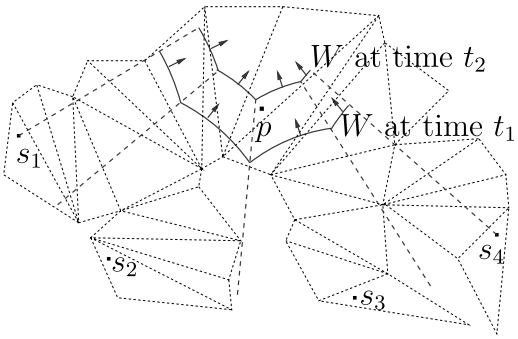


Fig. 33 The wavefront W at simulation times t_1 and t_2 consists of four source images $s_1, \ldots, s_4$, all unfolded to one plane at time t_1 and to another plane at time t_2 (for this illustration, both planes are the same—this is the plane of the facet that contains the point p). In order to determine the generator of W that claims p, the SEARCH operation can be applied to the version of W at time t_2, when p is already claimed by s_3

2. *Priority queue operations*: We assign to each generator a priority (as defined below in Sect. 5.3.1; it is essentially the time at which the generator is eliminated by its two neighbors), and the data structure needs to update priorities and find the minimum priority in the list.
3. *Source unfolding operations*: (a) To compute explicitly each source image s_i in the wavefront at time t, we need to unfold the maximal polytope edge sequence of s_i at t—this operation is referred to as an "unfolding query"; the unfolding structure needs to be updated as the wavefront advances. (b) The bisectors between consecutive generators in the list, as long as they do not meet one another, partition a portion of the plane of unfolding into a linearly ordered sequence of regions, and we want to locate the region containing a query point q. That is, we SEARCH in the generator list for a claimer of q (without considering other wavefronts or possible visibility constraints); see Fig. 33, and see later for more precise details.

All these types of operations can be supported by a data structure based on balanced binary search trees, with the generators stored at the leaves [15]. In particular, the "bare" list operations (ignoring the maintenance of priorities and unfolding data) take $O(\log n)$ time each, using standard machinery [15, 37]. Moreover, one can also update the extra unfolding fields (described in the following paragraphs) as these list operations are executed (so that the operations retain their $O(\log n)$ time). Although not completely straightforward, the manipulation of the unfolding fields is still simple enough, so that we omit it here—we present the full details in [34]. The priority queue operations are supported by adding a priority field to each node of the binary tree, which records the minimum priority of the leaves in the subtree of that node (and the leaf with that priority). Each priority queue operation takes $O(\log n)$ time; the actual implementation details are fairly standard, and are therefore omitted.

Source Unfolding Operations The source unfolding queries are supported by adding an unfolding transformation field $U[v]$ to each node v of the binary tree, in such a way that, for any queried generator s_i, the unfolding of s_i is equal to the

product (composition) of the transformations stored at the nodes of the path from the leaf storing s_i to the root. That is, if the nodes on the path are $v_1 = \text{root}, v_2, \ldots, v_k =$ leaf storing s_i, then the unfolding of s_i is given by $U[v_1]U[v_2]\cdots U[v_k]$. We represent each unfolding transformation as a single 4×4 matrix in homogeneous coordinates (see [32, 34]), so composition of any pair of transformations takes $O(1)$ time. For each node v, and for any path $v = v_1, v_2, \ldots, v_k$ that leads from v to a leaf, the product $U[v_1]U[v_2]\cdots U[v_k]$ maps the generator stored at v_k to a fixed destination plane that depends only on v.

For each internal node v, let ($v = v_1, v_2, \ldots, v_k =$ the rightmost leaf of the left subtree of v) be the path from v to v_k, and let ($v = v'_1, v'_2, \ldots, v'_{k'} =$ the leftmost leaf of the right subtree of v) be the path from v to $v'_{k'}$. To perform the SEARCH operation efficiently, we store at v the *bisector image* $b[v] = b(U[v_1]U[v_2]\cdots U[v_k](s), U[v'_1]U[v'_2]\cdots U[v'_{k'}](s))$, which is the bisector between the source image stored at v_k and the source image stored at $v'_{k'}$, unfolded into the destination plane of $U[v_1]U[v_2]\cdots U[v_k]$ (or, *equivalently*, of $U[v'_1]U[v'_2]\cdots U[v'_{k'}]$). Note that, for any path π from v to a leaf in the subtree of v, the *destination plane* $\Lambda(v)$ of the resulting composition of the unfolding transformations stored at the nodes of π, in their order along π, is the same, and depends only on v (and independent of π). During any operation that modifies the data structure, we always maintain the invariant that $b[v]$ is unfolded onto $\Lambda(v)$. As already said, the updating of the fields $U[v], b[v]$, at nodes v affected by tree rebalancing rotations, is quite simple, and described in [34].

The procedure SEARCH with a query point q in $\Lambda(\mathsf{root})$ is performed as follows. We determine on which side of $b[\mathsf{root}]$ q lies, in constant time, and proceed to the left or to the right child of the root, accordingly. When we proceed from a node v to its child, we maintain the composition $U^*[v]$ of all unfolding transformations on the path from the root to v (by initializing $U^*[\mathsf{root}] := U[\mathsf{root}]$ and updating $U^*[w] := U^*[u]U[w]$ when processing a child w of a node u on the path). Thus, denoting by b the bisector whose corresponding image $b[v]$ is stored at v, we can determine on which side of b q lies, by computing the image $U^*[v]b[v]$, in $O(1)$ time. Since the height of the tree is only $O(\log n)$, it takes $O(\log n)$ time to SEARCH for the claimer of q.

Note that the result of the SEARCH operation is guaranteed to be correct only if the query point q is already covered by the wavefront (that is, the bisectors between consecutive generators in the list do not meet one another closer to s than the location of q). It is the "responsibility" of the algorithm to provide valid query points (in that sense).

Typical Manipulation of the Structure Initializing the unfolding fields is trivial when the unique singleton wavefront is initialized at $t = 0$ at s. In a typical step of updating some wavefront W, we have a contiguous subsequence W' of W, which we want to advance through a new polytope edge sequence $\mathcal{E}$ (given that all the source images in W are currently unfolded to the plane of the first facet of the corresponding facet sequence of $\mathcal{E}$; see Sect. 5.3 for further details). We perform two SPLIT operations that split T into three subtrees T^-, T', T^+, where T' stores W', and T^- (resp., T^+) stores the portion of W that precedes (resp., succeeds) W' (either of these two latter subtrees can be empty). Then we take the root r' of T', and replace $U[r']$

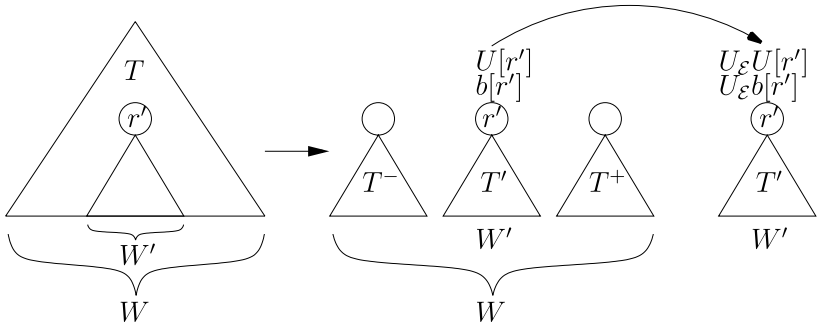


Fig. 34 T is split into three subtrees T^-, T', T^+, where T' stores the sub-wavefront W' of W. Then the unfolding fields stored at the root r' of T' are updated

by $U_{\mathcal{E}}U[r']$ and $b[r']$ by $U_{\mathcal{E}}b[r']$; see Fig. 34. Finally, we concatenate T^-, the new T', and T^+, into a common new tree T.

Remark The collection of the fields $U[v]$ and $b[v]$ in the resulting data structure is actually a dynamic version of the *incidence data structure* of Mount [28], which stores the incidence information between m nonintersecting geodesic paths and n polytope edges; the main novelty is the dynamic nature of the structure and the optimal construction time of $O((n+m)\log(n+m))$. (Mount constructs his data structure in time proportional to the number of intersections between the polytope edges and the geodesic paths, which is $\Theta(nm)$ in the worst case.)

Maintaining all Versions We also require our data structure to be *confluently persistent* [12]; that is, we need the ability to maintain, operate on, and modify past versions of any list (wavefront), and we need the ability to *merge* (in the terminology of [12]) existing distinct versions into a new version. Consider, for example, a transparent edge e and two transparent edges f, g in *output*(e). We propagate $W(e)$ to compute $W(e, f), W(e, g)$; the first propagation has modified $W(e)$, and the second propagation goes back to the old version of $W(e)$ and modifies it in a different manner. Moreover, later, when f, say, is ascertained to be covered, we merge $W(e, f)$ with other wavefronts that have reached f, to compute $W(f)$, and then propagate $W(f)$ further. At some later time g is ascertained to be covered, and we merge $W(e, g)$ with other wavefronts at g into $W(g)$. Thus, not only do we need to retrieve older versions of the wavefront, but we also need to merge them with other versions.

We also use the persistence of the data structure to implement the wavefront propagation through a block tree, as described in Sect. 5.3.1 below. Specifically, our propagation simulation uses a "trial and error" method; when an "error" is discovered, we restart the simulation from an earlier point in time, using an older version of the wavefront.

Each of the three kinds of operations, CONCATENATE, SPLIT and DELETE, uses $O(1)$ storage for each node of the binary tree that it accesses, so we can make the data structure confluently persistent by path-copying [20]. Each of our operations affects $O(\log n)$ nodes of the tree, including all the ancestors of every affected node.

Once we have determined which nodes an operation will affect, and before the operation modifies any node, we copy all the affected nodes, and then modify the copies as needed. This creates a new version of the tree while leaving the old version unchanged; to access the new version we can simply use a pointer to the new root, so traversing it is done exactly as in the ephemeral case. In summary, we have:

Lemma 5.1 *There exists a data structure that represents a one-sided wavefront and supports all the list operations, priority queue operations, and unfolding operations, as described above, in $O(\log n)$ worst-case time per operation. The size of the data structure is linear in the number of generators; it can be made confluently persistent at the cost of $O(\log n)$ additional storage per operation.*

5.2 Overview of the Wavefront Propagation Stage

Recall from Sect. 4 that the two main subroutines of the algorithm are wavefront propagation and wavefront merging. In this and the following subsection we describe the implementation details of the first procedure; the merging is discussed in Sect. 4.2, which, together with the data structure details presented in Sect. 5.1, implies that all the merging procedures can be executed in $O(n \log n)$ time.

Let e be a transparent edge. We now show how to propagate a given one-sided wavefront $W(e)$ to another edge $g \in \mathit{output}(e)$ (that is, $e \in \mathit{input}(g)$), denoting, as above, the resulting propagated wavefronts by $W_{H_1}(e, g), \ldots, W_{H_k}(e, g)$, where $H_1, \ldots, H_k$ are all the relevant homotopy classes that correspond to block sequences from e to g within $R(g)$ (see Sect. 3.3); note that a transparent endpoint "splits" a homotopy class, similarly to a vertex of P. In the process, we also determine the time of first contact between each such $W(e, g)$ and the endpoints of g.

The high-level description of the algorithm is a sequence of steps, each of which propagates a wavefront $W(e)$ from one transparent edge e to another $g \in \mathit{output}(e)$, within a fixed homotopy class H, to form $W_H(e, g)$.[11] Nevertheless, in the actual implementation, when we start the propagation from e, all the topologically constrained wavefronts $W_H(e, g)$, over all relevant g and H, are treated as *a single wavefront* W. At the beginning of the propagation, W is split into k_1 initial sub-wavefronts, where k_1 is the number of building blocks that e bounds (on the side into which we propagate W); during the propagation, these initial wavefronts are further split into a total of k sub-wavefronts, one per homotopy class.

Let c be the surface cell for which $e \subset \partial c$, and $W(e)$ enters c after reaching e. We describe in the next subsection a procedure for computing (all the relevant topologically constrained wavefronts) $W(e, g)$ for any transparent edge $g \subset \partial c$. To compute $W(e, g)$ for all transparent edges $g \in \mathit{output}(e)$, possibly not belonging to ∂c, we proceed as follows. We propagate $W(e)$ cell-by-cell inside $R(g)$ from e to g, and effectively split the wavefront into multiple *component wavefronts*, each labeled by the sequence of $O(1)$ transparent edges it traverses from e to g. We propagate a wavefront W from e to g inside a single surface cell, either when W is one of the two one-sided wavefronts merged at e, or when W has reached e on its way to g from

[11] The initial singleton wavefront $W(s)$ from s to a transparent edge g on the boundary of the cell that contains s is propagated similarly.

some other transparent edge $f \in \mathit{input}(g)$ (without being merged with other component wavefronts at e). In what follows, we treat W as in the former case; the latter case is similar.

5.3 Wavefront Propagation in a Single Cell

So far we have considered a wavefront as a static structure, namely, as a sequence of generators that reach a transparent edge. We now describe a "kinetic" form of the wavefront, in which we track changes in the combinatorial structure of the wavefront $W(e)$ as it sweeps from its origin transparent edge e across a single cell c. Our simulation detects and processes any bisector event in which a wave of $W(e)$ is eliminated by its two neighboring waves inside c; actually, the propagation may also detect some events that occur in $O(1)$ nearby cells, as described in detail below. Events are detected and processed in order of increasing distance from s, that is, in simulation time order. However, *the simulation clock t is not updated during the propagation inside c*; that is, the propagation from an edge e to all the edges in $\mathit{output}(e)$ is done without "external interruptions" of propagating from other fully covered transparent edges that need processing. The effect of the propagated wavefront $W(e, g)$, for $g \in \mathit{output}(e)$, on the simulation clock is in its updating of the values $\mathit{covertime}(g)$; the actual updating of t occurs only when we select a new transparent edge e' with minimum $\mathit{covertime}(e')$ for processing—see Sect. 4.1.

We propagate the wavefront separately in each of the $O(1)$ block trees of the Riemann structure $\mathcal{T}(e)$. Let $W(e)$ be the one-sided wavefront that reaches e from outside c; it is represented as an ordered list of source images, each claiming some (contiguous and *nonempty*) portion of e. To prepare $W(e)$ for propagation in c, we first SPLIT $W(e)$ into $O(1)$ sub-wavefronts, according to the subdivision of e by building blocks of c. A sub-wavefront that claims the segment of e that bounds a building block B of c is going to be propagated in the block tree $T_B(e) \in \mathcal{T}(e)$.

By propagating $W(e)$ from e in all the trees of $\mathcal{T}(e)$ within c, we compute $O(1)$ new component wavefronts that reach other transparent edges of ∂c. If e is the initial edge in this propagation step, then, by Corollary 3.17, these component wavefronts collectively encode all the shortest paths from s to points p of c that enter c through e and do not leave c before reaching p. In general, this property holds for all the cells c' in $R(e)$, as follows easily from the construction. Hence, these component wavefronts, collected over all propagation steps that traverse c, contain all the needed information to construct (an implicit representation of) SPM(s) within c.

5.3.1 Wavefront Propagation in a Single Block Tree

Let $T_B(e)$ be a block tree in $\mathcal{T}(e)$, and denote by e_B the sub-edge $\partial B \cap e$. Denote by $W(e_B)$ the sub-list of generators of $W(e)$ that claim points on e_B (recall that $W(e)$ claims a single connected portion of e, which may or may not contain the endpoints of e, or of e_B). Let $W = W(t)$ denote the kinetic wavefront within the blocks of $T_B(e)$ at any time t during the simulation; initially, $W = W(t_0) = W(e_B)$. Note that even though we need to start the propagation from e at simulation time $t_0 = \mathit{covertime}(e)$, the actual starting time may be strictly smaller, since there may have been bisector events beyond e that have occurred before time $\mathit{covertime}(e)$. In

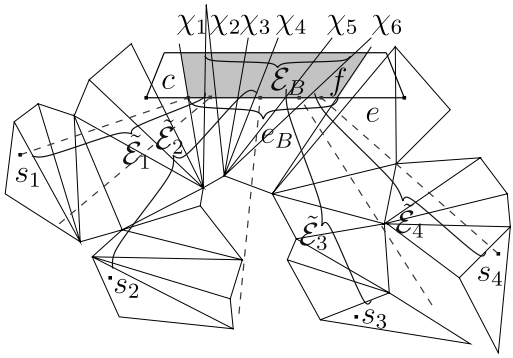


Fig. 35 The block B is *shaded*; the edge sequence associated with B is $\mathcal{E}_B = (\chi_1, \ldots, \chi_6)$. $W(e_B)$ consists of four source images $s_1, \ldots, s_4$, all unfolded to the plane of the facet f before the simulation of the propagation into $T_B(e)$ starts (that is, the last facet of the facet sequence corresponding to each $\mathcal{E}_i$ is f). Specifically, $\mathcal{E}_1 = \tilde{\mathcal{E}}_1 || (\chi_2, \ldots, \chi_6)$, $\mathcal{E}_2 = \tilde{\mathcal{E}}_2 || (\chi_5, \chi_6)$, $\mathcal{E}_3 = \tilde{\mathcal{E}}_3 \setminus (\chi_6, \chi_5)$ and $\mathcal{E}_4 = \tilde{\mathcal{E}}_4 \setminus (\chi_6)$

this case, these events need now to be processed (up to now, they have been detected by the algorithm but not processed yet), and we set t_0 to be the time when the earliest among them takes place.

Denote by $\mathcal{E}_B$ an edge sequence associated with B (any one of the two oppositely ordered such sequences, for blocks of type II, III), and by $\mathcal{F}_B$ its corresponding facet sequence. We can then write $W = (s_1, s_2, \ldots, s_k)$, so that, for each i, we have $s_i = U_{\mathcal{E}_i}(s)$, where $\mathcal{E}_i$ is defined as follows. Denote by $\tilde{\mathcal{E}}_i$ the maximal polytope edge sequence traversed by the wave of s_i from s_i to the points that it claims on e; $\tilde{\mathcal{E}}_i$ must overlap either with a portion of $\mathcal{E}_B$ or with a portion of the reverse sequence $\mathcal{E}_B^{\text{rev}}$. In the former case we extend $\tilde{\mathcal{E}}_i$ by the appropriate suffix of $\mathcal{E}_B$ (which takes us to f in Fig. 35). In the latter case we truncate $\tilde{\mathcal{E}}_i$ at the first polytope edge of $\mathcal{E}_B^{\text{rev}}$ that it meets, and then extend it by the appropriate suffix of $\mathcal{E}_B$. However, the algorithm does not compute these sequences explicitly (and does not perform the "extend" or "truncate" operations); it only stores and composes their unfolding transformations, as described in Sect. 5.1. Denote by $\Lambda(W)$ the (common) destination plane of all the $U_{\mathcal{E}_i}$. We do not alter $\Lambda(W)$ until the propagation of W in $T_B(e)$ is completed (and then $\Lambda(W)$ is updated, as described below). That is, as we traverse new blocks of $T_B(e)$, we unfold them all to the plane $\Lambda(W)$. When we propagate the initial singleton wavefront directly from s in $T_B(s)$, we initialize $W := (s)$, so that the maximal polytope edge sequence $\mathcal{E}$ of s is empty, and $U_{\mathcal{E}}$ is the identity transformation I. This setting is appropriate since s is assumed to be a vertex of P, and therefore all the polytope edges in $\mathcal{E}_B$ emerge from s, so it lies on all the facets of $\mathcal{F}_B$, and, particularly, on the last facet of $\mathcal{F}_B$.

The *boundary chain* $\mathcal{C}$ of $T_B(e)$ is recursively defined as follows. Initially, we put in $\mathcal{C}$ all the boundary edges of ∂B, other than e_B. We then proceed top-down through $T_B(e)$. For each node B' of $T_B(e)$ and for each child B'' of B', we remove from the current $\mathcal{C}$ the contact interval connecting B' and B'', and replace it by the remaining boundary portion of B''. This results in a connected (unfolded) polygonal boundary chain that shares endpoints with $B \cap e$. Since $T_B(e)$ has $O(1)$ nodes, and each block has $O(1)$ boundary elements, $\mathcal{C}$ contains only $O(1)$ elements; see Fig. 36.

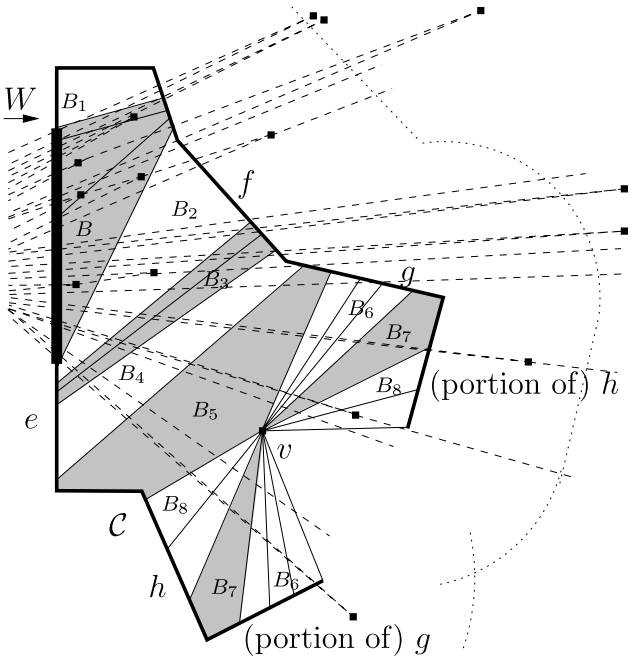


Fig. 36 Bisector events (the *thick square points*), some of which are processed during the propagation of the wavefront W from the transparent edge portion e_B (the *thickest segment* in this figure) through the building blocks (their shadings alternate) of the block tree $T_B(e)$. The unfolded transparent edges are drawn as *thick solid lines*, while the unfolded contact intervals are *thin solid lines*. The bisectors of the generators of W, as it sweeps through the unfolded blocks, are shown dashed. The union of all the blocks in $T_B(e)$ is bounded by e_B and the boundary chain $\mathcal{C}$ (which is non-overlapping in this example). The *dotted lines* indicate the distance from the transparent edges in $\mathcal{C}$ within which we still process bisector events of W. For each transparent edge f of $\mathcal{C}$, we can stop propagating the wavefront portion $W(e_B, f)$ that has reached f after it crosses the dotted line (which lies at distance $2|f|$ from f), since f must have already been fully swept at that time by the waves of $W(e_B, f)$

When W is propagated towards $\mathcal{C}$, the most important property is that each transparent edge or contact interval of $\mathcal{C}$ can be reached only by a *single topologically constrained sub-wavefront* of W, since, if W splits on its way, the new sub-wavefronts reach different elements of $\mathcal{C}$. (The property does not hold for ∂c, since, when c contains holes and/or a vertex of P, there is more than one way to reach a transparent edge $f \in \partial c$—in such cases f appears more than once in $\mathcal{C}$, each time as a distinct element, as illustrated in Fig. 36.) In the rest of this section, whenever a resulting wavefront $W(e, f)$ is mentioned for some $f \in \mathcal{C}$, we interpret $W(e, f)$ as $W_H(e, f)$ for the unique homotopy class H that constrains W on its way from e to this specific incarnation of f along $\mathcal{C}$.

We denote by *range*(W) the subset of segments of $\mathcal{C}$ that can potentially be reached by W, initialized as *range*$(W) := \mathcal{C}$. As W is propagated (and split), *range*(W) is updated (that is, split and/or truncated) accordingly, as described below.

Critical Events and Simulation Restarts We simulate the continuous propagation of W by updating it at the (discrete) critical events that change its topology during its

propagation in $T_B(e)$. There are two types of these events—bisector events (of the first kind), when a wave of W is eliminated by its two neighbors, and vertex events, when W reaches a vertex of $\mathcal{C}$ (either transparent or a real vertex of P) and has to be split. Before we describe in detail the processing of these events, we provide here the intuition behind the (somewhat unorthodox implementation of the) low-level procedures.

The purpose of the propagation of W in $T_B(e)$ is to compute the wavefronts $W(e_B, f)$, for each transparent edge f in $\mathcal{C}$ that W reaches. To do so, we have to correctly update W at those critical events that are *true with respect to the propagation of W in $T_B(e)$*; that is, events that take place in $T_B(e)$ that would have been vertices of SPM(s) if there were no other wavefronts except W. For the sake of brevity, in the rest of this section we refer to these events simply as *true events*. Unfortunately, it is difficult to determine in "real time" the exact set of true events (mainly because of vertex events—see below). Instead, we determine on the fly a larger set of *candidates* for critical events, which is guaranteed to contain all the true events, but which might also contain events that are *false with respect to the propagation of W in $T_B(e)$*; in the rest of this section we refer to events of the latter kind as *false events*. The candidates that turn out to be false events either are bisector events that involve at least one generator s' of W so that the path from s' to the event location intersects $\mathcal{C}$, or take place later than some earlier true event that has not yet been detected (and processed).

Let x be such a *candidate bisector event* that takes place at simulation time t_x. If all the true events of W that *have taken place before t_x* were *processed before t_x*, then x can be *foreseen* at the last critical event at which one of the bisectors involved in x was updated before time t_x, using the *priorities* assigned to the source images in W. The priority of a source image s' is the distance from s' to the point at which the two (unfolded) bisectors of s' intersect beyond e_B, either in B or beyond it. The priority is $+\infty$ if the bisectors do not intersect beyond e_B. (Initially, when W contains the single wave from s, the priority of s is defined to be $+\infty$.) Whenever a bisector of a source image s' is updated (as detailed below), the priority of s' is updated accordingly.

A *candidate vertex event* cannot be foreseen so easily, since we do not know which source image of W claims a vertex v (because of the critical events that might change W before it reaches v), until v is actually reached by W. Even when v is reached by W, we do not have in the data structure a "warning" that this vertex event is about to take place. Instead, we detect the vertex event that occurs at v only later and indirectly, either when processing some later candidate event (which is false as it was computed without taking into account the event at v—see Fig. 37(a), (b)), or when the propagation of W in $T_B(e)$ is stopped at a later simulation time, when a segment f of $\mathcal{C}$ incident to v is ascertained to be fully covered, as illustrated in Fig. 37(c).

When we detect a vertex event at some vertex v which is reached by W at time t_v, so that at least one candidate critical event of W that takes place later than t_v has already been processed, *all the versions of the (persistent) data structure that encode W after time t_v become invalid*, since they do not reflect the update that occurs at t_v. To correct this situation, we discard all the invalid versions of W, and *restart the simulation of the propagation of the last valid version of W from time t_v*. This time, however, we SPLIT W at v (at simulation time t_v) into two new sub-wavefronts, as detailed below. Note that this step does not guarantee that the current event at v is a true event, since there might still exist undetected earlier vertex events, which, when

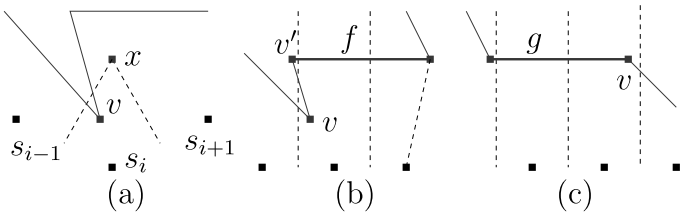


Fig. 37 An earlier vertex event at $v \in \mathcal{C}$ can be detected later: (**a**) while processing a false bisector event x; (**b**) while processing a vertex event at an endpoint v' of a segment $f \subset \mathcal{C}$, when f is ascertained to be covered by W; (**c**) when the segment $g \subset \mathcal{C}$, incident to v, is ascertained to be covered by W

eventually detected later, will cause the simulation to be restarted again, making the current event at v invalid (and we will have to wait until the wavefront reaches v again).

Path Tracing Let $x \in \Lambda(W)$ be an (unfolded) image of some point of c, and let $s' \in W$ be a source image. To determine whether the path to x from s' does or does not meet $\mathcal{C}$, and, in the former case, to also determine the first intersection point (along the path $\pi(s', x)$) with $\mathcal{C}$, we *trace* $\pi(s', x)$ either up to x, or until it intersects $\mathcal{C}$—whichever occurs first—as follows.[12]

The tracing is done by following the sequence of blocks traversed by $\pi(s', x)$, which forms a path in $T_B(e)$. At each block B' that we encounter, we test whether $\pi(s', x)$ terminates within B', and, if not, we find the edge of $\partial B'$ through which $\pi(s', x)$ leaves B'. If we reach x, or if the exit edge of $\partial B'$ is a portion of $\mathcal{C}$, we stop the tracing. Otherwise we exit B' through a contact interval I, and proceed to the next block beyond I. (It is also possible that we reach a contact interval I which is a "dead-end" in $T_B(e)$, and is thus a portion of $\mathcal{C}$.)

At each step we proceed in $T_B(e)$ from a node to its child; since the depth of $T_B(e)$ is $O(1)$, we are done after $O(1)$ steps. Since at each step we compute $O(1)$ unfoldings of paths and transparent edges, and each unfolding operation takes $O(\log n)$ time to perform, using the data structures described in Sects. 2.4 and 5.1, the whole tracing procedure takes $O(\log n)$ time.

Corollary 5.2 *Tracing the path $\pi(s', p)$ from a generator $s' \in W$ to a point p without intersecting $\mathcal{C}$, correctly determines the distance $d(s', p)$.*

Proof Follows from the description of the tracing procedure. □

Note that we can similarly trace any path π of W until it intersects $\mathcal{C}$, without specifying any terminal point on π, as long as the starting direction of π in $\Lambda(W)$ is well defined.

[12]Here and in the rest of this section, whenever we say that a *path* π from a generator $s' \in W$ intersects $\mathcal{C}$, we actually mean that only the portion of π from s' to the first intersection point $x = \pi \cap \mathcal{C}$ is a valid geodesic path; the portion of π beyond x is merely a straight segment along the direction of π on $\Lambda(W)$. Still, for the sake of simplicity, we call π (including possibly a portion beyond x) a *path* (from s' to the terminal point of π).

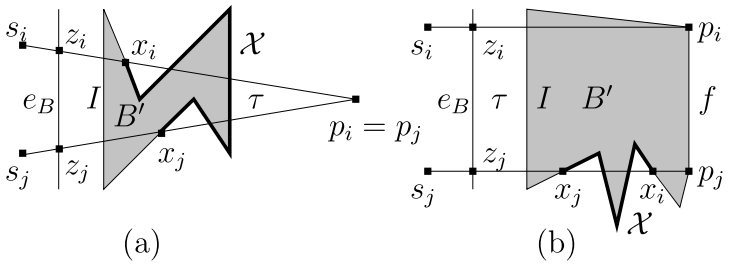


Fig. 38 (**a**) $\pi(s_i, p_i), \pi(s_j, p_j)$ leave B' through two different contact intervals of $\partial B'$. Here $p_i = p_j$, and τ is the triangle $z_i p_i z_j$. (**b**) $\pi(s_i, p_i)$ reaches $p_i \in B'$ and $\pi(s_j, p_j)$ leaves B' at the point x_j. Here $p_i \neq p_j$, and τ is the quadrilateral $z_i p_i p_j z_j$. The portion $\mathcal{X}$ of $\partial B'$ is highlighted in both cases

The following technical lemma is needed later for the correctness analysis of the simulation algorithm—in particular, for the analysis of critical event processing. See Fig. 38.

Lemma 5.3 *Let s_i, s_j be a pair of generators in W, and let p_i, p_j be a pair of (possibly coinciding) points in $\Lambda(W)$, so that $\pi(s_i, p_i)$ and $\pi(s_j, p_j)$ do not intersect each other (except possibly at their terminal point, if $p_i = p_j$), and if $p_i \neq p_j$ then $f = \overline{p_i p_j}$ is a straight segment of $\mathcal{C}$. Denote by z_i (resp., z_j) the intersection point $\pi(s_i, p_i) \cap e_B$ (resp., $\pi(s_j, p_j) \cap e_B$), and denote by τ the unfolded convex quadrilateral (or triangle) $z_i p_i p_j z_j$. Let B' be the last building block of the maximal common prefix block sequence along which both $\pi(s_i, p_i)$ and $\pi(s_j, p_j)$ are traced (before possibly diverging into different blocks).*

If only one of the two paths leaves B', or if $\pi(s_i, p_i)$ and $\pi(s_j, p_j)$ leave B' through different contact intervals of $\partial B'$, then the region $B' \cap \tau$ contains at least one vertex of $\mathcal{C}$ that is visible, within the unfolded blocks of $T_B(e)$, from every point of $\overline{z_1 z_2} \subseteq e_B$.

Proof Assume for simplicity that $B' \neq B$. The paths $\pi(s_i, p_i), \pi(s_j, p_j)$ must enter B' through a common contact interval I of $\partial B'$. Consider first the case where $\pi(s_i, p_i), \pi(s_j, p_j)$ leave B' through two respective different contact intervals I_i, I_j of $\partial B'$, and denote their first points of intersection with $\partial B'$ by x_i and x_j, respectively—see Fig. 38(a). Denote by $\mathcal{X}$ the portion of $\partial B'$ between x_i and x_j that does not contain I; $\mathcal{X}$ must contain at least one vertex of $\partial B'$. By definition, each vertex of a building block is a vertex of $\mathcal{C}$; note that the extreme vertices of $\mathcal{X}$ are x_i and x_j, which may or may not be vertices of $\mathcal{C}$. Since the unfolded image of $\mathcal{X}$ is a simple polygonal line that connects $\pi(s_i, x_i)$ and $\pi(s_j, x_j)$, and intersects neither $\pi(s_i, x_i)$ nor $\pi(s_j, x_j)$, it is easily checked that we can sweep τ by a line parallel to e_B, starting from e_B, until we encounter a vertex v of $\mathcal{X}$ within τ, which is also a vertex of $\mathcal{C}$: Either x_i or x_j is such a vertex, or else τ must contain an endpoint of either I_i or I_j. Therefore v is visible from each point of $\overline{z_1 z_2}$.

Consider next the case in which only one of $\pi(s_i, p_i), \pi(s_j, p_j)$ leaves B', and assume, without loss of generality, that $\pi(s_i, p_i)$ reaches $p_i \in B'$ and $\pi(s_j, p_j)$ leaves B' at the point x_j before reaching p_j—see Fig. 38(b). Denote by $\pi(p_j, s_j)$ the path $\pi(s_j, p_j)$ directed from p_j to s_j, and denote by π' the concatenation

$\pi(s_i, p_i)||\overline{p_i p_j}||\pi(p_j, s_j)$. The path $\pi(s_i, p_i)$ does not leave B', and, by assumption, the segment $\overline{p_i p_j}$ is either an empty segment or a segment of $\partial B'$, and therefore the only portion of π' that leaves B' is $\pi(p_j, s_j)$. Denote by x_i the first point along $\pi(p_j, s_j)$ (beyond p_j itself) that lies on $\partial B'$; if $\pi(p_j, s_j)$ leaves B' immediately, we do take $x_i = p_j$. Since (the unfolded) $\pi(p_j, s_j)$ is a straight segment, and since, for each segment f' of $\partial B'$, B' lies locally only on one side of f', it follows that x_i and x_j lie on different segments of $\partial B'$. Define $\mathcal{X}$ as above; here it connects the prefixes of π' and $\pi(s_j, p_j)$, up to x_i and x_j, respectively, and the proof continues as in the previous case. □

Stopping Times and Their Maintenance The simulation of the propagation of W in the blocks of $T_B(e)$ processes candidate bisector events in order of increasing priority, up to some time $t_{\text{stop}}(W)$, which is initialized to $+\infty$, and is updated during the propagation.[13] When the time $t_{\text{stop}}(W)$ is reached, the following holds: Either $t_{\text{stop}}(W) = +\infty$ (see Fig. 39(a)), all the known candidate critical events of W in the blocks of $T_B(e)$ have been processed, and all the waves of W that were not eliminated at these events have reached $\mathcal{C}$; or $t_{\text{stop}}(W) < +\infty$ (see Fig. 39(b)), and there exists some sub-wavefront $W' \subseteq W$ that claims some segment (a transparent edge or a contact interval) f of *range*(W) (that is, f is ascertained to have been covered by W' not later than at time $t_{\text{stop}}(W)$), such that all the currently known candidate events of W' have been processed before time $t_{\text{stop}}(W)$. In the former case we split W into sub-wavefronts $W(e, f)$ for each segment $f \in$ *range*(W); in the latter case, we extract from W (by splitting it) the sub-wavefront $W(e, f) = W'$ that has covered f. When we split W into a pair of sub-wavefronts W_1, W_2, the time $t_{\text{stop}}(W_1)$ (resp., $t_{\text{stop}}(W_2)$) replaces $t_{\text{stop}}(W)$ in the subsequent propagation of W_1 (resp., W_2), following the same rule, while $t_{\text{stop}}(W)$ plays no further role in the propagation process.

For each segment f in $\mathcal{C}$, we maintain an individual time $t_{\text{stop}}(f)$, which is a conservative upper estimate of the time when f is completely covered by W during the propagation in $T_B(e)$. Initially, we set $t_{\text{stop}}(f) := +\infty$ for each such f. As

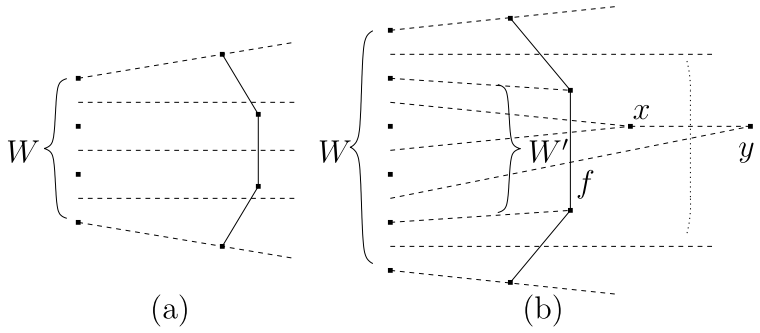


Fig. 39 (**a**) The stopping time $t_{\text{stop}}(W) = +\infty$. (**b**) The stopping time $t_{\text{stop}}(W') = t_{\text{stop}}(f) < +\infty$; the *dotted line* indicates the stopping time (or distance) at which we stop processing bisector events: the event at x has been processed before $t_{\text{stop}}(W')$, while the event at y has been detected but not processed

[13] The present description also applies to appropriate sub-wavefronts that have already been split from W—see below.

detailed below, we update $t_{\text{stop}}(f)$ whenever we trace a path from a generator in W that reaches f (without reaching $\mathcal{C}$ beforehand); by Corollary 5.2, these updates are always valid (i.e., do not depend on simulation restarts). The time $t_{\text{stop}}(W)$ is the minimum of all such times $t_{\text{stop}}(f)$, where f is a segment of $range(W)$. Whenever $t_{\text{stop}}(f)$ is updated for such an f, we also update $t_{\text{stop}}(W)$ accordingly. When the simulation clock reaches $t_{\text{stop}}(W)$, either some f of $range(W)$ is completely covered by the wavefront W, so that $t_{\text{stop}}(f) = t_{\text{stop}}(W)$, or the priority of the next event of W in the priority queue is $+\infty$, in which case $t_{\text{stop}}(W) = +\infty$.

As shown below, $range(W)$ is maintained correctly, independently of simulation restarts; therefore, when $range(W)$ contains only one segment, no further vertex events may cause a restart of the simulation of the propagation of W (since a simulation restart of a wavefront that is separated from W does not affect W, and the vertex events at the endpoints of f have already been processed, since W and $range(W)$ have already been split at them).

Note that there is a gap of at most $|f|$ time between the time t_f when the segment f of $\mathcal{C}$ is *first reached* by W and the time when f is *completely covered* by W. In particular, it is possible that both endpoints of f are reached by W before f is completely covered by W—see Fig. 40(a). It is also possible, because of visibility constraints, that W reaches only a portion of f in our propagation algorithm (and then there must be other topologically constrained wavefronts that reach the portions of f that are not reached by W). Still we say that f is *covered by* W at time $t_f + |f|$, as if we were propagating also the non-geodesic paths that progress along f from the first point of contact between W and f. See Fig. 40(b).

The algorithm does not necessarily detect the first time t_f when f is reached by W. Instead, we detect a time t'_f, when *some* path encoded in some wave of W reaches f. However, in order to estimate the time when f is completely covered by W correctly (although somewhat conservatively), the algorithm sets $t_{\text{stop}}(f) := t'_f + |f|$. We show below that t'_f is greater than t_f by at most $|f|$, hence the total gap between the time when f is first reached by W, and the time when the algorithm ascertains that f is completely covered, is at most $2|f|$.

Consider W', the sub-wavefront of W that covers a segment f of $\mathcal{C}$. If f is a transparent edge, the well-covering property of f ensures that during these $2|f|$ simulation time units (since t_f) no wave of W' has reached "too far" beyond f. That is, all the bisector events of W' beyond f that have been detected and processed before $t_{\text{stop}}(f)$ occur in $O(1)$ cells near c (see Fig. 36). This invariant is crucial for the time complexity of the algorithm, as it implies that no bisector event is detected more than $O(1)$ times—see below. If f is a contact interval, the paths encoded in W that reach f in our propagation do not reach f in the real SPM(s), by Corollary 3.17; therefore

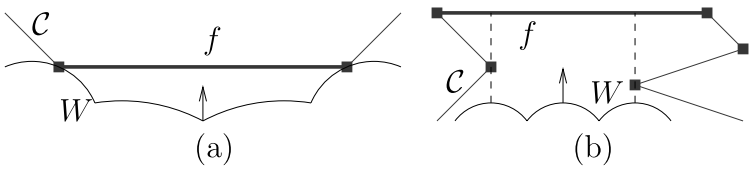


Fig. 40 (**a**) Both endpoints of f are reached by W before f is covered by W. (**b**) W actually reaches only a portion of f (between the two *dashed lines*), because of visibility constraints

these paths do not leave c (as shortest paths), and need not be encoded in the one-sided wavefronts that leave c. This property is also used below in the time complexity analysis of the algorithm.

Processing Candidate Bisector Events As long as the simulation clock has not yet reached $t_{\text{stop}}(W)$, at each step of the simulation we extract from the priority queue of W the candidate bisector event which involves the generator s_i with the minimum priority in the queue, and process it according to the high-level description in Sect. 4.3, the details of which are given next. Let x denote the unfolded image of the location of the candidate event (the intersection point of the two bisectors of s_i), and denote by W' the constant-size sub-wavefront of W that encodes the paths involved in the event. If s_i is neither the first nor the last source image in W, then $W' = (s_{i-1}, s_i, s_{i+1})$. The generator s_i cannot be the only source image in W, since in this case its two bisectors would be rays emanating from s_i, and two such rays cannot intersect (beyond e). If s_i is either the first or the last source image in W, then W' is either (s_i, s_{i+1}) or (s_{i-1}, s_i), respectively. Denote by π_1 (resp., π_2) the path from the first (resp., last) source image of W' to x, or, more precisely, the respective unfolded straight segments of (common) length $\mathit{priority}(s_i)$.

We use the tracing procedure defined above for each of the paths π_1, π_2. For any path π, denote by $\mathcal{C}(\pi)$ the first element of $\mathcal{C}$ (along π) that π intersects, if such a point exists. The following two cases can arise:

Case (i): The bisector event at x is *true with respect to the propagation of W in $T_B(e)$* (see Fig. 41(a)), which means that neither π_1 nor π_2 intersects $\mathcal{C}$, and both paths are traced along a common block sequence in $T_B(e)$. (Recall that the unfolded blocks of $T_B(e)$ might overlap each other; see Fig. 41(b).) By definition of a block tree, this is a necessary and *sufficient* condition for the event to be true (with respect to the propagation of W in $T_B(e)$); however, a following simulation restart might still discard this candidate event, forcing the simulation to reach it again. If s_i is neither the first nor the last source image in W, we DELETE s_i from W, and recompute the priorities of its neighbors s_{i-1}, s_{i+1}, as follows. Since all the source images of W are currently unfolded to the same plane $\Lambda(W)$, we can compute, in constant time, the intersection point p, if it exists, of the new bisector $b(s_{i-1}, s_{i+1})$ (stored in the data structure during the DELETE operation) with the bisector of s_{i-1} that is not incident to x. If the two bisectors do not intersect each other (p does not exist), we put $\mathit{priority}(s_{i-1}) := +\infty$; otherwise $\mathit{priority}(s_{i-1})$ is the length of the straight line from s_{i-1} to p, ignoring any visibility constraints, or the possibility that the two bisectors reach p through different block sequences. The priority of s_{i+1} is recomputed similarly.

If $s_i = s_1$ is the first but not the last source image in W, we DELETE s_1 from W (that is, s_2 becomes the first source image in W), and define the first (unfolded) bisector b of W as a ray from s_2 through x; the priority of s_2 is recomputed as above, intersecting b with the other bisector of s_2. If s_i is the last but not the first source image in W, it is handled symmetrically.

Case (ii): The bisector event at x is *false with respect to the propagation of W in $T_B(e)$*: Either at least one of the paths π_1, π_2 intersects $\mathcal{C}$, or π_1, π_2 are traced towards x along different block sequences in $T_B(e)$, reaching the location x in different layers of the Riemann structure that overlap at x. See Fig. 41(b–d) for an illustration.

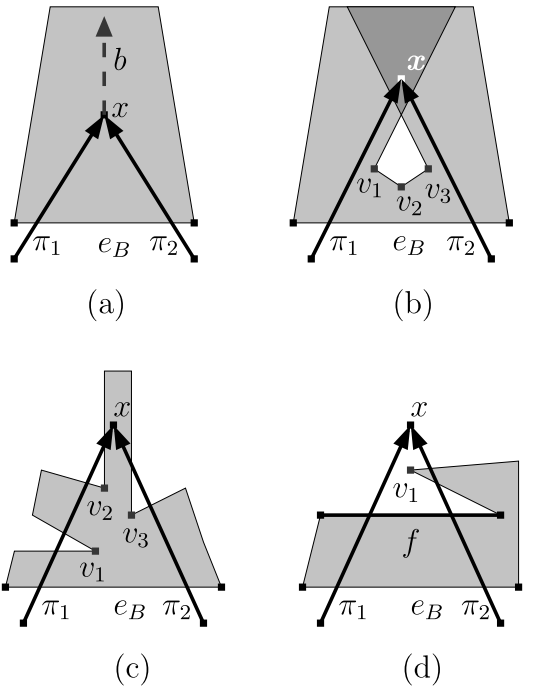


Fig. 41 In (**a**) x is a true bisector event; the new bisector b between the generators of π_1, π_2 is shown dashed. In (**b–d**) x is a false candidate. (**b**) π_1, π_2 do not intersect $\mathcal{C}$, but reach x through different layers of the Riemann structure that overlap each other. At least one vertex of $\mathcal{V} = \{v_1, v_2, v_3\}$ is visible from the portion of e_B between π_1 and π_2; the same is true in (**c**), where both π_1, π_2 intersect $\mathcal{C}$ (before reaching x). (**d**) $\mathcal{C}(\pi_1) = \mathcal{C}(\pi_2) = f$. No vertex of $\mathcal{V}$ (here $\mathcal{V} = \{v_1\}$) is visible from the portion of e_B between π_1 and π_2

If π_1 intersects $\mathcal{C}$, denote the first such intersection point (along π_1) by z and the segment $\mathcal{C}(\pi_1)$, which contains z, by f. We compute z and update $t_{\text{stop}}(f) := \min\{t_{\text{stop}}(f), d_z + |f|\}$, where d_z is the distance from s to z along π_1. As described above, and with the visibility caveats noted there, the expression $d_z + |f|$ is a time at which W will certainly have swept over f. We also update $t_{\text{stop}}(W) := \min\{t_{\text{stop}}(f), t_{\text{stop}}(W)\}$. If, as the result of this update, $t_{\text{stop}}(W)$ becomes less than or equal to the current simulation time, we conclude that f is already fully covered. We then stop the propagation of W and process f as a covered segment of $\mathcal{C}$ (as described below), immediately after completing the processing of the current bisector event. Note that in this case, that is, when $t_{\text{stop}}(f)$ gets updated because of the detection of the crossing of the wavefront of f at z, which causes $t_{\text{stop}}(W)$ to go below the current simulation clock t, we have $t_{\text{stop}}(W) = t_{\text{stop}}(f) = d_z + |f| \le t = d_z + d(z, x)$, where $d(z, x)$ is the distance from z to x along π_1; see Fig. 42. Hence $d(z, x) \ge |f|$. This however violates the invariant that we want to maintain, namely, that we only process bisector events that lie no farther than $|f|$ from an edge f of $\mathcal{C}$. Nevertheless, this can happen at most once per edge f, because from now on $t_{\text{stop}}(W)$ will not exceed $t_{\text{stop}}(f)$. We will use this property in the time complexity analysis below.

If π_2 intersects $\mathcal{C}$, we treat it similarly.

Regardless of whether π_1, π_2, or neither of them, intersects $\mathcal{C}$, we then proceed as follows. Denote by τ the triangle bounded by the images of e, π_1 and π_2, unfolded to $\Lambda(W)$, and denote by $\mathcal{V}$ the set of the (at most $O(1)$) vertices of $\mathcal{C}$ that lie in the interior of τ. Since it takes $O(\log n)$ time to unfold each segment of $\mathcal{C}$, it takes $O(\log n)$ time to compute $\mathcal{V}$.

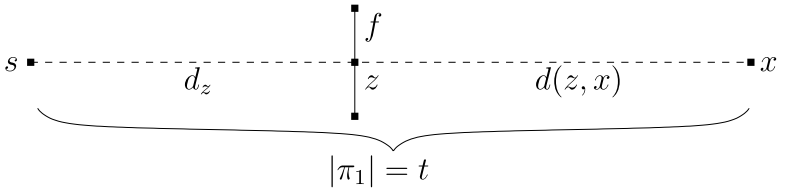


Fig. 42 If $d_z + |f| \leq t = d_z + d(z, x)$, then $d(z, x) \geq |f|$

Assume first that π_1, π_2 satisfy the assumptions of Lemma 5.3; it follows that $\mathcal{V}$ is not empty (see Fig. 41(b), (c)). We trace the path from each generator in W' to each vertex v of $\mathcal{V}$, and compute $claimer(v)$ (which satisfies $d(claimer(v), v) = \min\{\{d(s', v) \mid v \text{ is visible from } s' \in W'\} \cup \{+\infty\}\}$). Denote by u the vertex of $\mathcal{V}$ so that $t_u := d(claimer(u), u) = \min_{v \in \mathcal{V}} d(claimer(v), v)$; by Lemma 5.3, at least one vertex of $\mathcal{V}$ is visible from at least one generator in W', and therefore t_u is finite. As we will shortly show in Corollary 5.8, $t_u < t_x$ (where $t_x = priority(s_i)$ is the current simulation time). This implies that the propagation is invalid for $t \geq t_u$. We thus *restart* the propagation at time t_u, as follows.

Let W_u denote the last version of (the data structure of) W that has been computed before time t_u. We SPLIT W_u into sub-wavefronts W_1, W_2 at $s' := claimer(u)$ at the simulation time t_u, so that $range(W_1)$ is the prefix of $range(W_u)$ up to u, and $range(W_2)$ is the rest of $range(W_u)$ (to retrieve the range that is consistent with the version W_u we can simply store all the versions of $range(W)$—recall that each uses only constant space, because we can keep it unfolded). Discard all the later versions of W. We set $t_{\text{stop}}(W_1)$ (resp., $t_{\text{stop}}(W_2)$) to be the minimal $t_{\text{stop}}(f)$ value among all segments f in $range(W_1)$ (resp., $range(W_2)$). We replace the last (resp., first) unfolded bisector image of W_1 (resp., W_2) by the ray from s' through u, and correspondingly update the priority of s' in both new sub-wavefronts (recall from Sect. 5.1 that the SPLIT operation creates two distinct copies of s').

Assume next that the assumptions of Lemma 5.3 do not hold, which means that both π_1 and π_2 intersect $\mathcal{C}$, and that $\mathcal{C}(\pi_1) = \mathcal{C}(\pi_2)$, which is either a contact interval I or a transparent edge f of $\mathcal{C}$ (see Fig. 41(d)). In the former case (a contact interval), the wave of s_i is not part of any sub-wavefront of W that leaves c (as shortest paths), and it should not be involved in any further critical event inside c, as discussed above. To ignore s_i in the further simulation of the propagation of W in $T_B(e)$, we reset $priority(s_i) := +\infty$ (instead of deleting s_i from W, which would involve an unnecessary recomputation of the bisectors involving the neighbors of s_i). In the latter case, the following similar technical operation must be performed. Since s_i is a part of the resulting wavefront $W(e, f)$ (as will follow from the correctness of the bisector event processing, proved in Lemma 5.9 below), we do not want to delete s_i from W; yet, since s_i is not involved in any further critical event inside c, we want to ignore s_i in the further simulation of the propagation of W in $T_B(e)$ (that is, to ignore its priority in the priority queue), and therefore we update $priority(s_i) := +\infty$. However, this artificial setting must be corrected later, when the propagation of W in $T_B(e)$ is finished, to ensure that the priority of s_i in $W(e, f)$ is correctly set—we must then reset $priority(s_i)$ to its true (current) value. We *mark* s_i to remember that its

priority must be reset later, and keep a list of pointers to all the currently marked generators; when their priorities must be reset, we go over the list, fixing each generator and removing it from the list).

To summarize, in Case (i) we trace two paths and perform one DELETE operation and $O(1)$ priority queue operations, hence it takes $O(\log n)$ time to process a true bisector event. In Case (ii) we trace $O(1)$ paths, compute at most $O(1)$ unfolded images, and perform at most one SPLIT operation and $O(1)$ priority queue operations; hence it takes $O(\log n)$ time to process a false (candidate) bisector event. The correctness of the above procedure is established in Lemma 5.9 below, but first we describe the detection and the processing of the candidate vertex events that were not detected and processed during the handling of false candidate bisector events. This situation arises when the priority of the next event of W in the priority queue is at least $t_{\text{stop}}(W)$, in which case we stop processing the bisector events of W in $T_B(e)$, and proceed as described next.

Processing a Covered Segment of $\mathcal{C}$ Consider the situation in which the algorithm stops propagating W in $T_B(e)$ at simulation time $t_{\text{stop}}(W) \neq +\infty$. We then must have $t_{\text{stop}}(W) = t_{\text{stop}}(f)$, for some segment f in $range(W)$, so that all the currently known candidate events which occur in c and involve the sub-wavefront of W that claims f have already been processed.

Another case in which the algorithm stops the propagation of W is when $t_{\text{stop}}(W) = +\infty$. This means that all the currently known candidate events of W have already been processed; that is, the former situation holds for each segment f' in $range(W)$. Therefore, to treat the latter case, we process each f' in $range(W)$ in the same manner as we process the (only relevant) segment f in the former case; and so, we consider only the former situation.

Let f be such a segment of $range(W)$. We compute the static wavefront $W(e, f)$ from the current dynamic wavefront W—if f is a transparent edge, then $W(e, f)$ is needed for the propagation process in further cells; otherwise (f is a contact interval) we do not need to compute $W(e, f)$ to propagate it further, but we need to know the extreme generators of $W(e, f)$ to ensure correctness of the simulation process, a step that will be explained in the proof of Lemma 5.9 below. Since the computation in the latter case is almost identical to the former, we treat both cases similarly (up to a single difference that is detailed below).

Since $f \in \mathcal{C}$ defines a unique homotopy class of paths from e_B to f within $T_B(e)$, the sub-wavefront of W that claims points of f is indeed a single contiguous sub-wavefront $W' \subseteq W$. We determine the *candidate* extreme claimers of f by performing a SEARCH in W for each of the endpoints a, b of f (note that the candidates are not necessarily true, since SEARCH does not consider visibility constraints). If the candidate claimer of a does not exist, we denote by a' the point of f closest to a which is intersected by an extreme bisector of W—see Fig. 43(a). (If there is no such a', we can already determine that W claims no points on f, and no further processing of f is needed—see Fig. 43(c).) Symmetrically, we SEARCH for the claimer of b, and, if it is not found, we define b' similarly. If a (resp., b) is claimed by W, denote by π_1 (resp., π_2) the path $\pi(claimer(a), a)$ (resp., $\pi(claimer(b), b)$); otherwise denote by π_1 (resp., π_2) the path $\pi(claimer(a'), a')$ (resp., $\pi(claimer(b'), b')$). Denote by W' the sub-wavefront of W between the generators of π_1 and π_2 (inclusive), and use

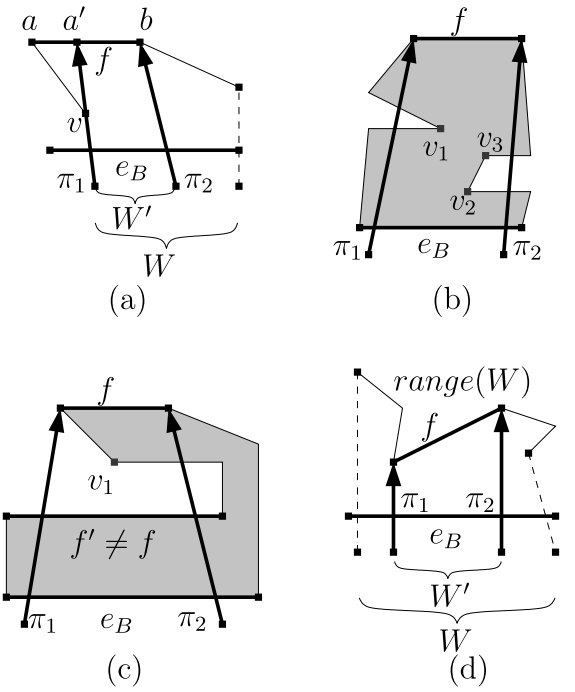


Fig. 43 Processing a covered segment f of $range(W)$. (**a**) The endpoint a of f is not claimed by W, and π_1 is the shortest path to the point a' closest to a and claimed by W; the generator of π_1 is extreme in W (which has already been split at v). (**b**) At least one vertex of $\mathcal{V} = \{v_1, v_2, v_3\}$ (namely, v_2) is visible from the entire portion of e_B between π_1 and π_2. (**c**) f is not reached by W at all. No vertex of $\mathcal{V}$ is visible from the portion of e_B between π_1 and π_2. (**d**) Since $d_f = |\pi_1| < |\pi_2|$, W is first split at the generator of π_1

π_1, π_2 to define (and compute) $\mathcal{V}$ as in the processing of a candidate bisector event (described above).

Assume first that π_1, π_2 satisfy the assumptions of Lemma 5.3. It follows that $\mathcal{V}$ is not empty, and at least one vertex of $\mathcal{V}$ is visible from its claimer in W' (see, e.g., Fig. 43(b)). Then the case is processed as Case (ii) of a candidate bisector event, with the following difference: Instead of tracing a path from each source image in W' to each vertex $v \in \mathcal{V}$ (which is too expensive now, since W' may have non-constant size), we first SEARCH in W' for the claimer of each such v and then trace only the paths $\pi(claimer(v), v)$. (Then we restart the simulation from the earliest time when a vertex v of $\mathcal{V}$ is reached by W, splitting W at $claimer(v)$.)

Assume next that the assumptions of Lemma 5.3 do not hold, which means that both π_1 and π_2 intersect $\mathcal{C}$, and that $\mathcal{C}(\pi_1) = \mathcal{C}(\pi_2)$, which is either f or a segment $f' \neq f$ of $\mathcal{C}$. In the latter case, since f is not reached by W at all, no further processing of f is needed (see Fig. 43(c))—we *ignore* f in the rest of the present simulation, and update $t_{\text{stop}}(W) := \min\{t_{\text{stop}}(f') | f' \in range(W) \setminus \{f\}\}$. In the former case, if both π_1, π_2 are extreme in W, then we have $W' = W$; the further processing of f is described below. Otherwise (at least one of π_1, π_2 is not extreme in W), we first have to split W, as follows. If π_1 and π_2 are not extreme in W, denote by d_f the minimum of $|\pi_1|, |\pi_2|$; if only one path $\pi \in \{\pi_1, \pi_2\}$ is non-extreme in W, let $d_f := |\pi|$. Without loss of generality, assume that $d_f = |\pi_1|$ (see Fig. 43(d)). We restart the simulation from time $|\pi_1|$, splitting W at the generator of π_1, as described in Case (ii) of the processing of a candidate bisector event.

It is only left to describe the case where $W' = W$ and f is the only (not *ignored*) segment of $range(W)$. If f is a contact interval, no further processing of f is needed.

Otherwise (f is a transparent edge), we have to make the following final updates (to prepare $W(e, f)$ for the subsequent merging procedure at f and for further propagation into other cells). First, we recalculate the priority of each *marked* source image (recall that it was temporarily set to $+\infty$), and update the priority queue component of the data structure accordingly. Next, we update the source unfolding data (and $\Lambda(W)$), as follows. Let $\mathcal{B}$ be the block sequence traversed by W from e to f along $T_B(e)$, including (resp., excluding) B if the first (resp., last) facet of B lies on $\Lambda(W)$, and let $\mathcal{E}$ be the edge sequence associated with B. We compute the unfolding transformation $U_{\mathcal{E}}$, by composing the unfolding transformations of the $O(1)$ blocks of $\mathcal{B}$. We update the data structure of $W(e, f)$ to add $U_{\mathcal{E}}$ to the unfolding data of all the source images in $W(e, f)$, as described in Sect. 5.1. As a result, for each generator s_i of $W(e, f)$, the polytope edge sequence $\mathcal{E}_i$ is the concatenation of its previous value with $\mathcal{E}$, and all the generators in $W(e, f)$ are unfolded to the plane of an extreme facet incident to f.

To summarize, we trace $O(1)$ paths and perform at most $O(1)$ SPLIT and SEARCH operations, for each of $O(1)$ segments of $\mathcal{C}$. Then we perform at most one source unfolding data update for each transparent edge in $\mathcal{C}$. All these operations take a total of $O(\log n)$ time. However, we also perform a single priority update operation for each marked generator that has participated in a candidate bisector event beyond a transparent edge of $\mathcal{C}$. A linear upper bound on the total number of these generators, as well as the number of the processed candidate events, is established next.

Correctness and Complexity Analysis We start by observing, in the following lemma, a basic property of W that asserts that distances from generators *increase* along their bisectors.

Lemma 5.4 *Let $s_i, s_j \in W$ be a pair of generators that become neighbors at a bisector event x during the propagation of W through $T_B(e)$, where an intermediate generator s' gets eliminated. Then* (i) *the portion of the bisector $b(s_i, s_j)$ that is closer to s' than x is claimed, among s_i, s' and s_j, by s', and* (ii) *the distances from s_i and s_j to points y on the portion of $b(s_i, s_j)$ that is not claimed by s', increase as y moves away from x.*

Proof In the plane $\Lambda(W)$, consider the Voronoi diagram of the three sites s_i, s', s_j, whose sole vertex is x. The line containing e intersects exactly two Voronoi edges, because it meets the Voronoi cells $V(s_i)$, $V(s')$, $V(s_j)$ of all the three sites. Moreover, by assumption, $e \cap V(s')$ lies between $e \cap V(s_i)$ and $e \cap V(s_j)$. Hence, the Voronoi edge that e misses is between $V(s_i)$ and $V(s_j)$, implying that $b(s_i, s_j)$, between e and x, is fully contained in $V(s')$, as asserted—see Fig. 44(a). The same argument also implies (ii). □

Lemma 5.5 *Assume that all bisector events of W that have occurred up to some time t have been correctly processed, and that the data structure of W has been correctly updated. Let p be a point tentatively claimed by a generator $s_i \in W$ at time $d(s_i, p) \leq t$, meaning that the claim is only with respect to the current generators in W (at time t), and that we ignore any visibility constraints of $\mathcal{C}$. Denote by $R(s_i)$*

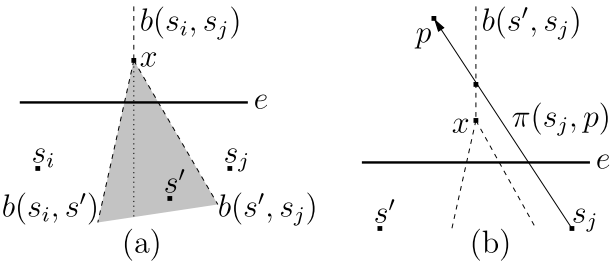


Fig. 44 (**a**) The bisector $b(s_i, s_j)$, between e and x, must be fully contained in $V(s')$ (*shaded*). (**b**) If the bisector $b(s', s_j)$ is already computed in the wavefront W, then the path $\pi(s_j, p)$, which intersects $b(s', s_j)$, cannot be encoded in W

the unfolded region that is enclosed between the bisectors of s_i currently stored in the data structure. Then $p \in R(s_i)$, and $p \notin R(s_j)$, for any other generator $s_j \neq s_i$ in W.

Proof The claim that $p \in R(s_i)$ is trivial, since the bisectors of s_i that are currently stored in the data structure have been computed before time t, and are therefore correct, by assumption; hence, p is enclosed between them.

For the second claim, assume to the contrary that there exists a generator $s_j \neq s_i$ in W so that $p \in R(s_j)$ too. Denote by q the first point along $\pi(s_j, p)$ that is equally close to s_j and to some other generator $s' \in W$ (such q and s' must exist, since $d(s_i, p) < d(s_j, p)$); that is, $q = \pi(s_j, p) \cap b(s', s_j)$. The fact that in the data structure p lies in $R(s_j)$ means that the bisector $b(s', s_j)$ is not correctly stored in the data structure, and thus it cannot be part of $W(e_B)$; therefore $b(s', s_j)$ emanates from a bisector event location x that lies within c—see Fig. 44(b). By Lemma 5.4, $d(s', x) < d(s', q) < d(s', p) \leq t$; hence, the bisector event when $b(s', s_j)$ is computed occurs before time t, and therefore, by assumption, $b(s', s_j)$ is correctly stored in the data structure—a contradiction. □

In particular, Lemma 5.5 shows that when a vertex event at v is discovered during the processing of another event at simulation time t, or is processed when a segment of $\mathcal{C}$ that is incident to v is covered at time t, the tentative claimer of v (among all the current generators in W) is correctly computed, assuming that all bisector events of W that have occurred up to time t have been correctly processed. We will use this argument in Lemma 5.9 below.

Lemma 5.6 *Assume that all bisector events of W that have occurred up to some time t have been correctly processed, and that the data structure of W has been correctly updated at all these events. If two waves of a common topologically constrained portion of W are adjacent at t, then their generators must be adjacent in the generator list of W at simulation time t.*

Proof Assume the contrary. Then there must be two source images s_i, s_j in a common topologically constrained portion $W' \subseteq W$ such that their respective waves w_i, w_j are adjacent at some point x at time t (that is, $d(s_i, x) = d(s_j, x) = t \leq d(s_k, x)$ for all other generators s_k in W), but there is a positive number of source images $s_{i+1}, \ldots, s_{j-1}$ in the generator list of W' at time t between s_i and s_j, whose distances to x are necessarily larger than $d(s_i, x)$ (and their waves in W' at time t are nontrivial arcs). See Fig. 45 for an illustration.

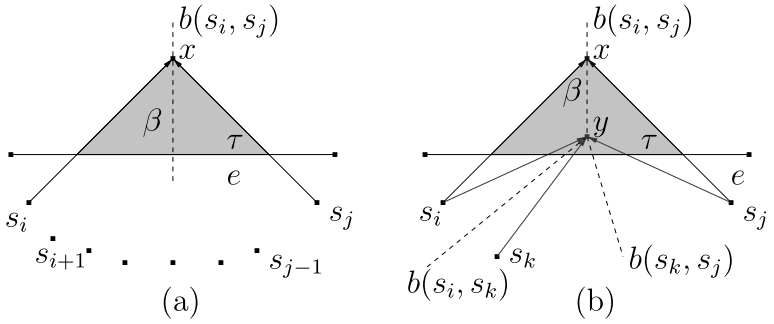


Fig. 45 The waves from the source images s_i, s_j collide at x. Each of the two following cases contradicts the assumption in the proof of Lemma 5.6: (**a**) The portion β of $b(s_i, s_j)$ intersects the transparent edge e; (**b**) The generator s_k is eliminated at time $t_y = d(s_i, y) < d(s_i, x) = t$

Consider the situation at time t. Since w_i, w_j belong to a common topologically constrained W', it follows that e, $\pi(s_i, x)$ and $\pi(s_j, x)$ unfold to form a triangle τ in an unfolded block sequence of $T_B(e)$ (so that τ is not intersected by $\mathcal{C}$).

Consider the "unfolded" Voronoi diagram $\mathrm{Vor}(\{s_i, \ldots, s_j\})$ within τ. By assumption, x lies in the Voronoi cells $V(s_i)$, $V(s_j)$ of s_i, s_j, respectively, separated by a Voronoi edge β, which is a portion of $b(s_i, s_j)$. If β intersects e (see Fig. 45(a)), then s_i and s_j claim consecutive portions of e in $W(e)$, so s_i and s_j must be consecutive in W already at the beginning of its propagation within $T_B(e)$, a contradiction.

Otherwise, β ends at a Voronoi vertex y within τ—see Fig. 45(b). Clearly, y is the location of a bisector event in which some generator $s_k \in W$ is eliminated at time $t_y = d(s_i, y) = d(s_j, y)$. By Lemma 5.4, $t_y < t$, and therefore, by our assumption, the bisector event at y has been correctly processed, so s_i and s_j must be consecutive in W already before time t—a contradiction. □

Lemma 5.6 shows that if all the events considered by the algorithm are processed correctly, then all the true bisector events of the first kind are processed by the algorithm, since, as the lemma shows, such events occur only between generators of W that are consecutive at the time the bisector event occurs. Let W' be a topologically constrained portion of W, and denote by $R(W', t)$ the region within $T_B(e)$ that is covered by W' from the beginning of the simulation in $T_B(e)$ up to time t. By definition of topologically constrained wavefronts, $\partial R(W', t)$ consists only of e_B and of the unfolded images of the waves and of the extreme bisectors of W' at time t. Another role of Lemma 5.6 is in the proof of the following observation.

Corollary 5.7 *$R(W', t)$ is not punctured (by points that are not covered by W' at time t).*

Proof Consider the first time at which $R(W', t)$ becomes punctured. When this happens, $R(W', t)$ must contain a point q where a pair of waves, generated by the respective generators s_i, s_j, collide, and e_B and the paths $\pi(s_i, q), \pi(s_j, q)$ enclose an island within the (unfolded) triangle that they form. This however contradicts the proof of Lemma 5.6. □

Corollary 5.8 *When a vertex event at v is discovered during the processing of a candidate event at simulation time t (either a bisector event x or an event involving a covered segment f of $\mathcal{C}$), the vertex v is reached by W no later than time t.*

Proof By the way vertex events are discovered, v must lie in an unfolded triangle τ formed as in the proof of Lemma 5.6, where the waves of the respective generators s_i, s_j either collide at x, or are adjacent in the wavefront that covers the segment f. Since the two sides of τ incident to x belong to $R(W', t)$, for some topologically constrained portion W' of W that contains s_i, s_j, Corollary 5.7 implies that all of τ is contained in $R(W', t)$, which implies the claim. □

We are now ready to establish the correctness of the simulation algorithm. Since this is the last remaining piece of the inductive proof of the whole Dijkstra-style propagation (Lemmas 4.2 and 4.5), we may assume that all the wavefronts were correctly propagated to some transparent edge e, and consider the step of propagating from e. This implies that $W(e_B)$ encodes all the shortest paths from s to the points of e_B from one fixed side. Now, let $x_1, \ldots, x_m$ be all the *true critical events* (that is, both bisector and vertex events that are true with respect to the propagation of W in $T_B(e)$), ordered according to the times $t_1, \ldots, t_m$ at which the locations of these events are first reached by W. Since we assume general position, $t_1 < \cdots < t_m$.

Before we show the correctness of the processing of the true critical events, let us discuss the processing of the false candidates. First, note that the simulation can be aborted at time t' (during the processing of a false candidate event) and restarted from an earlier time $t'' < t'$ only if there exists some true vertex event x that should occur at time $t \leq t''$ and has not been detected prior to time t' (in the aborted version of the simulation). Note that whenever a false candidate event $x' \notin \{x_1, \ldots, x_m\}$ is processed at time t', one of the three following situations must arise.

(i) It might be that x' is not currently (at time t') determined to be false, since both paths involved in x' are traced along the same block sequence and do not intersect $\mathcal{C}$; x' is false "just" because there is some earlier true vertex event x'' that is still undetected. In this case, we create a new version of W at time t', but it will later be declared invalid, when we finally detect x''.

Otherwise, x' is immediately determined to be false (since either one of the involved paths intersects $\mathcal{C}$ or the paths are traced along different block sequences). In this case either (ii) an earlier candidate vertex event x'' (occurring at some time $t'' < t'$) is currently detected and the simulation is restarted from t'', or (iii) x' is a bisector event which occurs outside $T_B(e)$, so it involves only bisectors that do not participate in any further critical event inside $T_B(e)$. In this case a new version of W, corresponding to the time t', is created, the generator that is eliminated at x' is marked in it, and its priority is set to $+\infty$.

In any of the above cases, none of the existing true (valid) versions of W is altered (although some invalid versions may be discarded during a restart); moreover, a new invalid version corresponding to time t' may be created (without restarting the simulation yet) only if there is some true event that occurred at time $t < t'$ but is still undiscovered at time t'.

Assume now that at the simulation time t_k (for $1 \leq k \leq m$) all the true events that occur before time t_k have been correctly processed; that is, for each such bisector

event x_i, the corresponding generator has been eliminated from W at simulation time t_i, and for each such vertex event x_j, W has been split at simulation time t_j at the generator that claims the corresponding vertex. Note that the assumption is true for simulation time t_1, since the processing of false candidate events does not alter $W(e_B)$ (which does not encode events within $T_B(e)$; its validity follows from the inductive correctness of the merging procedure and is not violated by the processing of false events).

Lemma 5.9 *Assuming the above inductive hypothesis, the next true critical event x_k is correctly processed at simulation time t_k, possibly after a constant number of times that the simulation clock has reached and passed t_k (to process a later false candidate event) without detecting x_k, each time resulting in a simulation restart.*

Proof There are two possible cases. In the first case, x_k is a true bisector event, in which the wave of a generator s' in W is eliminated by its neighbors at propagation time t_k. Any possible false candidate event that is processed before x_k and after the processing of all true events that take place before time t_k may only create new invalid versions that correspond to times that are later than time t_k (since a false candidate event can arise only when an earlier true event is still undetected). This implies that s' has not been deleted from any valid version of W that corresponds to time t_k or earlier, and all such valid versions exist. By this fact and by the inductive hypothesis, the bisectors of s' have been computed correctly either already in $W(e_B)$, or during the processing of critical events that took place before time t_k.

In the second case, x_k is a true vertex event that takes place at a vertex $v \in \mathcal{C}$, which is claimed by some generator s_v in W. By the argument used in the first case, s_v has not been deleted from W at an earlier (than t_k) simulation time, and each point on the path $\pi(s_v, v)$ is claimed by s_v at time t_k or earlier. Therefore s_v can only be deleted from a version of W at time later than t_k when a false bisector event involving s_v is processed. Moreover, a sub-wavefront including s_v can be split from a version of W at time later than t_k (and v can be removed from $range(W)$) when a false vertex event is processed. We show next that in both cases, x_k is detected and the simulation is restarted from time t_k, causing x_k to be processed correctly.

Consider first the case where s_v is not deleted in any later false candidate event. In that case, when we stop the propagation of W, v is in $range(W)$, and therefore at least one segment f of the segments of $range(W)$ that is incident to v is ascertained to be covered at that time. Since s_v is in W, Lemma 5.5 implies that the SEARCH procedure that the algorithm uses to compute the claimer of v outputs s_v, and, by Corollary 5.2, the tracing procedure correctly computes $d(s_v, v)$ to be t_k. Since x_k is the next true vertex event, the distance from the other endpoint of f to its claimer is larger than or equal to t_k, and, since W has not yet been split at v, $\pi(s_v, v)$ is not an extreme bisector of W. Hence the algorithm sets $d_f := t_k$, and W is split at s_v at time t_k, as required.

Consider next the case where s_v is deleted (or split) from W at a false event x' at time $t' \geq t_k$. Suppose first that x' is a false bisector event. Then v must lie in the interior of the region τ bounded by e and by the paths to the location of x' from the outermost generators of W involved in x'. The algorithm traces the paths to v and to (some of) the other vertices of $\mathcal{C}$ in τ from all the generators of W that are involved

in x', including s_v (see Fig. 41(b), (c)); then all such distances are compared. Only distances from each such generator s' to each vertex that is visible from s' (within the unfolded blocks of $T_B(e)$) are taken into account, since, by Corollary 5.2, all visibility constraints are detected by the tracing procedure. The vertex v must be visible from s_v and the distance $d(s_v, v)$ must be the shortest among all compared distances, since, by the inductive hypothesis, all vertex events that are earlier than x_k have already been processed (and W has already been split at these events). By Lemma 5.5 and by Corollary 5.2, the tentative claimer (among all current generators in W) of each vertex u is computed correctly. No generator s' that has already been eliminated from W can be closer to u than the computed $claimer(u)$, since, by Corollary 5.8 and by the inductive hypothesis, u would have been detected as a vertex event no later than the bisector event of s', which is assumed to have been correctly processed. Therefore the distance $d(claimer(u), u)$ is correctly computed for each such vertex u (including v), and therefore the distance $d(s_v, v) = t_k$ is determined to be the shortest among all such distances. Hence the simulation is restarted from time t_k, and W is split at s_v at simulation time t_k, as asserted.

Otherwise, x' is a false vertex event processed when a segment f of $\mathcal{C}$ is ascertained to be fully covered by W, and v must lie in the interior of the region τ bounded by e, f, and by the paths from the outermost generators of W claiming f to the extreme points of f that are tentatively claimed by W (see Fig. 43(b)). The algorithm performs the SEARCH operation in the sub-wavefront $W' \subseteq W$ that claims f for v and for all the other vertices of $\mathcal{C}$ in τ, and then compares the distances $d(claimer(u), u)$, for each such vertex u that is visible from its claimer (including v). By the same arguments as in the previous case, the distance $d(s_v, v) = t_k$ is determined to be the shortest among all such distances, the simulation is restarted from time t_k, and W is split at s_v at simulation time t_k, as asserted. □

The above lemma completes the proof of the correctness of our algorithm. Now we show that the total number of the processed candidate events is only linear. Order the $O(1)$ vertices of $\mathcal{C}$ that are reached by W (that is, the locations of the true vertex events) as $v_1, \ldots, v_m$, where W reaches v_1 first, then v_2, and so on; denote by t_j, for $1 \leq j \leq m$, the time at which W reaches v_j. Note that if the simulation is restarted because of a vertex event at v_j, then, by Lemma 5.9, the simulation is restarted exactly from time t_j—that is, t_j depends only on W and on the previous true candidate events. Note also that the simulation is only restarted from times $t_1, \ldots, t_m$.

Lemma 5.10 *When the vertex events at vertices $v_1, \ldots, v_k$, for $1 \leq k \leq m$, are already detected and processed by the algorithm, the simulation is never restarted from time t_k or earlier.*

Proof Since the simulation restart from time t discards all existing versions of W that correspond to times $t' \geq t$, the claim of the lemma is equivalent to the claim that all the versions of W that were created at time t_k or earlier will never be discarded by the algorithm if all the vertex events at vertices $v_1, \ldots, v_k$ have already been detected and processed. We prove the latter claim by induction on k.

For $k = 1$, the version of W created at time t_1 can only be discarded if a vertex event that occurs earlier than t_1 is discovered, which is impossible since v_1 is the

first vertex reached by W. Now assume that the claim is true for $v_1, \ldots, v_{k-1}$, and consider the version W_k of W that is created at time t_k when the vertex events at vertices $v_1, \ldots, v_k$ are already detected and processed. The algorithm may discard W_k only when at some time $t' > t_k$ a vertex v is discovered, such that v is reached by W at time $t_v < t_k$, and therefore v must be in $\{v_1, \ldots, v_{k-1}\}$. But then, restarting the propagation from time t_v contradicts the induction hypothesis. □

Lemma 5.11 *For each $1 \leq j \leq m$, the simulation is restarted from time t_j at most 2^{j-1} times.*

Proof By induction on j. By Lemma 5.10, the simulation is restarted from time t_1 at most once. Now assume that $j \geq 2$ and that the claim is true for times $t_1, \ldots, t_{j-1}$.

By Lemma 5.9, the vertex event at v_j is eventually processed at time t_j; by Lemma 5.10, there are no further restarts from time t_j after we get a version of W that encodes all the events at $v_1, \ldots, v_j$. Hence the simulation may be restarted from time t_j only once each time that W ceases to encode the vertex event at v_j, and this may only happen either at the beginning of the simulation, or when the simulation is restarted from a time earlier than t_j. Since, by the induction hypothesis, the simulation is restarted from times $t_1, \ldots, t_{j-1}$ at most $\sum_{i=1}^{j-1} 2^{i-1} = 2^{j-1} - 1$ times, the simulation may be restarted from time t_j at most 2^{j-1} times. □

Remark From a practical point of view, the algorithm can be significantly optimized, by using the information computed before the restart to speed up the simulation after it is restarted. Moreover, we suspect that, in practice, the number of restarts that the algorithm will perform will be very small, significantly smaller than the bounds in the lemma.

By Lemma 5.11, the algorithm processes only $O(1)$ candidate vertex events (within a fixed $T_B(e)$), and, since the simulation is restarted only at a vertex event, it follows that each bisector event x has at most $O(1)$ "identical copies," which are the same event, processed at the same location (and at the same simulation time t_x) after different simulation restarts. At most one of these copies of x remains encoded in valid versions of W, and the rest are discarded (that is, there is at most one valid version of W that has been created at simulation time t_x to reflect the correct processing of x, and the following valid versions of W are coherent with this version). Hence for the purpose of further asymptotic time complexity analysis, it suffices to bound the number of the processed candidate bisector events that take place at distinct locations.

Note that each candidate bisector event x processed by the propagation algorithm falls into one of the five following types:

(i) x is a true bisector event.
(ii) x is a false candidate bisector event, during the processing of which an earlier-reached vertex of $\mathcal{C}$ has been discovered, and the simulation has been restarted.
(iii) x is a false candidate bisector event of a generator $s' \in W$, so that all paths in the wave from s' cross a common contact interval of $\mathcal{C}$ (a "dead-end") before the wave is eliminated at x.

(iv) x is a false candidate bisector event of a generator $s' \in W$, so that all paths in the wave from s' cross a common transparent edge f of $\mathcal{C}$ before the wave is eliminated at x, and the distance from f to x along $d(s', x)$ is greater than $2|f|$.
(v) x is a false candidate bisector event, as in (iv), except that the distance from f to x along $d(s', x)$ is at most $2|f|$.

Lemma 5.12 *The total number of processed true bisector events* (*events of type* (*i*)), *during the whole wavefront propagation phase*, *is* $O(n)$.

Proof First we bound the total number of waves that are created by the algorithm. The wavefront W is always propagated from some transparent edge e, within the blocks of a tree $T_B(e)$, for some block B incident to e, in the Riemann structure $\mathcal{T}(e)$ of e. A wave of W is split during the propagation only when W reaches a vertex of $\mathcal{C}$, the corresponding boundary chain of $T_B(e)$. Each such vertex is reached at most once (ignoring restarts) by each topologically constrained wavefront that is propagated in $T_B(e)$. There are only $O(1)$ such wavefronts, since there are only $O(1)$ paths in $T_B(e)$ (and corresponding homotopy classes). Each (side of a) transparent edge e is processed exactly once (as the starting edge of the propagation within $R(e)$), by Lemma 4.2, and e may belong to at most $O(1)$ well-covering regions of other transparent edges that may use e at an intermediate step of their propagation procedures. There are $O(1)$ vertices in any boundary chain $\mathcal{C}$, hence at most $O(1)$ wavefront splits can occur within $T_B(e)$ during the propagation of a single wavefront. Since there are only $O(n)$ transparent edges e in the surface subdivision, and there are only $O(1)$ trees $T_B(e)$ for each e, we process at most $O(n)$ such split events. (Recall from Lemma 5.11 that a split at a vertex is processed at most $O(1)$ times.) Since a new wave is added to the wavefront only when a split occurs, at most $O(n)$ waves are created and propagated by the algorithm.

In each true bisector event processed by our algorithm, an existing wave is eliminated (by its two adjacent waves). Since a wave can be eliminated exactly once and only after it was earlier added to the wavefront, we process at most $O(n)$ true bisector events. □

Lemma 5.13 *The algorithm processes only* $O(n)$ *candidate bisector events during the whole wavefront propagation phase*.

Proof There are at most $O(n)$ events of type (i) during the whole algorithm, by Lemma 5.12. By Lemma 5.11, there are only $O(1)$ candidate events of type (ii) that arise during the propagation of W in any single block tree $T_B(e)$. Since a candidate event of type (iii), within a fixed surface cell c, involves at least one wave that encodes paths that enter c through e_B but never leave c (that is, they traverse a facet sequence that contains a loop, and are therefore known not to be shortest paths beyond some contact interval in the loop), the total number of these candidate events during the whole propagation is bounded by the total number of generated waves, which is $O(n)$ by the proof of Lemma 5.12.

Consider a candidate event of type (iv) at a location x at time t_x, in some fixed $T_B(e)$, and let f denote the transparent edge of $\mathcal{C}$ that is crossed by the wave from the generator s' eliminated at x. Denote by d_1 (resp., d_2) the distance along $\pi(s', x)$ from

s' to f (resp., from f to x); that is, $d_2 > 2|f|$ and $d_1 + d_2 = d(s', x) = t_x$. Before the update of $t_{\text{stop}}(f)$, caused by the processing of this event, the value of $t_{\text{stop}}(f)$ must have been equal to or greater than $t_x > d_1 + 2|f|$, since otherwise f would have been ascertained to be covered before time t_x, and therefore the event at t_x would not have been processed; hence, after the update, we have $t_{\text{stop}}(f) = d_1 + |f| < t_x$. Therefore, immediately after the processing of the event at t_x we detect that f has been covered; by Lemma 5.11 each f is detected to be covered at most $O(1)$ times, and, since there are only $O(1)$ transparent edges in $\mathcal{C}$, there are at most $O(1)$ events of type (iv) during the propagation of W in $T_B(e)$.

Consider now a candidate event of type (v) that occurs at a location x at time t_x after crossing the transparent edge f of $\mathcal{C}$. This event may also be detected during the propagation of the wavefront through f into further cells, and therefore it must be counted more than once during the whole wavefront propagation phase. However, on $\Lambda(W)$, x lies no further than $2|f|$ from the image of f, and therefore the shortest-path distance from f to the location of x on ∂P cannot be greater than $2|f|$; hence, by the well-covering property of f, x lies within $k = O(1)$ cells away from the cell c. Therefore the event at x is detected during the simulation in the cell which contains x, where the event at x is considered a *true event*, and during the simulation in at most k other cells; hence, by Lemma 5.12, the total number of these candidate events during the whole algorithm is bounded by $O(kn) = O(n)$. □

We summarize the main result of the preceding discussion in the following lemma.

Lemma 5.14 *The total number of candidate events processed during the wavefront propagation is $O(n)$. The wavefront propagation phase of the algorithm takes a total of $O(n \log n)$ time and space.*

5.4 Shortest Path Queries

Preprocessing Building Blocks Let B be a building block of a surface cell c. A generator of a wavefront W is called *active in B* if it was detected by the algorithm to be involved in a bisector event inside B. The wavefront propagation algorithm lets us compute the active generators for all pairs (W, B) in a total of $O(n \log n)$ time.

We next define the partition $local(W, B)$ of the unfolded portion of B that was covered by W (and the wavefronts that W has been split into during its propagation within B), which will be further preprocessed for point location for shortest path queries.[14] The partition $local(W, B)$ consists of *active* and *inactive regions*, defined as follows. The active regions are those portions of B that are claimed by generators of W that are active in B, and each inactive region is claimed by a contiguous band of waves of W that cross B in an "uneventful" manner, delimited by a sequence of pairwise disjoint bisectors. See Fig. 46 for an illustration.

[14]Note that if W has been split in another preceding building block of c into two sub-wavefronts W_1, W_2 that now traverse B as two distinct topologically constrained wavefronts, no interaction between W_1 and W_2 in B is detected or processed (the two traversals are processed at two distinct nodes of a block tree, or of different block trees of $\mathcal{T}(e)$, both representing B). Moreover, if W has been split in B (which might happen if B is a nonconvex type I block—see Sect. 3.1), the split portions cannot collide with each other inside B; see Fig. 46.

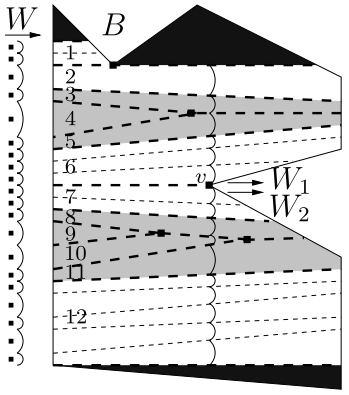


Fig. 46 The wavefront W enters the building block B (in this example, B is a nonconvex block of type I, bounded by *solid lines*) from the left. The partition $local(W, B)$ is drawn by thick *dashed lines*; *thin dashed lines* denote bisectors of W that lie fully in the interior of the inactive regions. The regions of the partition are numbered from 1 to 12; the active regions are *lightly shaded*, the inactive regions are *white*, and the portions of B that were not traversed by W due to visibility constraints are *darkly shaded*. The locations of the bisector events of W and the reflex vertices reached by W in B are marked. W is split at v into W_1 and W_2, and $local(W, B)$ includes these sub-wavefronts too

Here are several comments concerning this definition. The edges of $local(W, B)$ are those bisectors of pairs of generators of W, at least one of which is active in B. The first and the last bisectors of W are also defined to be edges of $local(W, B)$. If, during the propagation in B, W has been split (into sub-wavefronts W_1, W_2) at a reflex vertex v of B, then the ray from the generator of W, whose wave has been split at v, through v (an artificial extreme bisector of both W_1, W_2) is also defined to be an edge of $local(W, B)$. If W has been split into sub-wavefronts W_1, W_2 in such a way, we treat also the bisectors of W_1, W_2, within B, as if they belonged to W (that is, embed $local(W_1, B)$, $local(W_2, B)$ as extensions of $local(W, B)$, and preprocess them together as a single partition of B).

Note that the complexity of $local(W, B)$ is $O(k + 1)$, where k is the number of true critical (bisector and vertex) events of W in B. The partition can actually be computed "on the fly" during the propagation of W in B, in additional $O(k)$ time.

We preprocess each such partition $local(W, B)$ for point location [13, 22], so that, given a query point $p \in B$, we can determine which region r of $local(W, B)$ contains the unfolded image q of p (that is, if B is of type II or III and $\mathcal{E}$ is the edge sequence associated with B, $q = U_{\mathcal{E}}(p)$; if B is of type I or IV then $q = p$). If r is traversed by a single wave of W (which is always the case when r is active, and can also occur when r is inactive), it uniquely defines the generator of W that claims p (if we ignore other wavefronts traversing B). This step of locating r takes $O(\log k)$ time. If q is in an inactive region r of $local(W, B)$ that was traversed by more than one wave of W, then r is the union of several "strips" delimited by bisectors between waves that were propagated through B without events. We can then SEARCH for the claimer of q in the portion of W corresponding to the inactive region, in $O(\log n)$ time (see Sect. 5.1).

Preprocessing S_{3D} In order to locate the cell of S that contains the query point, we also preprocess the 3D-subdivision S_{3D} for point location, as follows. First, we subdivide each perforated cube cell into six rectilinear boxes, by extending its inner horizontal faces until they reach its outer boundary, and then extending two parallel vertical inner faces until they reach the outer boundary too, in the region between the extended horizontal faces. Next, we preprocess the resulting 3-dimensional rectilinear subdivision in $O(n \log n)$ time for 3-dimensional point location [10]. The resulting data structure takes $O(n \log n)$ space, and a point location query takes $O(\log n)$ time.

Answering Shortest-Path Queries To answer a shortest-path query from s to a point $p \in \partial P$, we perform the following steps.

1. Query the data structure of the preprocessed S_{3D} to obtain the 3D-cell c_{3D} that contains p.
2. Query the surface unfolding data structure (defined in Sect. 2.4) to find the facet f of ∂P that contains p in its closure.
3. Since the transparent edges are close to, but not necessarily equal to, the corresponding intersections of subfaces of S_{3D} with ∂P, p may lie either in a surface cell induced by c_{3D} or by an adjacent 3D-cell, or in a surface cell derived from the intersection of transparent edges of $O(1)$ such cells. To find the surface cell containing p, let $I(c_{3D})$ be the set of the $O(1)$ surface cells induced by c_{3D} and by its $O(1)$ neighboring 3D-cells in S_{3D} (whose closures intersect that of c_{3D}). For each cell $c \in I(c_{3D})$, check whether $p \in c$, as follows.
   (a) Using the surface unfolding data structure, find the transparent edges of ∂c that intersect f, by finding, for each transparent edge e of ∂c, the polytope edge sequence $\mathcal{E}$ that e intersects, and searching for f in the corresponding facet sequence of $\mathcal{E}$ (see Sect. 2.4).
   (b) Calculate the portion $c \cap f$ and determine whether p lies in that portion.
   If p is contained in more than one surface cell, assign it to an arbitrary cell among them.
4. Among the $O(1)$ building blocks of c, find a block B that contains p. For each wavefront W that has traversed B, we find the generator s_i that claims p in W, using the point location structure of $local(W, B)$ as described above, and compute the distance $d(s_i, p)$. We report the minimal distance from s to p among all such claimers of p.
5. If the corresponding shortest path has to be reported too, we report all polytope edges that are intersected by the path from the corresponding source image to p. In case there are several such paths, each can be reported in the same manner.

Steps 1–3 take $O(\log n)$ time, using [10] and the data structure defined in Sect. 2.4. As argued above, it takes $O(\log n)$ time to perform Step 4. This concludes the proof of our main result (modulo the construction of the 3D-subdivision, given in the next section):

Theorem 5.15 (Main Result) *Let P be a convex polytope with n vertices. Given a source point $s \in \partial P$, we can construct an implicit representation of the shortest path map from s on ∂P in $O(n \log n)$ time and space. Using this structure, we can identify, and compute the length of, the shortest path from s to any query point $q \in \partial P$ in*

$O(\log n)$ time (in the real RAM model). A shortest path $\pi(s,q)$ can be computed in additional $O(k)$ time, where k is the number of straight edges in the path.

6 Constructing the 3D-Subdivision

This section briefly sketches the proof of Theorem 2.1, by describing an algorithm for constructing a conforming 3D-subdivision for a set V of n points in $\mathbb{R}^3$. Since this is a straightforward generalization of the construction of a similar conforming subdivision in the plane [18], we only describe the details that are different from those in [18], and provide a few necessary definitions.

The main part of the algorithm constructs a *1-conforming* 3D-subdivision S^1_{3D} of size $O(n)$ in $O(n \log n)$ time, which is then transformed into a conforming 3D-subdivision S_{3D} by subdividing each face of S^1_{3D} into 16×16 square subfaces, in $O(n)$ additional time.

Constructing the 1-Conforming 3D-Subdivision We fix a Cartesian coordinate system in $\mathbb{R}^3$. For any whole number i, the ith-order grid $\mathcal{G}_i$ in this system is the arrangement of all planes $x = k2^i$, $y = k2^i$ and $z = k2^i$, for $k \in \mathbb{Z}$. Each cell of $\mathcal{G}_i$ is a cube of size $2^i \times 2^i \times 2^i$, whose near-lower-left corner lies at a point $(k2^i, l2^i, m2^i)$, for a triple of integers k, l, m. We call each such cell an *i-box*. Any $4 \times 4 \times 4$ contiguous array of i-boxes is called an *i-quad*. Although an i-quad has the same size as an $(i+2)$-box, it is not necessarily an $(i+2)$-box because it need not be a cell in $\mathcal{G}_{i+2}$. The eight non-boundary i-boxes of an i-quad form its *core*, which is thus a $2 \times 2 \times 2$ array of i-boxes; see Fig. 47. Observe that an i-box b has exactly eight i-quads that contain b in their cores.

The algorithm constructs a conforming partition of the point set V in a bottom-up fashion. It simulates a growth process of a cube box around each data point, until their union becomes connected. The simulation works in discrete *stages*, numbered $-2, 0, 2, 4, \ldots$. It produces a subdivision of space into axis-parallel cells. The key object associated with a data point p at stage i is an i-quad containing p in its core. In fact, the following stronger condition holds inductively: Each $(i-2)$-quad constructed at stage $(i-2)$ lies in the core of some i-quad constructed at stage i.

In each stage i, only a minimal set $\mathcal{Q}(i)$ of quads is maintained. This set is partitioned into equivalence classes under the transitive closure of the *overlap* relation, where two i-quads overlap if they have a common i-box (not necessarily in their cores). The portion of space covered by quads in one class of this partition is called a *component*. Each component at stage i is either an i-quad or a connected union of (open) i-quads. We classify each component as being either *simple* or *complex*.

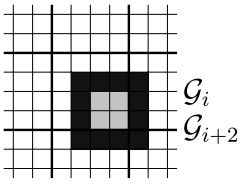


Fig. 47 The planar analog of an i-quad (*darkly shaded*) and its core (*lightly shaded*)

A component at stage i is *simple* if (1) its outer boundary is an i-quad and (2) it contains exactly one $(i-2)$-quad of $\mathcal{Q}(i-2)$ in its core. Otherwise, the component is *complex*.

The algorithm consists of two main parts. The first part grows the $(i-2)$-quads of stage $(i-2)$ into i-quads of stage i, and the other part computes and updates the equivalence classes, and constructs subdivision subfaces. These tasks are performed by the procedures *Growth* and *Build-subdivision*, respectively. We omit the description of *Growth* (which is a duplicate three-dimensional version of the same procedure in [18]), but briefly review some of its features, to facilitate the description of *Build-subdivision*.

Given an i-quad q, *Growth*(q) is an $(i+2)$-quad containing q in its core. For a *family* S of i-quads, *Growth*(S) is a *minimal* (but not necessarily the minimum) set of $(i+2)$-quads such that each i-quad in S is contained in the core of a member of *Growth*(S). Let *Growth*(q), or $\tilde{q}$, denote the unique $(i+2)$-quad returned by the procedure *Growth* with input q (see [18, 34] for details concerning the choice of $\tilde{q}$ among eight possible $(i+2)$-quads).

The Build-Subdivision *Procedure* By appropriate scaling and translation of 3-space, we may assume that the L_∞-distance between each pair of points in V is at least 1, and that no point coordinate is a multiple of $\frac{1}{16}$. For each point $p \in V$, we *construct* (to distinguish from other quads that we only *compute* during the process, constructing a quad means actually adding it to the 3D-subdivision) a (-4)-quad with p at the near-lower-left (-4)-box of its core; this choice ensures that the minimal distance from p to the boundary of its quad is at least $\frac{1}{4}$ of the side length of the quad. (This step is needed for the (MVC) property, and does not exist in [18].) Around each of these quads q, we compute (but *not* construct yet) a (-2)-quad with q in its core, so that when there is more than one choice to do that (there are one, two, four, or eight possibilities to choose the (-2)-quad if ∂q is coplanar with none, two, four, or six planes of $\mathcal{G}_{-2}$, respectively), we always choose the (-2)-quad whose position is extreme in the near-lower-left direction. This ensures that the (-2)-quads associated with distinct points are openly disjoint (because the points of V are at least 1 apart from each other in the L_∞-distance; without the last constraint, one could have chosen two (-2)-quads whose interiors have nonempty intersection). These quads form the set $\mathcal{Q}(-2)$, which is the initial set of quads in the *Build-subdivision* algorithm described below. Each quad in $\mathcal{Q}(-2)$ forms its own singleton component under the equivalence class in stage -2. As above, we regard all quads in $\mathcal{Q}(-2)$ as open, and thus forming distinct simple components, even though some pairs might share boundary points.

Let V_S be the set of points of V in the cores of the i-quads of a component $S \subseteq \mathcal{Q}(i)$. The implementation of *Build-subdivision* is based on the observation that the longest edge of the L_∞-minimum spanning tree of V_S has length less than $6 \cdot 2^i$. To make this observation more precise, we define $G(i)$ to be the graph on V containing exactly those edges whose L_∞ length is at most $6 \cdot 2^i$, and define MSF(i) to be the minimum spanning forest of $G(i)$.

The algorithm is based on an efficient construction of MSF(i) for all i such that MSF$(i) \neq$ MSF$(i-2)$. We first find all the $O(n)$ edges of the final MSF of V (a single tree), using the $O(n \log n)$ algorithm of Krznaric et al. [23] for computing an

L_∞-minimum spanning tree in three dimensions. (In the planar case of [18], the classical algorithm of Kruskal is used instead.) Then, for each edge e constructed by the algorithm, we compute the stage $k = 2\lceil \frac{1}{2} \log_2 \frac{1}{6}|e| \rceil$, at which e is added to MSF(k). By processing the edges in increasing length order, we obtain the entire sequence of forests MSF(i), for those i for which MSF(i) $\neq$ MSF($i-2$).

Only stages at which something happens are processed: MSF(i) changes, or there are complex components of $\mathcal{Q}(i)$ whose *Growth* computation is nontrivial. *Growth*(S) is only computed for complex components and for simple components that are about to be merged with another component, and maintain the equivalence classes of $\mathcal{Q}(i)$ only for this same subset of quads. Simple components that are well separated from other components are not involved at stage i.

The equivalence classes of $\mathcal{Q}(i)$ are computed by finding $k = 7^3 - 1$ nearest neighbors of each i-quad q, using the well-separated pair decomposition of [7], and by testing which of them overlaps q.[15] This is different from the planar case of [18], where the nearest-neighbors algorithm is not needed (instead, the plane is simply swept).

To recap, at each "interesting" stage i, we construct $\mathcal{Q}(i)$ from $\mathcal{Q}(i-2)$, by invoking the *Growth* procedure on the set of complex components and simple components that are about to merge with other components. As argued in [18], repeated applications of *Growth* decrease the size of $\mathcal{Q}(i)$ (specifically, after each pair of consecutive steps of *Growth*, $|\mathcal{Q}(i)|$ is at most $\frac{3}{4}$ of its previous size), until we reach a single quad containing all of V.

The running time of the L_∞-minimum spanning tree algorithm in [23] is $O(n \log n)$. The k-nearest-neighbors algorithm of [7] requires $O(m_i \log m_i + k m_i) = O(m_i \log m_i)$ time to process $m_i = |\mathcal{Q}(i)|$ quads, when computing the equivalence classes of $\mathcal{Q}(i)$. As argued in [18], $\sum_i m_i = O(n)$; hence, it takes $O(n \log n)$ total time to perform this step. The space requirements of the MST construction in [23], and of the k-nearest-neighbors computation, are both $O(n)$, as well as the space requirements of the other stages of the algorithm. Other steps of the algorithm *Build-subdivision* are similar to those in [18], and therefore, the algorithm *Build-subdivision* can be implemented to run using $O(n \log n)$ standard operations on a real RAM, plus $O(n)$ floor and base-2 logarithm operations. As shown in [18], the total cost of all the calls to *Growth* is $O(n \log n)$, and this procedure requires only linear space; hence, S_{3D} can be constructed in overall $O(n \log n)$ time, using $O(n)$ space.

7 Extensions and Concluding Remarks

We have presented an optimal-time algorithm for computing an implicit representation of the shortest path map from a fixed source on the surface of a convex polytope with n facets in three dimensions. The algorithm takes $O(n \log n)$ preprocessing time and $O(n \log n)$ storage, and answers a shortest path query (which identifies the path and computes its length) in $O(\log n)$ time. We have used and adapted the ideas of Hershberger and Suri [18], solving Open Problem 2 of their paper, to construct "on the fly" a dynamic version of the incidence data structure of Mount [28], answering in the affirmative the question that was left open in [28].

[15] For each i-quad q, at most $7^3 - 1$ different i-quads $q' \neq q$ can be packed so that q' overlaps q.

As in the planar case (see [18]), our algorithm can also easily be extended to a more general instance of the shortest path problem that involves *multiple sources* on the surface of P, which is equivalent to computing their (implicit) *geodesic Voronoi diagram*. This is a partition of ∂P into regions, so that all points in a region have the same nearest source and the same combinatorial structure (i.e., maximal edge sequence) of the shortest paths to that source. We only compute this diagram implicitly, so that, given a query point $q \in \partial P$, we can identify the nearest source point s to q, and return the shortest path length and starting direction (and, if needed, the shortest path itself) from s to q; this is an easy adaptation of the algorithm presented in this paper, with minor (and obvious) modifications. One can show that, for m given sources, the algorithm processes $O(m+n)$ events in total $O((m+n)\log(m+n))$ time, using $O((m+n)\log(m+n))$ storage; afterwards, a nearest-source query can be answered in $O(\log(m+n))$ time.

It is natural to extend the wavefront propagation method to the shortest path problem on the surface of a *nonconvex* polyhedral surface. As our more recent results [33] show, such an extension, which still runs in optimal $O(n\log n)$ time, exists for several restricted classes of "realistic" polyhedra, such as a polyhedral terrain whose maximal facet slope is bounded, and a few other classes. However, the question of whether a subquadratic-time algorithm exists for the most general case of nonconvex polyhedra, remains open.

Finally, we conclude with two less prominent open problems.

1. Can the space complexity of the algorithm be reduced to linear? Note that our $O(n\log n)$ storage bound is a consequence of only the need to perform path copying to ensure persistence of the surface unfolding data structure in Sect. 2.4 and the source unfolding data structure in Sect. 5.1. Note also that the related algorithms of [18] and [28] also use $O(n\log n)$ storage.
2. Can an unfolding of a surface cell of S overlap itself?

Acknowledgements We thank Joseph O'Rourke for his thorough review of this paper, as well as for the valuable comments and material on surface unfolding and overlapping, and for remarks on Kapoor's paper. We are also grateful to Haim Kaplan for his invaluable help in designing the data structures, to Joe Mitchell for his comments of Kapoor's paper, and to an anonymous referee for a careful review and many useful suggestions. We also acknowledge, with thanks, a recent correspondence with Sanjiv Kapoor concerning some of the details in his paper.

References

1. Agarwal, P.K., Aronov, B., O'Rourke, J., Schevon, C.A.: Star unfolding of a polytope with applications. SIAM J. Comput. **26**, 1689–1713 (1997)
2. Agarwal, P.K., Har-Peled, S., Sharir, M., Varadarajan, K.R.: Approximate shortest paths on a convex polytope in three dimensions. J. ACM **44**, 567–584 (1997)
3. Aleksandrov, L., Lanthier, M., Maheshwari, A., Sack, J.-R.: An ϵ-approximation algorithm for weighted shortest paths on polyhedral surfaces. In: 6th Scand. Workshop Algorithm Theory. Lecture Notes in Computer Science, vol. 1432, pp. 11–22. Springer, Berlin (1998)
4. Aleksandrov, L., Maheshwari, A., Sack, J.-R.: An improved approximation algorithm for computing geometric shortest paths. In: 14th FCT. Lecture Notes in Computer Science, vol. 2751, pp. 246–257. Springer, Berlin (2003)
5. Aloupis, G., Demaine, E.D., Langerman, S., Morin, P., O'Rourke, J., Streinu, I., Toussaint, G.: Unfolding polyhedral bands. In: Proc. 16th Canad. Conf. Comput. Geom., pp. 60–63 (2004)

6. Aronov, B., O'Rourke, J.: Nonoverlap of the star unfolding. Discrete Comput. Geom. **8**, 219–250 (1992)
7. Callahan, P.B., Kosaraju, S.R.: A decomposition of multidimensional point sets with applications to k-nearest-neighbors and n-body potential fields. J. ACM **42**(1), 67–90 (1995)
8. Chen, J., Han, Y.: Shortest paths on a polyhedron, part I: computing shortest paths. Int. J. Comput. Geom. Appl. **6**, 127–144 (1996)
9. Chen, J., Han, Y.: Shortest paths on a polyhedron, part II: storing shortest paths. Tech. Rept. 161-90, Comput. Sci. Dept., Univ. Kentucky, Lexington, KY, February 1990
10. de Berg, M., van Kreveld, M., Snoeyink, J.: Two- and three-dimensional point location in rectangular subdivisions. J. Algorithms **18**, 256–277 (1995)
11. Demaine, E.D., O'Rourke, J.: Geometric Folding Algorithms: Linkages, Origami, and Polyhedra. Cambridge University Press, Cambridge (2007)
12. Driscoll, J.R., Sleator, D.D., Tarjan, R.E.: Fully persistent lists with catenation. J. ACM **41**(5), 943–949 (1994)
13. Edelsbrunner, H., Guibas, L.J., Stolfi, J.: Optimal point location in a monotone subdivision. SIAM J. Comput. **15**, 317–340 (1986)
14. Guibas, L., Hershberger, J., Leven, D., Sharir, M., Tarjan, R.E.: Linear time algorithms for visibility and shortest path problems inside simple polygons. Algorithmica **2**, 209–233 (1987)
15. Guibas, L.J., Sedgewick, R.: A dichromatic framework for balanced trees. In: Proc. 19th IEEE Sympos. Found. Comput. Sci., pp. 8–21 (1978)
16. Har-Peled, S.: Approximate shortest paths and geodesic diameters on convex polytopes in three dimensions. Discrete Comput. Geom. **21**, 216–231 (1999)
17. Har-Peled, S.: Constructing approximate shortest path maps in three dimensions. SIAM J. Comput. **28**(4), 1182–1197 (1999)
18. Hershberger, J., Suri, S.: An optimal algorithm for Euclidean shortest paths in the plane. SIAM J. Comput. **28**(6), 2215–2256 (1999). Earlier versions: in Proc. 34th IEEE Sympos. Found. Comput. Sci., pp. 508–517 (1993); Manuscript, Washington Univ., St. Louis (1995)
19. Hershberger, J., Suri, S.: Practical methods for approximating shortest paths on a convex polytope in $\mathbb{R}^3$. Comput. Geom. Theory Appl. **10**(1), 31–46 (1998)
20. Italiano, G.F., Raman, R.: Topics in Data Structures. In: Atallah, M.J. (ed.) Handbook on Algorithms and Theory of Computation, Chap. 5. CRC Press, Boca Raton (1998)
21. Kapoor, S.: Efficient computation of geodesic shortest paths. In: Proc. 32nd Annu. ACM Sympos. Theory Comput., pp. 770–779 (1999)
22. Kirkpatrick, D.: Optimal search in planar subdivisions. SIAM J. Comput. **12**, 28–35 (1983)
23. Krznaric, D., Levcopoulos, C., Nilsson, B.J.: Minimum spanning trees in d dimensions. Nord. J. Comput. **6**(4), 446–461 (1999)
24. Lanthier, M., Maheshwari, A., Sack, J.-R.: Approximating shortest paths on weighted polyhedral surfaces. Algorithmica **30**(4), 527–562 (2001)
25. Mata, C., Mitchell, J.S.B.: A new algorithm for computing shortest paths in weighted planar subdivisions. In: Proc. 13th Annu. ACM Sympos. Comput. Geom., pp. 264–273 (1997)
26. Mitchell, J.S.B., Mount, D.M., Papadimitriou, C.H.: The discrete geodesic problem. SIAM J. Comput. **16**, 647–668 (1987)
27. Mount, D.M.: On finding shortest paths on convex polyhedra. Tech. Rept., Computer Science Dept., Univ. Maryland, College Park, October 1984
28. Mount, D.M.: Storing the subdivision of a polyhedral surface. Discrete Comput. Geom. **2**, 153–174 (1987)
29. O'Rourke, J.: Computational geometry column 35. Int. J. Comput. Geom. Appl. **9**, 513–515 (1999); also in SIGACT News, 30(2), 31–32 (1999)
30. O'Rourke, J.: On the development of the intersection of a plane with a polytope. Tech. Rept. 068, Smith College, June 2000
31. O'Rourke, J., Suri, S., Booth, H.: Shortest paths on polyhedral surfaces. Manuscript, The Johns Hopkins Univ., Baltimore, MD (1984)
32. Paul, R.P.: Robot Manipulators: Mathematics, Programming, and Control. MIT Press, Cambridge (1981)
33. Schreiber, Y.: Shortest paths on realistic polyhedra. In: Proc. 23rd Annu. ACM Sympos. Comput. Geom., pp. 74–83 (2007)
34. Schreiber, Y., Sharir, M.: An optimal-time algorithm for shortest paths on a convex polytope in three dimensions, http://www.tau.ac.il/~michas/ShortestPath.pdf. Also in Y. Schreiber, PhD thesis, http://www.tau.ac.il/~syevgeny/SchreiberThesis.pdf

35. Sharir, M.: On shortest paths amidst convex polyhedra. SIAM J. Comput. **16**, 561–572 (1987)
36. Sharir, M., Schorr, A.: On shortest paths in polyhedral spaces. SIAM J. Comput. **15**, 193–215 (1986)
37. Tarjan, R.E.: Data structures and network algorithms. SIAM CBMS, p. 44 (1983)
38. Varadarajan, K.R., Agarwal, P.K.: Approximating shortest paths on a nonconvex polyhedron. In: Proc. 38th Annu. IEEE Sympos. Found. Comput. Sci., pp. 182–191 (1997)
39. Weisstein, E.W.: Riemann surface. MathWorld—a Wolfram web resource. http://mathworld.wolfram.com/RiemannSurface.html

General-Dimensional Constrained Delaunay and Constrained Regular Triangulations, I: Combinatorial Properties

Jonathan Richard Shewchuk

Abstract Two-dimensional constrained Delaunay triangulations are geometric structures that are popular for interpolation and mesh generation because they respect the shapes of planar domains, they have "nicely shaped" triangles that optimize several criteria, and they are easy to construct and update. The present work generalizes constrained Delaunay triangulations (CDTs) to higher dimensions and describes constrained variants of regular triangulations, here christened *weighted CDTs* and *constrained regular triangulations*. CDTs and weighted CDTs are powerful and practical models of geometric domains, especially in two and three dimensions.

The main contributions are rigorous, theory-tested definitions of CDTs and piecewise linear complexes (geometric domains that incorporate nonconvex faces with "internal" boundaries), a characterization of the combinatorial properties of CDTs and weighted CDTs (including a generalization of the Delaunay Lemma), the proof of several optimality properties of CDTs when they are used for piecewise linear interpolation, and a simple and useful condition that guarantees that a domain has a CDT. These results provide foundations for reasoning about CDTs and proving the correctness of algorithms. Later articles in this series discuss algorithms for constructing and updating CDTs.

Supported in part by the National Science Foundation under Awards CMS-9318163, ACI-9875170, CMS-9980063, CCR-0204377, CCF-0430065, and EIA-9802069, in part by the Advanced Research Projects Agency and Rome Laboratory, Air Force Materiel Command, USAF under agreement number F30602-96-1-0287, in part by an Alfred P. Sloan Research Fellowship, and in part by gifts from the Okawa Foundation and Intel. The views in this document are those of the author. They are not endorsed by the sponsors or the U.S. Government.

J.R. Shewchuk (✉)
Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720, USA
e-mail: jrs@cs.berkeley.edu

1 Introduction

Many geometric applications can benefit from triangulations that have properties similar to Delaunay triangulations, but are constrained to contain specified edges or faces. Delaunay triangulations have virtues when they are used to interpolate multivariate functions [13, 28, 37, 50], including a tendency to favor "round" simplices over "skinny" ones. However, some applications rely on the presence of faces that represent specified discontinuities, as illustrated in Fig. 1, and the Delaunay triangulation might not respect these constraints. Triangulations also serve as meshes that represent objects for rendering or for the numerical solution of partial differential equations. For these purposes, Delaunay triangulations have many advantages, but the triangulations are required to assume the shapes of the objects being modeled, and perhaps to resolve interfaces where different materials meet or where boundary conditions are applied.

In two dimensions there are two popular alternatives for creating a Delaunay-like triangulation that respects constraints. In either case, the input is a *planar straight line graph* (PSLG), such as the one illustrated in Fig. 2(a). A PSLG X is a set of vertices and segments (constraining edges) that satisfy two restrictions: both endpoints of every segment in X are members of X, and a segment in X may intersect other segments and vertices in X only at its endpoints. A triangulation is sought that contains the vertices in X and respects the segments in X.

The first alternative is to form a *conforming Delaunay triangulation* (Fig. 2(c)). The vertices of X are augmented by additional vertices (sometimes called *Steiner points*) carefully chosen so that the Delaunay triangulation of the augmented vertex set conforms to all the segments—in other words, so that each segment is represented by a contiguous linear sequence of edges of the triangulation. Edelsbrunner

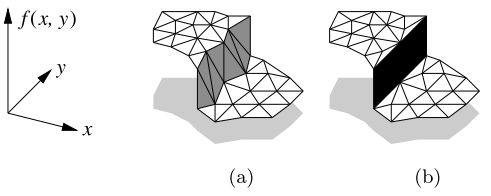


Fig. 1 A triangulation that respects a discontinuity in a function (b) can be a better interpolating surface than one that does not (a)

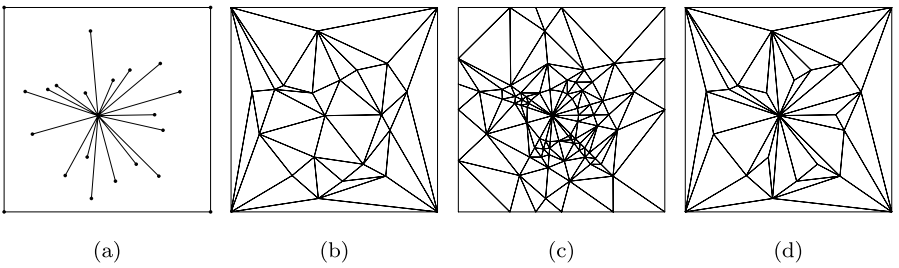


Fig. 2 The Delaunay triangulation (b) of the vertices of a PSLG (a) might not respect the segments of the PSLG. These segments can be incorporated by adding vertices to obtain a conforming Delaunay triangulation (c), or by forgoing Delaunay triangles in favor of constrained Delaunay triangles (d)

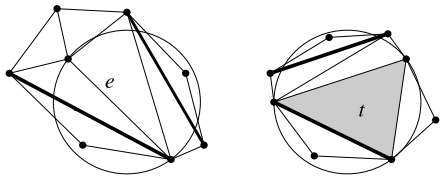


Fig. 3 The edge e and the triangle t are both constrained Delaunay. Bold lines represent segments

and Tan [20] show that a PSLG X can be triangulated with the addition of $\mathcal{O}(m^2n)$ augmenting vertices, where m is the number of segments in X, and n is the number of vertices in X. For many PSLGs, their algorithm uses far fewer augmenting vertices, but the numbers required in practice are often undesirably large. PSLGs are known that have no conforming Delaunay triangulation with fewer than $\Theta(mn)$ augmenting vertices. Closing the gap between the $\mathcal{O}(m^2n)$ and $\Omega(mn)$ bounds remains an open problem.

The second alternative is to form a *constrained Delaunay triangulation* (CDT) [9, 29, 43], illustrated in Fig. 2(d). A CDT of X has no vertices not in X, and every segment in X is a single edge of the CDT. However, a CDT, despite its name, is not a Delaunay triangulation. In an ordinary Delaunay triangulation, every simplex (triangle, edge, or vertex) is *Delaunay*. A simplex is Delaunay if its vertices are in X and there exists a *circumcircle* of the simplex—a circle that passes through all its vertices—that encloses no vertex in X. (Any number of vertices is permitted on the circle.) In a CDT this requirement is waived, and instead every simplex must either be a segment specified in X or be *constrained Delaunay*. A simplex is constrained Delaunay if it has a circumcircle that encloses no vertex in X that is *visible* from any point in the relative interior of the simplex—here visibility is occluded only by segments in X—and furthermore, the simplex does not "cross" any segment. (For a formal definition, see Section 1.1.)

Figure 3 demonstrates examples of a constrained Delaunay edge e and a constrained Delaunay triangle t. Segments in X appear as bold lines. Although there is no empty circle that encloses e, the depicted circumcircle of e encloses no vertex that is visible from the relative interior of e. There are two vertices inside the circle, but both are hidden behind segments. Hence, e is constrained Delaunay. Similarly, the sole circumcircle of t encloses two vertices, but both are hidden from the interior of t by segments, so t is constrained Delaunay.

The advantage of a CDT over a conforming Delaunay triangulation is that it has no vertex other than those in X. The advantage of a conforming Delaunay triangulation is that its triangles are Delaunay, whereas those of a CDT are not. Nevertheless, CDTs retain many of the desirable properties of Delaunay triangulations. For instance, a two-dimensional CDT maximizes the minimum angle in the triangulation, compared with all other constrained triangulations of X [29].

We live in a three-dimensional world, and those who model it have a natural interest in constructing constrained and conforming triangulations in three or more dimensions. Algorithms by Murphy et al. [33], Cohen-Steiner et al. [12], Cheng and Poon [8], and Pav and Walkington [34] can construct a conforming Delaunay tetrahedralization of any three-dimensional polyhedron by inserting carefully chosen vertices on the boundary of the polyhedron. (Their algorithms work not only on polyhedra, but also on a more general input called a *piecewise linear complex*, defined

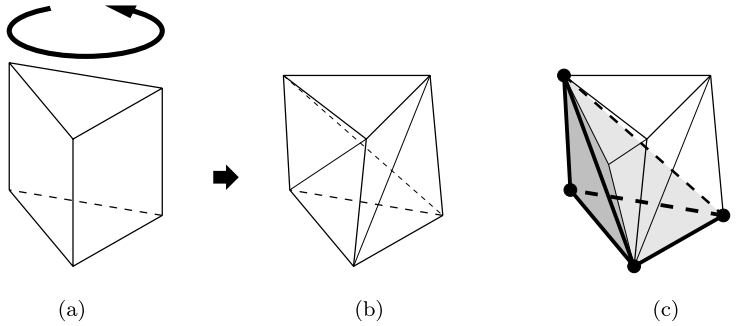


Fig. 4 Schönhardt's untetrahedralizable polyhedron (b) is formed by rotating one end of a triangular prism (a), thereby creating three diagonal reflex edges. Every tetrahedron defined on the vertices of Schönhardt's polyhedron sticks out (c)

below.) These algorithms might introduce a huge number of new vertices. No known algorithm for finding conforming Delaunay tetrahedralizations is guaranteed to introduce only a polynomial number of new vertices, and no algorithm of any complexity has been offered for four- or higher-dimensional conforming Delaunay triangulations.

Prior to the present work (in its first incarnation [45]), CDTs had not been generalized to dimensions higher than two. One reason is that in three or more dimensions, there are polytopes that cannot be triangulated at all without additional vertices. Schönhardt [41] furnishes a three-dimensional example depicted in Fig. 4(b). The easiest way to envision this polyhedron is to begin with a triangular prism (Fig. 4(a)). Imagine twisting the prism so that the top triangular face rotates slightly like the lid of a jar, while the bottom triangular face is fixed in place. Each of the three square faces is broken along a diagonal *reflex edge* (an edge at which the polyhedron is locally nonconvex) into two triangular faces. After this transformation, the upper left corner and lower right corner of each (formerly) square face are separated by a reflex edge, and the line segment connecting them is outside the polyhedron. Any four vertices of the polyhedron include two separated by a reflex edge; thus, any tetrahedron whose vertices are vertices of the polyhedron does not lie entirely within the polyhedron, as illustrated in Fig. 4(c). Schönhardt's polyhedron cannot be tetrahedralized without an additional vertex. (One extra vertex at its center will do.)

Ruppert and Seidel [40] add to the difficulty by proving that it is NP-hard to determine whether a three-dimensional polyhedron is tetrahedralizable. Even among polyhedra that can be triangulated without additional vertices, there is not always a triangulation that is in any reasonable sense "constrained Delaunay."

What features of polytopes make them amenable to being triangulated with Delaunay-like simplices? This article offers a partial answer by proposing a conservative extension of the definition of CDT to higher dimensions, and by demonstrating that there is an easily tested and enforced, sufficient (but not necessary) condition that guarantees that a CDT exists. This article also shows that CDTs optimize several criteria for the accuracy of piecewise linear interpolation of certain classes of functions. These results extend to *weighted CDTs* (a constrained generalization of regular triangulations), wherein each vertex is assigned a numerical weight that influences the triangulation.

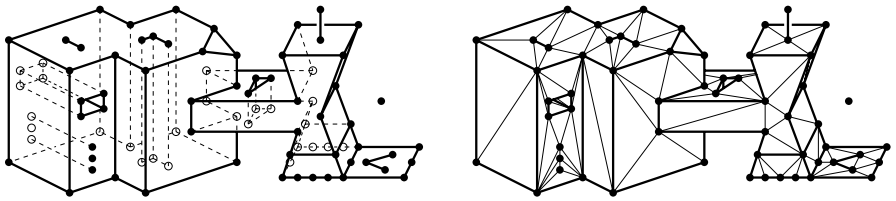


Fig. 5 Each facade of a PLC (a) may have holes, slits, and interior vertices, which are used to constrain a triangulation or to support intersections with other facades. (b) The constrained Delaunay triangulation of the PLC in (a). It is a PLC, too

There is more than one way in which the notion of "constrained Delaunay" might generalize to three or more dimensions. The choices made here yield useful CDTs and efficient algorithms for their construction, though other generalizations of CDTs might be discovered in the future.

This article is the first in a three-part series. The second article discusses sweep and gift-wrapping algorithms for constructing the CDT of any piecewise linear complex that has one, except for a class of difficult, "nongeneric" inputs. It also demonstrates the NP-completeness of determining whether a nongeneric polyhedron has a CDT. The third article discusses algorithms for updating a CDT to reflect the insertion or deletion of a $(d-1)$-facade, a vertex, or several vertices, as well as an incremental algorithm for constructing CDTs that have a property called "ridge protection" (described in the next section).

1.1 Summary of Results

The input is a *piecewise linear complex* (PLC), following Miller et al. [32].[1] A PLC is a finite set of *facades* in an ambient space E^d. A facade is a polytope (roughly speaking) of any dimension from zero to d, possibly with holes and lower-dimensional facades inside it. Figure 5 illustrates a three-dimensional PLC. As the figure shows, a facade may have any number of sides, may be nonconvex, and may have holes, slits, or vertices inside it. A *k-facade* is a k-dimensional facade. 0-Facades are vertices, and 1-facades are segments. Observe that a PSLG is a two-dimensional PLC without 2-facades.

PLCs have restrictions similar to those of PSLGs or any other type of complex. For each facade f in a PLC X, the boundary of f must be composed of lower-dimensional facades in X. Nonempty intersections of facades in X must be facades in X. For details, see Section 2.1, where the terms *facade* and *PLC* are defined with full mathematical rigor.

The purpose of most facades is to constrain a triangulation. A d-dimensional PLC typically includes d-facades, whose purpose is to specify what region the triangulation should fill. The union of all the facades in a PLC X is the *triangulation domain*

[1] Miller et al. call it a *piecewise linear system*, but their construction is so obviously a complex that a change in name seems obligatory.

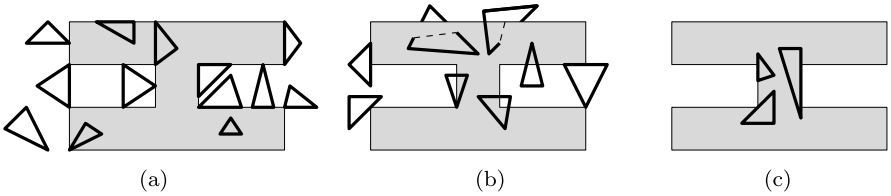


Fig. 6 (a) Examples of triangles that respect a shaded facade. (b) Examples of triangles that do not respect the facade. (c) Examples of triangles that respect the facade, but do not respect all its edges and vertices (which are facades themselves)

$|X|$, the portion of space a user wishes to triangulate. The specification of a triangulation domain is sometimes crucial, because there are PLCs for which a CDT of the triangulation domain exists but a CDT of its convex hull does not. For example, it is easy to tetrahedralize the region sandwiched between Schönhardt's polyhedron and a suitable bounding box, even though the interior of the polyhedron is not tetrahedralizable.

The complement of the triangulation domain, $E^d \backslash |X|$, is called the *exterior domain* and includes any hollow cavities enclosed by the triangulation domain, as well as outer space. Because X is a complex, some of its $(d-1)$-facades separate the interior of the triangulation domain from the exterior domain. However, not all $(d-1)$-facades play this role. Figure 5 includes several *dangling* lower-dimensional facades that are not part of any d-facade. Some facades are *internal facades*, which do not lie on the boundary of the exterior domain. These facades allow PLCs to represent multiple-component domains and domains with nonmanifold boundaries.

The goal of this work is to subdivide a domain into simplices. A *k-simplex* is a k-dimensional simplex—the convex hull of $k+1$ affinely independent points. A *triangulation* or *simplicial complex* T is a finite set of simplices that intersect each other "nicely": T contains every face of every simplex in T, and the intersection of any two simplices in T is either empty or a face of both simplices. A triangulation T *fills* a PLC X if $\bigcup_{t \in T} t = \bigcup_{f \in X} f$; that is, if the union of simplices in T is the triangulation domain $|X|$.

Of course, not all triangulations that fill X are equally good. Facades constrain what sort of simplex is acceptable. A simplex s *respects* a facade f if $s \cap f$ is a union of faces of s (possibly empty). As Fig. 6 illustrates, the intersection of a nonconvex facade and a triangle that respects it might be the empty set, a vertex, an edge, the entire triangle, the union of two or all three edges, the union of two or all three vertices, or the union of an edge and opposite vertex of the triangle. Loosely speaking, if s respects f, then s cannot "cross" f or f's boundary.

A simplex s *respects* X if $s \subseteq |X|$ and s respects every facade in X, except perhaps some of the vertices. (Weighted CDTs may omit some of the vertices in X, unlike ordinary CDTs, but some designated vertices must be respected; see Section 2.3 for details.)

A triangulation T is a *triangulation of* X if T fills X, every simplex in T respects X, and every vertex in T is in X. This definition implies that every facade in X (except perhaps the vertices) is a union of simplices in T. (See Section 2.3 for a discussion of why the definition does not explicitly require every vertex in X to be

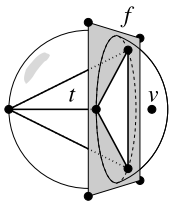


Fig. 7 A constrained Delaunay tetrahedron t

in T. This requirement arises implicitly if every vertex is designated as one that must be respected.)

Sometimes it is desirable to permit a triangulation to have vertices not present in X—and sometimes it is necessary, as Schönhardt demonstrates. A triangulation T is a *conforming triangulation* or *Steiner triangulation* of X if T fills X and every simplex in T respects X. This article is devoted to pure triangulations in which extra vertices are not permitted, but Steiner triangulations are investigated elsewhere [47, 49].

Within a PLC X, the visibility between two points p and q is *occluded* if $pq \not\subseteq |X|$ or there is a facade between p and q whose affine hull contains neither p nor q. (Note, however, that some vertices do not obstruct visibility—namely those that the triangulation is not required to respect. See Section 2.4.) The points p and q are *visible* from each other if $pq \subseteq |X|$ and X contains no occluding facade.

Let s be a k-simplex (for any k) whose vertices are in X (though s is not necessarily a facade in X). Let S be a (full-dimensional) hypersphere in E^d. S is a *circumsphere* of s if S passes through all the vertices of s. If $k = d$, then s has a unique circumsphere; otherwise, s has infinitely many circumspheres. The simplex s is *Delaunay* if there exists a circumsphere S of s that encloses no vertex in X. The simplex s is *strongly Delaunay* if there exists a circumsphere S of s such that no vertex in X lies inside *or on* S, except the vertices of s. Every 0-simplex (vertex) is trivially strongly Delaunay.

A simplex s is *constrained Delaunay* if

- the vertices of s are in X,
- s respects X, and
- there is a circumsphere S of s such that no vertex of X inside S is visible from any point in the relative interior of s.

Figure 7 depicts a constrained Delaunay tetrahedron t in E^3. The intersection of t with the facade f is a face of t, so t respects X. The circumsphere of t encloses one vertex v, but v is not visible from any point in the interior of t, because f occludes its visibility.

A *constrained Delaunay triangulation* T of X is a triangulation of X in which every d-simplex is constrained Delaunay. If X has dangling facades, this characterization is insufficient, and we must resort to the (less readable) true definition: a CDT T of X is a triangulation of X in which each simplex is constrained Delaunay "within" the lowest-dimensional facade in X that includes it. For example, a three-dimensional CDT fills each 2-facade (dangling or not) with triangles that are constrained Delaunay "within" that 2-facade, and collectively comprise a two-dimensional CDT of the 2-facade. However, those triangles might not be constrained Delaunay within the three-dimensional PLC—they might have empty circumcircles,

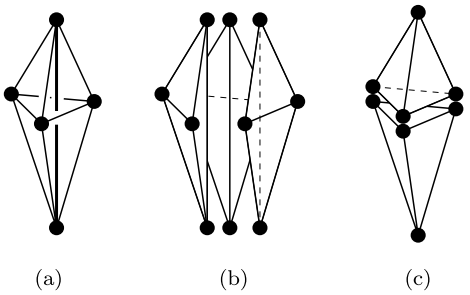


Fig. 8 (a) A PLC with no CDT. (b) The sole tetrahedralization of this PLC. Its three tetrahedra are not constrained Delaunay. (c) The two Delaunay tetrahedra do not respect the central segment

but not empty circumspheres. (That is why the 2-facade is there—to enforce the presence of triangles that might otherwise be absent.) The definition of "CDT" is therefore recursive in the dimension. See Section 2.4 for details.

The first main result of this article is a characterization of many basic properties of constrained Delaunay and weighted constrained Delaunay triangulations, analogous to the well-known properties of Delaunay triangulations. (Weighted CDTs are defined in Section 2.4.) For example, every face of a constrained Delaunay simplex is itself constrained Delaunay within some facade. A CDT of a facade includes CDTs of all the facade's faces (Section 3.1). If a PLC has no $d+2$ vertices lying on a common hypersphere, then its constrained Delaunay simplices have disjoint relative interiors and form a simplicial complex, and it has at most one CDT (Section 3.3). The Delaunay Lemma, which guarantees that a triangulation of a vertex set is Delaunay if and only if its facets are locally Delaunay [14], generalizes to CDTs (Section 3.2). The Delaunay Lemma is a fundamental tool for verifying that a triangulation is a CDT, and for dynamically maintaining the CDT of a PLC whose vertices are moving or changing their weights.

The second main result is that CDTs are optimal by several criteria (described in Section 4) when they are used for piecewise linear interpolation. This fact is among the reasons why CDTs are so valuable.

The third main result is a condition that guarantees the existence of a CDT. The main impediment to the existence of CDTs is the difficulty of respecting facades of dimension $d-2$ or less. Figure 8 offers an example of a three-dimensional PLC with no CDT. There is one segment that runs through the interior of the PLC. There is only one tetrahedralization of this PLC—composed of three tetrahedra encircling the central segment—and its tetrahedra are not constrained Delaunay, because each of them has a visible vertex inside its circumsphere. If the central segment were removed, the PLC would have a CDT made up of two tetrahedra.

The condition that guarantees that a PLC has a CDT is easiest to describe, and easiest to enforce, in three dimensions. A three-dimensional PLC X is *ridge-protected* if every segment (1-facade) in X is strongly Delaunay. (See Section 2.4 for the general-dimensional definition.) Every ridge-protected PLC has a CDT. This result, called the *CDT Theorem*, makes three-dimensional CDTs useful in geometric modeling applications.

It is not sufficient for every segment to be Delaunay. If Schönhardt's polyhedron is embedded so that all six of its vertices lie on a common sphere, then all of its edges (and its triangular faces as well) are Delaunay, but it still does not have a tetrahedral-

ization. It is not possible to place the vertices of Schönhardt's polyhedron so that all three of its reflex edges are strongly Delaunay (though any two may be).

Here is a stronger and even more useful version of the CDT Theorem. In three dimensions a segment may serve as a boundary to several 2-facades, which can be sorted by their rotary order around the segment. A segment is *grazeable* if two consecutive 2-facades in the rotary order are separated by an interior angle of 180° or more, or if the segment is included in fewer than two 2-facades and is internal, not dangling. (An *interior angle* subtends the interior of the triangulation domain. Exterior angles of 180° or more are irrelevant to the CDT Theorem.) Only the grazeable segments need to be strongly Delaunay to guarantee a CDT. A three-dimensional PLC X is *weakly ridge-protected* if every grazeable segment in X is strongly Delaunay. Every weakly ridge-protected PLC has a CDT.

Segments that are not grazeable are common. For instance, in a complex of convex polyhedra, no segment is grazeable. The stronger result exempts the segments of the complex from the need to be strongly Delaunay.

Testing whether a PLC is ridge-protected, or weakly ridge-protected, is straightforward. See the comments following Definition 23.

This article's results extend to weighted CDTs, which are described in Section 2.4. Weighted CDTs are central in the design of flip algorithms for updating and constructing CDTs; see the third article in this series. Several researchers have shown that weighted Delaunay triangulations are useful for three-dimensional mesh generation, because some undesirable tetrahedra can be removed by adjusting the vertex weights [6, 7, 16]. Weighted CDTs share this virtue and are even more powerful, because of the ease with which they respect the shape of a domain.

The definition of "ridge-protected" generalizes to weighted PLCs, and every weakly ridge-protected, weighted PLC has a weighted CDT. Interestingly, even in two dimensions there are weighted PLCs that do not have weighted CDTs.

1.2 Benefits of the CDT Theorem

Why is it useful to know that weakly ridge-protected PLCs have CDTs? Although a given PLC X might not be weakly ridge-protected, the insertion of additional vertices can transform it into a weakly (or fully) ridge-protected PLC Y, which has a CDT. The CDT of Y is not a CDT of X, because it has vertices that X lacks, but it is a *conforming CDT* or *Steiner CDT* of X: "conforming" or "Steiner" because boundary conformity is obtained by inserting new vertices (Steiner points), and "CDT" because the simplices of the Steiner CDT are constrained Delaunay (rather than Delaunay).

Compare this idea with the most common methods of recovering missing facades in three-dimensional Delaunay-based mesh generation algorithms, which insert additional vertices into all the missing facades. Some of these algorithms produce conforming Delaunay meshes [8, 34, 39], and some recover the missing facades by bisecting and flipping tetrahedra, yielding a mesh that is not necessarily Delaunay nor constrained Delaunay, although you might say it is "almost" Delaunay [22, 27, 52, 53]. Figure 9 illustrates the advantage of a Steiner CDT. All the procedures use vertex insertions to recover missing grazeable segments, but the customary approaches require additional vertex insertions to recover missing 2-facades and non-grazeable segments. A Steiner CDT does not need these extra vertices.

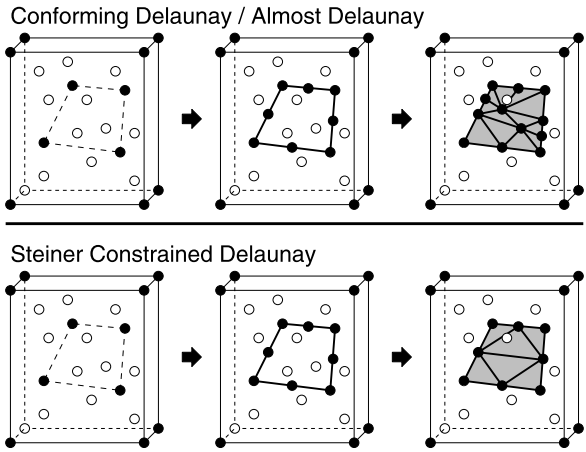


Fig. 9 Two methods for recovering a 2-facade in the interior of a cubical triangulation domain. The initial Delaunay tetrahedralization does not respect the facade. (For clarity, the tetrahedra are not shown.) Both methods insert new vertices to recover missing segments. Next, the customary method is to insert more vertices to recover missing 2-facades (top), but no additional vertices are needed if constrained Delaunay tetrahedra are used (bottom)

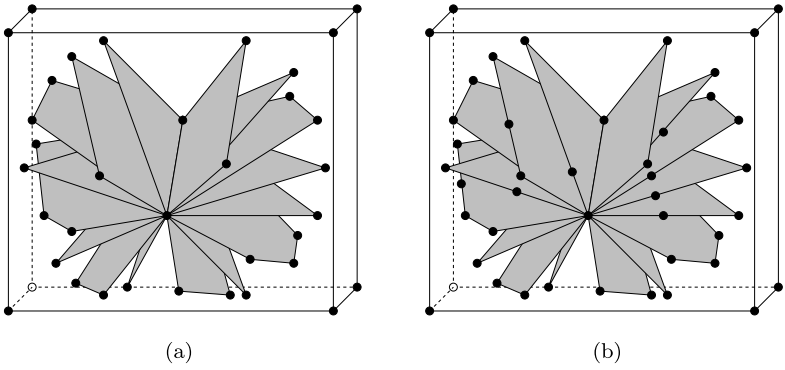


Fig. 10 (a) It is difficult to mesh the interior of this box with Delaunay tetrahedra that conform to all the facades. (b) The box can be meshed with constrained Delaunay tetrahedra with the addition of just the vertices shown

Figure 10(a) depicts an example of a PLC for which a Steiner CDT is much more effective than a conforming Delaunay tetrahedralization. In the interior of the box, many oddly shaped 2-facades adjoin a single shared segment. The triangulation domain is the entire box. Vertices inserted to recover one 2-facade—so that it is a union of triangular faces of the Delaunay tetrahedralization—are likely to knock out triangles from the adjacent 2-facades. The aforementioned algorithms of Murphy et al. and others [8, 12, 33, 34] can construct conforming Delaunay tetrahedralizations of this PLC, but they require many more vertices than are needed to form a Steiner CDT, most of them in the 2-facade interiors. The PLC augmented with a modest number of vertices (Fig. 10(b)) is weakly ridge-protected and has a CDT.

I conjecture that for the worst three-dimensional PLCs, conforming Delaunay triangulations need asymptotically more vertices than Steiner CDTs. It is an open question whether this is true, but based on the two-dimensional complexity results, it seems like a safe gamble.

An algorithm that decides how to choose new vertices so that there are provably good bounds on the edge lengths of the Steiner CDT (i.e. edges are not made unnecessarily short) is described elsewhere [47]. This algorithm does not guarantee a polynomial bound on the number of new vertices, but its guarantees on edge lengths are in some ways more useful, because the Steiner CDT is an excellent starting triangulation for several algorithms for three-dimensional mesh generation. One algorithm uses the constrained Delaunay property to guarantee its ability to tetrahedralize any PLC [46], and another uses it to establish provable bounds on the quality of the tetrahedra it produces and on the edge lengths of the final mesh [44]. The results in this article underpin those algorithms.

Why does this article take PLCs as the input rather than, for simplicity, boundary triangulations? Consider finding a tetrahedralization of a cube. The edges of the cube are strongly Delaunay, so the CDT Theorem guarantees that the cube has a CDT. By contrast, consider a boundary triangulation of a cube. Any boundary triangulation bisects each square face of the cube with a diagonal edge. These diagonals are not strongly Delaunay, so the CDT Theorem does not apply. Moreover, a tetrahedralization respecting the boundary triangulation might not exist (depending on the choice of diagonals). Thus, the option to specify facades more general than simplices is an advantage both for the theorem and for CDT construction algorithms, which can choose a compatible set of diagonals.

If a PLC is ridge-protected, its CDT can be built by a simple incremental facade insertion algorithm described in the third article in this series. PLCs that are not ridge-protected (but have CDTs) currently require a more complicated sweep algorithm or a slower gift-wrapping algorithm, described in the second article in this series.

2 Complexes

This section defines the geometric constructions and ideas at the center of this work. The input structures—facades and PLCs—are formalized in Section 2.1. The output structures, a generalization of CDTs called *weighted CDTs*, are described in Sections 2.2–2.4. Definitions are often a perfunctory part of a mathematics article, so it is worth noting that 8 years of trial and error led to the definitions given here. "Constrained Delaunay" and the notion of visibility are defined differently here than in the earlier incarnation of this work [45], and the present definitions are more sound. These and other definitions in this article evolved with the proofs of the theorems here and in the sequel articles.

Throughout this article, the terms "simplex," "triangle," "tetrahedron," and "convex hull" refer to closed, convex sets of points; for instance, a "triangle" is not just three edges, but the points inside as well. The notation $\mathrm{conv}(S)$ represents the convex hull of the point set S.

Some simplices of specific dimensions have their own names. Of course, a vertex is a 0-simplex, an edge is a 1-simplex, a triangle is a 2-simplex, and a tetrahedron is a 3-simplex. In a d-dimensional ambient space, a $(d-2)$-dimensional convex polytope or $(d-2)$-simplex is called a *ridge*, and a $(d-1)$-dimensional convex polytope or $(d-1)$-simplex is called a *facet*.

The notation pq denotes a line segment with endpoints p and q. The notation $p \cdot q$ denotes the Euclidean inner product, $|p| = \sqrt{p \cdot p}$ is the Euclidean norm, and $|pq| = |p - q|$ is the Euclidean length of pq. It might help the reader to know that this article strictly distinguishes between the verbs *contain* for set membership ($\ni$) and *include* for set inclusion ($\supseteq$).

2.1 Piecewise Linear Complexes

Consider points in an ambient space E^d. A *k-flat* (k-dimensional flat) is the affine hull of $k + 1$ affinely independent points. (A flat is also known as an *affine subspace*—unlike a true subspace it is not required to contain the origin. For readers familiar with flats but not affine hulls, the *affine hull* of a point set is the lowest-dimensional flat that includes it.) A set of points $S \subseteq E^d$ is *k-dimensional* if the affine hull of S is a k-flat. (In other words, S contains $k + 1$ affinely independent points, but does not contain $k + 2$ affinely independent points.) A *hyperplane* is a $(d - 1)$-flat. The set of points on one designated side of a hyperplane, excluding every point of the hyperplane itself, is an *open halfspace*. By contrast, a *closed halfspace* includes the hyperplane as well.

An *open convex k-polyhedron* is the nonempty intersection of a k-flat and a finite number of open halfspaces. It is *bounded* if it does not include a ray (equivalently, if its diameter is finite). A *closed convex k-polyhedron* is the closure of an open convex k-polyhedron. The *closure* of a polyhedron has its usual meaning from real analysis—the set of all the points and accumulation points of the polyhedron—and more intuitively is a point set containing all the points of the polyhedron, plus all the points on its boundary.

Definition 1 (Facade) An *open k-facade* is the union of a finite number of bounded, open, convex k-polyhedra, all included in some common k-flat. A *closed k-facade* is the closure of an open k-facade.

Observe that a facade is not required to be connected. A 0-facade (open or closed—there is no difference) is a vertex, and a 1-facade is either a segment or a sequence of collinear segments.

A closed facade is equivalent to Hadwiger's classic polyhedron [26], which is defined to be a union of closed convex polyhedra. It is the open facades that motivate the new name. In geometric modeling, open facades are more versatile than closed facades as abstractions of geometric domains and their boundaries, because an open facade can have internal boundaries. Internal boundaries serve at least two purposes: they support intersections between surfaces, as Fig. 11 illustrates, and they constrain the permissible triangulations of the facade—for instance, to support the application of boundary conditions to a finite-element mesh, or to model discontinuities in the lighting of a surface for computer graphics. Internal boundaries are necessary to model some domains with nonmanifold boundaries, like the domain in Fig. 5.

Definition 2 (External and Internal Boundaries) The *external boundary* of a facade is the boundary of the closure of the facade. (Observe that the external boundary includes boundaries of holes.) The *internal boundary* of an open facade is the boundary

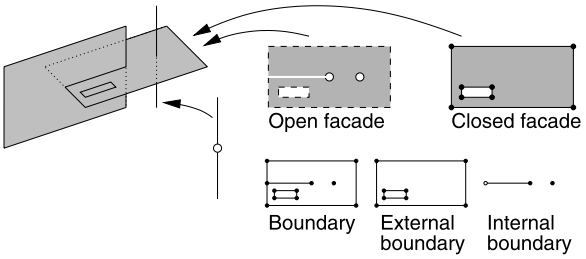


Fig. 11 At left are two connected 2-facades and a 1-facade (composed of two segments). At center and right appear one of the 2-facades, represented as both an open and a closed facade. Dashed lines and open circles represent points that are not part of the open facade. The internal boundary includes a slit and an isolated vertex, both of which are needed to support contacts with other facades. The internal boundary cannot be inferred from the closed facade alone

of the open facade minus the external boundary. (Equivalently, it is the intersection of the boundary with the relative interior of the closure of the facade.) See Fig. 11.

Throughout this article, *relative interior* has its usual meaning from real analysis, but *boundary* is used as shorthand for *relative boundary*, and *open* for *relatively open*.

The faces of a facade are defined in a fundamentally different way than the faces of a convex polyhedron. The faces of a convex polyhedron are an intrinsic property of the polyhedron, whereas the faces of a facade are defined only in the context of a PLC. Compare the following two definitions.

Definition 3 (Face of a Convex Polyhedron) The *faces* of a closed, convex k-polyhedron P are P and every polyhedron found by taking the intersection of P with a hyperplane that does not intersect the relative interior of P. The *proper faces* of P are the faces of dimensionalities zero through $k-1$.

This standard construction also defines the faces of a simplex. For example, the faces of a tetrahedron include its four vertices, its six edges, its four triangular faces, and the tetrahedron itself. By convention, the empty set is considered to be a (-1)-dimensional face of every polyhedron. This article makes no use of this convention, but in some circumstances it is convenient to assume that $\emptyset$ is a member of every nonempty PLC and triangulation.

PLCs and the faces of a facade are defined in a way that gives a geometric model the power to constrain how the boundary of a facade can be triangulated.

Definition 4 (Piecewise Linear Complex; Face of a Facade) An *open piecewise linear complex* (PLC) X is a set containing a finite number of open facades that satisfy the following two restrictions:

- For every facade $f \in X$, the boundary of f is a union of facades in X.[2] For example, X contains both endpoints of every segment in X, and every 2-facade's boundary is a union of segments and vertices in X.

[2]The boundary of a vertex is the empty set, which is a union of zero facades.

- For any two facades $f, g \in X$, $f \cap g = \emptyset$.

The *faces* of a facade f are $\{g \in X : \ g \subseteq \text{closure}(f)\}$. They include f itself and its vertices. The *proper faces* of f are all its faces except f and $\emptyset$.

For any open PLC X, $\{\text{closure}(f) : \ f \in X\}$ is a *closed piecewise linear complex*.

It is possible to reverse the transformation and convert a closed PLC into an open PLC by subtracting from each facade every facade of lower dimension. Hence, for a closed facade in a closed PLC, define the *internal boundary* of the closed facade to be the internal boundary of the corresponding open facade. The internal boundary of a closed facade is not really part of the boundary of the closed facade, and it is defined only in the context of a PLC.

Definition 5 (Triangulation Domain) Let $|X|$ denote the union of facades $\bigcup_{f \in X} f$. $|X|$ is called the *triangulation domain*, or simply the *domain*. (It is also known as the *underlying space of* X.)

A corollary of the definition of PLC is that $\bigcup_{f \in X} f$ is the same for an open PLC and the corresponding closed PLC. Another corollary is that a closed PLC Y satisfies the restrictions that Miller et al. [32] specified when they introduced the notion of a PLC.

- For every facade $f \in Y$, the boundary of f is a union of facades in Y.
- For any two facades $f, g \in Y$, $f \cap g$ is a union of facades in Y. (Usually $f \cap g$ is a single facade or the empty set, but imagine two nonconvex 2-facades that intersect each other at several isolated vertices and along several line segments. Each of these vertices and line segments must be in Y.)
- For any two facades $f, g \in Y$, if $f \cap g$ has the same dimensionality as f, then $f \subset g$, and f is of lower dimensionality than g.

Miller's third restriction is somewhat cryptic; its main effect is to prevent two facades of the same dimensionality from having overlapping relative interiors. The formulation of PLCs in terms of open facades is more elegant, because no similarly cryptic restriction is needed. However, closed PLCs offer a more elegant model for the incremental update of a PLC (discussed in the third article of this series). The insertion or deletion of a facade in a closed PLC can imply several modifications to the corresponding open PLC. For instance, when a vertex is added to an open PLC, if a facade contains the vertex, the facade must have that point removed.

This formal hair-splitting between open and closed facades is necessary because it is the open facades that determine the facade boundaries, but it is the closed facades that occlude visibility, and simplices must respect the closed facades. The rest of this article maintains an uneasy duality, wherein every use of the word "facade" refers to both the open and the closed versions of the facade. Fortunately, the bijective map between open and closed PLCs usually makes it unnecessary to specify which type of PLC is under discussion.

The reader should be aware that every reference to the "boundary of a facade" or the "faces of a facade" regards the boundary of the open facade, including the internal boundary. Similarly, the "relative interior of a facade" refers to the open

facade. However, wherever this article states that a facade contains a point, a facade obstructs visibility, or a simplex respects a facade, the closed facade is implied.

It makes no difference to most of this article's results whether or not facades are connected. An open facade made up of n connected components can be replaced with n separate facades without changing any essential properties of the PLC. Some components of a facade may be grazeable while others are not, so breaking up a facade into its components may improve the prospects for having a weakly ridge-protected PLC. However, there is an important convention for weighted CDTs. If a vertex in a PLC is an endpoint of two collinear segments and is not needed to support their intersection with some other facade, it is usually better to think of the two segments as parts of a single 1-facade, because the vertex might be absent from the weighted CDT. (See Definition 12 in Section 2.3 for details.)

There appear to be few publications exploring the properties of geometric partitions that permit the existence of faces with internal boundaries. An interesting exception by Grünbaum and Shephard [25] shows how to reliably compute Euler characteristics for a class of objects more general than PLCs. One can convert an open PLC into a "relatively open convex dissection" by partitioning its open facades into open convex facades (polyhedra), whereupon its Euler characteristic is easy to calculate. This method is particularly interesting when applied to an open facade with a complicated internal boundary, or to a subset of an open PLC that allows faces of facades to be absent.

The notion of a PLC generalizes to complexes of curved manifolds. For example, every semialgebraic or subanalytic set of points can be partitioned into a *stratification*—a set of *strata* (which generalize open facades), each of which is a manifold. See Gomes [23] for an excellent introduction to the topic.

How might a PLC be represented as a data structure? Here are a few suggestions. A 0-facade (vertex) is represented by its d coordinates. For $j \geq 1$, a j-facade f is most easily represented by a list of its proper faces. To conserve space, f can be represented by a list of every proper face of f that is not a proper face of a proper face of f; the unlisted faces can be inferred by reading the listed faces' lists. This representation differs in several ways from the usual face lattice representation of polyhedra and polyhedral complexes. First, the faces in f's list are not necessarily all $(j-1)$-faces, because f's internal boundary may include lower-dimensional faces that are not included in any $(j-1)$-face. For example, a 2-facade may have an isolated vertex inside it. Second, this representation is technically not a lattice. For example, two 2-facades might intersect at two separate vertices that are included in no other facades, so a pair of facades do not necessarily have a unique meet and join, contrary to the definition of "lattice." See Ziegler [54] for a definition and discussion of face lattices.

Within the affine hull of a j-facade f, each $(j-1)$-face of f has two sides. In an implementation of the sweep algorithm or gift-wrapping algorithm for CDT construction, the list of f's $(j-1)$-faces should include annotations that indicate which side (or sides) of each $(j-1)$-face adjoins f. A $(j-1)$-face on f's internal boundary adjoins f on both sides. If an open $(j-1)$-face is composed of several connected components, it needs one annotation for each side of each connected component.

For the algorithms described in the sequel articles, it is unnecessary to specify the d-facades explicitly as part of the input. Instead, each side of each $(d-1)$-facade should bear an annotation that indicates whether it adjoins the exterior domain or the

interior of the triangulation domain. A $(d-1)$-facade is part of the internal boundary of a PLC if both sides adjoin the triangulation domain, part of the external boundary if one side adjoins the exterior domain, and a dangling facade if both sides adjoin the exterior domain.

Definition 6 (Dangling Facade) Let X be a d-dimensional PLC. A facade in X is a *dangling facade* if it is not a face of any d-facade in X.

To a programmer, the distinction between open and closed facades is almost irrelevant. Any reasonable PLC data structure simultaneously represents both.

2.2 Weighted Delaunay Triangulations

This section reviews known facts about weighted Delaunay triangulations [2] and introduces new terminology as a preliminary to introducing weighted CDTs in Section 2.4. Consider the Euclidean space E^{d+1}, and let $x_1, x_2, \ldots, x_{d+1}$ be the coordinate axes. E^d is the subspace of E^{d+1} orthogonal to the x_{d+1}-axis. In the space E^{d+1}, a d-flat is *vertical* if it includes a line parallel to the x_{d+1}-axis.

Definition 7 (Polyhedral Complex; Triangulation) A *polyhedral complex* T is a set containing a finite number of closed, convex polyhedra that satisfy the following two restrictions:

- For every polyhedron $s \in T$, every face (in the sense of Definition 3) of s is in T.
- For any two polyhedra $s, t \in T$, if s and t are not disjoint, then $s \cap t$ is a face of both s and t.

A *triangulation* or *simplicial complex* is a polyhedral complex whose members are all simplices.

Every polyhedral complex is a PLC. Observe that polyhedral complexes are less general than PLCs whose facades are all convex, because they use a different definition of "face." In a PLC, one side of a tetrahedron might be subdivided into several triangular faces, and a (closed) tetrahedron might have an edge passing through its interior. In a polyhedral or simplicial complex, both circumstances are forbidden: a side of a tetrahedron is represented by exactly one triangular face, and a tetrahedron's interior intersects no other simplex of equal or lesser dimension.

A d-dimensional triangulation or polyhedral complex is *regular* if it is the vertical projection of one "side" of some convex $(d+1)$-polyhedron.

Definition 8 (Downward-Facing; Underside; Regular) Let P be a convex $(d+1)$-polyhedron in E^{d+1}. A face f of P is *downward-facing* if no point in P is directly below any point in f (i.e. having the same x_1- through x_d-coordinates but a lesser x_{d+1}-coordinate). The *underside* of P is the set of all its downward-facing faces.

A d-dimensional triangulation or polyhedral complex is *regular* if it can be formed by vertically projecting the underside of some convex $(d+1)$-polyhedron P into E^d (by dropping the x_{d+1}-coordinate of each vertex).

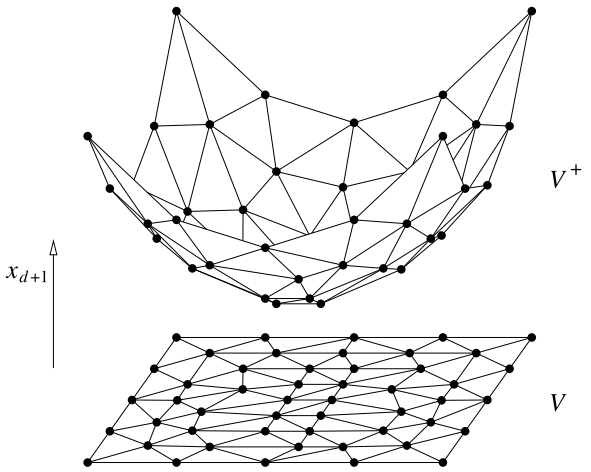


Fig. 12 The parabolic lifting map. In this illustration a two-dimensional vertex set V is lifted to a paraboloid in E^3. The underside of the convex hull of the lifted vertices is a lifted Delaunay triangulation

The best-known regular triangulation is the Delaunay triangulation. The regularity of most Delaunay triangulations is demonstrated by the well-known *parabolic lifting map* of Seidel [18, 42] (inspired by a spherical lifting map suggested by Brown [5]). Let V be a set of vertices in E^d for which a Delaunay triangulation is sought. The lifting map maps each vertex in V to a vertex on a paraboloid in a space one dimension higher, as Fig. 12 illustrates. Specifically, each vertex $v = (v_{x_1}, v_{x_2}, \ldots, v_{x_d}) \in V$ maps to a point $v^+ = (v_{x_1}, v_{x_2}, \ldots, v_{x_d}, v_{x_1}^2 + v_{x_2}^2 + \cdots + v_{x_d}^2)$ in E^{d+1}.

Definition 9 (Companion) The pair of vertices v and v^+ are called *companions*: v^+ is the *lifted companion* of v, and v is the *projected companion* of v^+.

If s is a k-simplex with vertices $v_0, v_1, \ldots, v_k$, then its lifted companion s^+ is the k-simplex embedded in E^{d+1} whose vertices are $v_0^+, v_1^+, \ldots, v_k^+$; and s is the projected companion of s^+. Note that s^+ is flat, and does not curve to hug the paraboloid.

Let $V^+ = \{v^+ : \ v \in V\}$. The Delaunay triangulation of V is regular because it has the same combinatorial structure as the underside of the convex hull of V^+, as the forthcoming Theorem 2 shows. Each downward-facing simplex of $\text{conv}(V^+)$ projects to a Delaunay simplex of V. This connection is routinely used to transform any $(d+1)$-dimensional convex hull construction algorithm into a d-dimensional Delaunay triangulation construction algorithm.

Lemma 1 *Let S be a hypersphere in E^d. Let $S^+ = \{p^+ : \ p \in S\}$ be the ellipsoid found by lifting S to the paraboloid. Then the points of S^+ lie on a non-vertical d-flat h. (Recall that a d-flat is vertical if it is parallel to the x_{d+1}-axis.) Furthermore, a point p inside S lifts to a point p^+ below h, and a point p outside S lifts to a point p^+ above h. Therefore, testing whether a point p is inside, on, or outside S is equivalent to testing whether the lifted point p^+ is below, on, or above h.*

Proof Let O and r be the center and radius of S, respectively. Let p be a point in E^d. The x_{d+1}-coordinate of p^+ is $|p|^2$. By expanding $|O - p|^2$, we have that $|p|^2 = 2O \cdot p - |O|^2 + |Op|^2$.

With O and r fixed and $x \in E^d$ varying, the equation $x_{d+1} = 2O \cdot x - |O|^2 + r^2$ defines a non-vertical d-flat h in E^{d+1}. For every point $p \in S$, $|Op| = r$, so $S^+ \subset h$. For every point $p \notin S$, if $|Op| < r$, then the lifted point p^+ lies below h, and if $|Op| > r$, then p^+ lies above h. □

Theorem 2 [42] *Let s be a simplex whose vertices are in V, and let s^+ be its lifted companion. Then s is Delaunay if and only if s^+ is included in some face of the underside of* $\text{conv}(V^+)$. *The simplex s is strongly Delaunay if and only if s^+ is a face of the underside of* $\text{conv}(V^+)$ *and no vertex in V^+ lies on s^+ except the vertices of s^+.*

Proof If s is Delaunay, there is a circumsphere S of s such that no vertex of V lies inside S. Let h be the unique d-flat in E^{d+1} that includes S^+. By Lemma 1, no vertex in V^+ lies below h. The d-flat h includes s^+ because the vertices of s^+ are in S^+. Therefore, s^+ is included in a downward-facing face of the convex hull of V^+. If s is strongly Delaunay, no vertex in V^+ lies below h, and no vertex in V^+ lies on h except the vertices of s^+. Therefore, s^+ is a downward-facing face of the convex hull of V^+.

The converse implications follow by reversing the argument. □

A *weighted Delaunay triangulation* is like a Delaunay triangulation, but each vertex $v \in V$ is assigned a real-valued *weight* w_v. A vertex v lifts to a companion $v^+ = (v_{x_1}, v_{x_2}, \ldots, v_{x_d}, v_{x_1}^2 + v_{x_2}^2 + \cdots + v_{x_d}^2 - w_v)$. The x_{d+1}-coordinate $|v|^2 - w_v$ is called the *height* of v. The weighted Delaunay triangulation of V is the projection to E^d of the underside of $\text{conv}(V^+)$. It follows that a weighted Delaunay triangulation is regular.

Some faces of $\text{conv}(V^+)$ might not be simplices, because some selection of $d+2$ or more of the lifted vertices might lie on a common non-vertical d-flat. (Observe that vertices that lie on a common vertical d-flat do not cause trouble, because a vertical face cannot be downward-facing. This is good news, because a typical real-world vertex set V includes large groups of cohyperplanar vertices.) These non-simplicial faces can be filled with any compatible triangulation, so V has more than one weighted Delaunay triangulation. However, some faces can be triangulated with triangulations that are not regular, so not all weighted (or unweighted) Delaunay triangulations are regular! Section 6 describes a simple way to perturb the weights to simulate the circumstance that no $d+2$ vertices in V^+ lie on a common non-vertical d-flat.

If its weight is sufficiently small, a lifted vertex v^+ might not be downward-facing—it might not lie on the underside of $\text{conv}(V^+)$—in which case the vertex v is absent from the weighted Delaunay triangulation of V, as illustrated in Fig. 13(a). Then v is said to be *submerged*. If every vertex has a weight of zero, the weighted Delaunay triangulation is the Delaunay triangulation, and no vertex is submerged, because every point on the paraboloid is on the underside of the convex hull of the paraboloid.

Weights necessitate a generalization of the notion of a "Delaunay simplex."

Definition 10 (Semiregular; Witness; Weighted Delaunay Triangulation) A simplex s whose vertices are in V is *semiregular* if s^+ is included in a downward-facing face

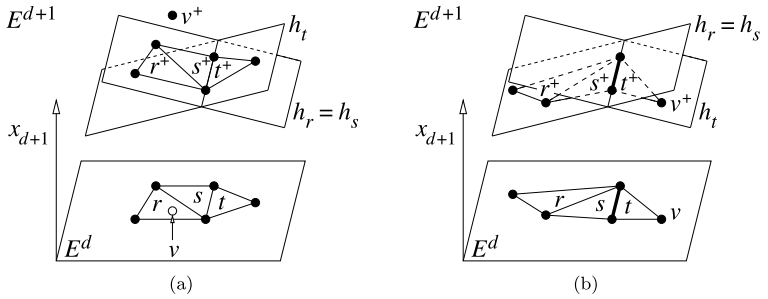


Fig. 13 (a) The triangles r, s, and t are all semiregular, but only t is regular. Triangles r and s have the same witness d-flat $h_r = h_s$, and t has a different witness h_t. The vertex v is submerged. (b) The bold edge is a constraining segment. The triangles r, s, and t are all constrained semiregular, but only t is constrained regular. No triangle is semiregular

of conv(V^+). In other words, there exists a non-vertical d-flat $h_s \subset E^{d+1}$ such that h_s includes s^+, and no vertex in V^+ lies below h_s. The d-flat h_s is called a *witness*[3] to the semiregularity of s.

A *weighted Delaunay triangulation* of V is a simplicial complex that fills conv(V) wherein every simplex is semiregular.

Figure 13(a) illustrates three semiregular triangles and their witnesses. All their edges and vertices are semiregular as well, but the submerged vertex v is not semiregular.

Definition 11 (Regular) A simplex s is *regular* if s^+ is a downward-facing face of conv(V^+), and no vertex in V^+ lies on s^+ except the vertices of s^+. In other words, there exists a non-vertical d-flat $h_s \subset E^{d+1}$ that is a *witness* to the regularity of s: h_s includes s^+, and every vertex in V^+ lies above h_s, except the vertices of s^+.

A triangulation is *regular* if there exists an assignment of weights to its vertices for which every simplex is regular.

Of the three triangles in Fig. 13(a), only t is regular. All the edges are regular except the edge shared by r and s. All the vertices are regular except v.

In a weighted Delaunay triangulation, a witness serves the same purpose that a circumsphere serves in an ordinary Delaunay triangulation. Theorem 2 shows that if all the weights are zero, "semiregular" is equivalent to "Delaunay" and "regular" is equivalent to "strongly Delaunay." If a simplex s is semiregular, it appears in at least one weighted Delaunay triangulation of V. If s is regular, it appears in *every* weighted Delaunay triangulation of V (see Theorem 19).

2.3 Triangulations of PLCs

For some geometric applications, the notion of a "constrained triangulation" of a PLC should permit some vertices to be left out, just as weighted Delaunay triangulations

[3] A witness for a semiregular or regular simplex is also known as a *supporting hyperplane* of conv(V^+), but a witness for a *constrained* semiregular simplex is not necessarily a supporting hyperplane.

submerge vertices with insufficient weight. However, some vertices cannot be omitted, because they support other facades. The following definition identifies vertices that could conceivably be submerged.

Definition 12 (Submersible) A vertex v in a closed PLC X is *submersible* if v is a proper face of some other facade (i.e. v is not isolated), and the removal of v from X (and possibly the merging of two collinear 1-facades) yields a valid closed PLC. Equivalently, either

- v lies on the internal boundary of a facade $f \in X$ such that f is a face of every facade (except v) that contains v, or
- v is an endpoint of two collinear 1-facades in X, and the condition above is satisfied by merging them into a single 1-facade f. In this case, X should be modified to reflect the merger. A row of collinear segments might comprise one 1-facade with many submersible vertices in it.

The user of a PLC triangulation algorithm can arbitrarily designate vertices as being non-submersible, but a vertex can be designated as submersible only if Definition 12 permits it.

Definition 13 (Fill; Respect; Triangulation of a PLC) Let T be a set of simplices. T *fills* X if $|T| = |X|$, meaning that $\bigcup_{s\in T} s = \bigcup_{f\in X} f$.

Let f be a closed facade. Let s be a simplex or convex polyhedron. Then s *respects* f if $s \cap f$ is a union of faces of s.

There is an equivalent definition that is less clear, but easier to use in proofs: s respects f if, for every face t of s whose relative interior intersects f, $t \subseteq f$.

If f is an open facade, s is said to respect f if s respects the closure of f.

A simplex (or convex polyhedron) s *respects* a PLC X if $s \subseteq |X|$ and s respects every facade in X except perhaps the submersible vertices—after agglomerating the segments of X into 1-facades as described in Definition 12.

A triangulation T respects a PLC X if every simplex in T respects X.

A triangulation T is a *triangulation of a PLC* X if T fills and respects X, and T has no vertex not in X. A triangulation that fills and respects X, but may have vertices not present in X, is a *conforming triangulation* or *Steiner triangulation* of X.

This definition allows a triangulation T of X to submerge vertices in X. However, submersibility is a nuisance when it is not needed. For some applications, such as unweighted PLCs and ordinary CDTs (in which vertices are never submerged), it does no harm to designate every vertex in X as non-submersible. Then Definition 13 implicitly requires that if T is a triangulation of X, then T and X have exactly the same vertices, because T must respect every vertex in X.

Why must adjoining collinear segments be agglomerated for Definition 13? If a vertex is submerged, then a triangulation lacking that vertex cannot respect a segment that terminates at that vertex, but it can respect a 1-facade that passes through the vertex.

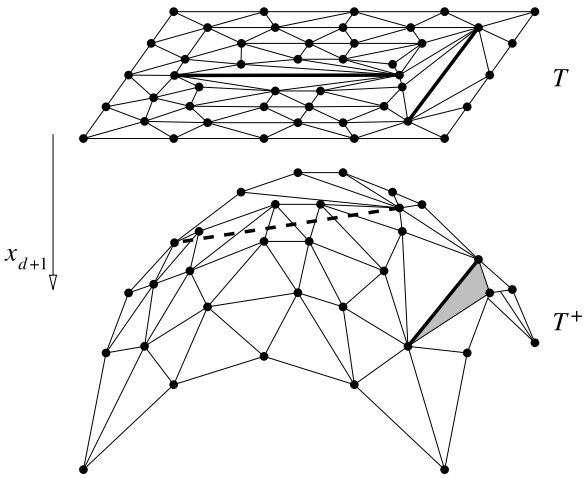


Fig. 14 A lifted CDT. The paraboloid is inverted to show its topography more clearly. The bold edges are constraining edges that are not Delaunay. They map to reflex edges of the lifted surface

2.4 Weighted CDTs

Before considering the formal definition of CDT, let us try to see intuitively what a CDT is, in terms of the parabolic lifting map. Suppose T is a CDT of a PLC X. Let $T^+ = \{s^+ : \ s \in T\}$ be the simplicial complex, embedded in E^{d+1}, defined by lifting T. As Fig. 14 illustrates, the lifted triangulation T^+ graphs a continuous piecewise linear function but, in general, is not the underside of a convex polyhedron: each facet of the CDT that is not constrained Delaunay is mapped to a reflex ridge in the lifted surface. (A $(d-1)$-simplex is called a *facet* if it exists in the ambient space E^d, and a *ridge* if it exists in the ambient space E^{d+1}.)

However, from any point p in the interior of a d-facade, the portions of the CDT visible from p appear convex on the lifting map. Only facets included in $(d-1)$-facades can lift to reflex ridges; every other facet is constrained Delaunay.

The next several definitions build toward the definition of a CDT or, more generally, a weighted CDT, which is a triangulation of a weighted PLC.

Definition 14 (Weighted PLC) A *weighted PLC* is a PLC in which each vertex is assigned a real-valued weight.

Sections 3.1 and 3.3 study the relationship between the weighted CDT of a weighted PLC and the weighted CDTs of its facades. Consider computing a triangulation of a two-dimensional PLC. Some algorithms need to "triangulate" the 1-facades of the PLC first—in other words, to decide which vertices on the 1-facades are submerged. The 1-facades may have both submersible and non-submersible vertices. A 1-facade in isolation does not reveal which of its vertices are submersible in the two-dimensional PLC. Therefore, it is best to think of submersibility as a global property of a vertex which remains fixed across all contexts, and is determined by the highest-dimensional PLC that contains the vertex. These observations motivate the following two policies. First, the internal boundary of a 1-facade may contain both submersible and non-submersible vertices (whereas the external boundary is a set of

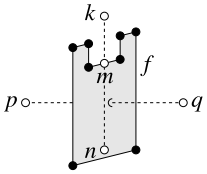


Fig. 15 In this three-dimensional example the 2-facade f occludes the visibility between p and q. The point m can see both k and n, but the visibility between k and n is occluded—not by f, but by an edge of f

non-submersible vertices). Second, non-submersible vertices occlude visibility and submersible vertices do not. This policy ensures that the weighted CDT of a 1-facade is consistent with the weighted CDT of any higher-dimensional facade that includes the 1-facade.

Visibility is occluded by constraining facades.

Definition 15 (Constraining Facade) A *constraining facade* in a d-dimensional PLC X is any facade in X that is not a submersible vertex or a d-facade.

Definition 15 omits submersible vertices because they do not occlude visibility or constrain the triangulation. It omits d-facades because they do not occlude visibility, and because a simplex or polyhedron that respects all the lower-dimensional facades automatically respects the d-facades.

Definition 16 (Occlusion; Visibility) Within a PLC X, the visibility between two points p and q is *occluded* if $pq \not\subseteq |X|$; or if there is a (closed) constraining facade $f \in X$ such that the line segment pq intersects f, and neither p nor q lie on the affine hull of f. See Fig. 15. The points p and q are *visible* from each other (equivalently, can *see* each other) if $pq \subseteq |X|$ and X places no constraining facade between them.

If no vertex is submersible, a more elegant characterization is that p and q can see each other if there is an open facade $f \in X$ that includes the open line segment pq. Open facades thus act as conductors of visibility. In this interpretation the d-facades play an essential role.[4]

There is a close relationship between visibility and the notion of respecting a PLC.

Theorem 3 *If a* (*closed*) *simplex or convex polyhedron s respects X, every point in s can see every other point in s.*

Proof Suppose for the sake of contradiction that the visibility between two points $p, q \in s$ is occluded by some facade f. Then pq intersects f at a point m, but f

[4] An attractive alternative formulation of a weighted PLC extends this characterization to PLCs with submersible vertices. Express a weighted PLC as two separate sets: a PLC X with no submersible vertices, and a set V of submersible vertices. In this formulation the open facades of X are both conductors and occluders of visibility, and there is a more elegant definition of "respect": a triangulation respects X if every *open* simplex of the triangulation is included in an *open* facade of X.

contains neither p nor q. Let t be the face of s whose relative interior contains m; then $pq \subseteq t$. Because s respects f, and f intersects the relative interior of a face t of s, it follows that $t \subseteq f$, contradicting the fact that f contains neither p nor q. □

Simplices in CDTs have the following property.

Definition 17 (Constrained Semiregular) A simplex s is *constrained semiregular* within X if

- the vertices of s are in X,
- s respects X, and
- there exists a d-flat $h_s \subset E^{d+1}$ that includes s^+, such that no vertex $v \in X$ that is visible from a point in the relative interior of s lifts to a point v^+ below h_s. The d-flat h_s is a witness to the constrained semiregularity of s.

The third condition is a bit difficult to visualize, because one must simultaneously picture the vertices in the ambient space E^d where visibility is determined, and in the ambient space E^{d+1} where witness d-flats are defined, as Fig. 13(b) illustrates. Think of it this way: if some lifted vertex v^+ lies below the d-flat that includes a lifted d-simplex s^+, then s is not semiregular, because s^+ is not on the underside of the convex hull of the lifted vertices. However, if some facade occludes the view of v from inside s, s may be constrained semiregular anyway and appear in the weighted CDT. The triangle s in Fig. 13(b) is an example: although v^+ lies below the witness h_s, v is not visible from the interior of s, so s is constrained semiregular. The shaded triangle in Fig. 14 is an example in an unweighted CDT (but note that the paraboloid in the figure is inverted for clarity, so "below" is "above").

In Fig. 13(b) all three triangles are constrained semiregular, and all the edges are constrained semiregular except the bold constraining segment.

Definition 18 (Constrained Regular) A simplex s is *constrained regular* within X if

- the vertices of s are in X,
- s respects X, and
- there exists a d-flat $h_s \subset E^{d+1}$ that includes s^+, such that every vertex $v \in X$ that is visible from a point in the relative interior of s, but is not a vertex of s, lifts to a point v^+ above h_s.

Of the three triangles in Fig. 13(b), only t is constrained regular. Neither the edge shared by r and s nor the constraining segment shared by s and t is constrained regular, but the other edges are.

The following implications hold. Statements in brackets are equivalent to the statements immediately above them in the unweighted case (i.e. when all the vertex weights are zero). *Locally semiregular* and *locally regular* are defined in Section 3.2 and apply to $(d-1)$-simplices only.

s is regular and respects X [s is strongly Delaunay and respects X]	$\longrightarrow$	s is semiregular and respects X [s is Delaunay and respects X]
$\downarrow$		$\downarrow$
s is constrained regular	$\longrightarrow$	s is constrained semiregular [s is constrained Delaunay]
$\downarrow$		$\downarrow$
s is locally regular	$\longrightarrow$	s is locally semiregular [s is locally Delaunay]

The statements in the right column become equivalent to the corresponding statements in the left column when the following condition holds. (Section 6 discusses a perturbation technique that enforces it.)

Definition 19 (Genericity) A d-dimensional PLC X is *generic* if no $d+2$ vertices in X lift to a common non-vertical d-flat (in the ambient space E^{d+1}).

If X is unweighted (or all the weights are equal), an equivalent statement is that no $d+2$ vertices in X lie on a common hypersphere (in the ambient space E^d).

Notions like constrained regularity are defined in the context of a specific PLC. The definition of "CDT" uses the notion that a simplex can be constrained semiregular within the context of some facade f of a PLC X, yet not be constrained semiregular within the context of X itself.

Definition 20 (Facade PLC) Let f be a k-facade in a PLC X (for any value of k). The *facade PLC* Y_f is a k-dimensional PLC containing f and all the faces of f (taken from X).

The vertices in a facade PLC often have coordinates from an ambient space E^d whose dimensionality is higher than that of the facade PLC itself (i.e. $d > k$). However, it is the latter dimensionality that defines constraining facades (facades of dimension $k-1$ or less that are not submersible vertices) and ridge protection (the protection of facades of dimension $k-2$ or less; see Definition 23) within Y_f. A simplex that is regular within Y_f might not be regular within X, and a segment that is grazeable within X might not be grazeable within Y_f. Hence, the word *within* is used wherever the context is not clear. Occasionally, this article will say that a simplex is "semiregular within the facade f" as shorthand for saying it is semiregular within the facade PLC Y_f. Likewise, a "triangulation of f" is a triangulation of Y_f.

At last, a definition of this article's central object of study.

Definition 21 (Weighted CDT) A *weighted constrained Delaunay triangulation* of a weighted PLC X is a simplicial complex that fills X wherein every simplex is constrained semiregular within the lowest-dimensional facade of X that includes it.

A *constrained Delaunay triangulation* of an unweighted PLC is a weighted CDT for which all the vertices in the PLC are implicitly assigned a weight of zero.

Figure 16 gives two examples of weighted CDTs, in one and two dimensions. In both triangulations, some vertices are submerged, and some collinear segments of the PLC are agglomerated into single edges of the triangulation. Observe that the lifted

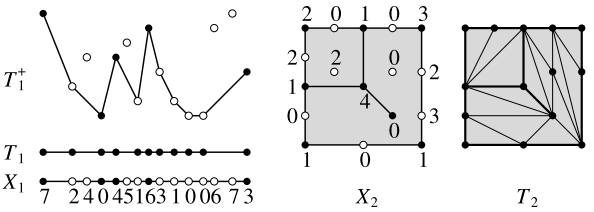


Fig. 16 T_1 and T_2 are weighted CDTs of the one- and two-dimensional weighted PLCs X_1 and X_2. White vertices are submersible; black vertices are non-submersible. The number by each vertex is the height (x_{d+1}-coordinate) to which it is lifted

one-dimensional triangulation T_1^+ is a sequence of convex hull undersides separated by non-submersible vertices. Note that $d = 1$ is the only dimensionality in which a PLC might have a new CDT if a vertex changes from submersible to non-submersible. For a higher-dimensional PLC with no dangling 1-facades, such a change might cause the PLC to have fewer CDTs (if a submerged vertex is proclaimed non-submersible), but it cannot cause the PLC to have a CDT it did not have before. (This claim is a consequence of the Delaunay Lemma in Section 3.2.)

In an unweighted CDT X (equivalently, if all the weights are equal), every vertex is regular and constrained regular, hence no vertex is submerged.

Definition 21 gives no reason to believe that the eligible simplices (those that are constrained semiregular within the lowest-dimensional facades that include them) can gel together to form a complex. Fortunately, if every facade can be filled with constrained regular simplices, Corollary 18 in Section 3.3 establishes that the facade CDTs match each other where they meet. Not every facade can be thus filled (recall Schönhardt's polyhedron). The next few definitions describe a class of PLCs that are guaranteed to have CDTs.

Definition 22 (Grazeable; Grazing Triangle) A facade f is *grazeable* if there is an open *grazing triangle* $L = \triangle pqr \subset |X|$ such that

- p can see every point in the open triangle L,
- pq intersects the open version of f (i.e. f with its external and internal boundaries removed), and
- neither p nor q lie on the affine hull of f.

Every point in an open grazing triangle $\triangle pqr$ is visible from p, but q is not (its visibility is occluded by f); so, loosely speaking, there is a line of visibility that grazes f. If f is a $(d-2)$-facade, Definition 22 is equivalent to the 180° angle condition described in Section 1.1, as Fig. 17 shows. Definition 22 extends the idea to facades of dimension less than $d-2$. Note that the proper faces of a grazeable facade are not necessarily grazeable themselves.

Recall from Section 1.1 that a three-dimensional PLC X is *ridge-protected* if every segment in X is strongly Delaunay. The extension of this definition to weighted PLCs accounts for the possibility that vertices might be submerged: X is ridge-protected if every 1-facade is a union of regular edges, and every non-submersible vertex is regular. (Submersible vertices do not need to be regular, because it is okay to let them be submerged.) The extension of this definition to higher dimensions requires that all constraining facades of dimension $d-2$ or less be "regular," but the definition of "regular" applies only to simplices. It suffices if the facades can be broken up into regular simplices that respect X.

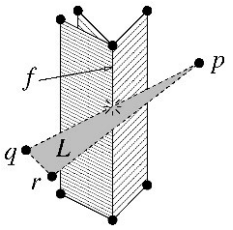


Fig. 17 Example of a grazeable segment f. Here p and q cannot see each other, but p sees every point in the open triangle L, so there is a line of visibility that grazes f

Definition 23 (Ridge Protection) A facade $f \in X$ is *protected* if there exists a triangulation of f whose simplices are regular within X and respect X.

A simpler definition is that f is *protected* if f is a union of simplices that are regular within X and respect X. (The equivalence of this definition with the first follows from the upcoming Theorem 4′ and Corollary 18.)

A weighted PLC X is *weakly ridge-protected* if every grazeable constraining facade in X of dimension $d-2$ or less is protected.

X is *ridge-protected* if *every* constraining facade in X of dimension $d-2$ or less is protected.

How can you tell if a facade f is protected? A weighted Delaunay triangulation T (unconstrained) of the vertices in X contains every simplex that is regular within X (by Theorem 19 in Section 3.3). So the answer is to construct T and search it for a subset of faces that fill f. If T contains such faces, check whether they respect f's faces and are regular. If X is not generic, the trickiest part is distinguishing the regular simplices in T from the merely semiregular. Dafna Talmor (personal communication) points out that simplices that are semiregular but not regular dualize to degenerate faces of the power diagram [2] (the Voronoi diagram if all the weights are zero). This observation does not offer the most numerically effective way to test them, though, and this is not the place to describe a better way. However, the simplest approach is to perturb the vertex weights as described in Section 6 before constructing T. Then all the simplices in T are regular, and there is no need to test. Theorem 31 in Section 6 shows that the CDT of the perturbed PLC is a CDT of the unperturbed PLC.

Ridge protection implies that T respects all the constraining k-facades in X for $k \leq d-2$, but might not respect the $(d-1)$-facades. Weak ridge protection implies that T respects the grazeable constraining facades of dimension $d-2$ or less (and their faces, whether grazeable or not), but perhaps not the other facades. One of the main results of this article is that every weakly ridge-protected weighted PLC has a weighted CDT, so the missing facades can be recovered without any need for additional vertices. See Section 5 for a proof.

Ridge protection requires non-submersible vertices to be regular. For $d = 2$, this is the sole requirement that defines ridge protection. In an unweighted PLC, every vertex is regular, which is why every unweighted two-dimensional PLC has a CDT. In the weighted PLC X_2 in Fig. 16, the sole grazeable non-submersible vertex is regular, so X_2 is weakly ridge-protected and has a CDT. (The vertex at the center of X_2 is not regular, but it is not grazeable.) Figure 18 depicts two two-dimensional weighted PLCs that are not weakly ridge-protected, and do not have weighted CDTs. Both examples include a grazeable non-submersible vertex that is not regular.

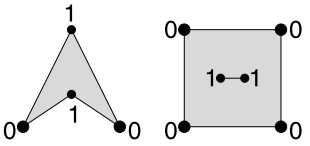


Fig. 18 Two weighted PLCs that do not have weighted CDTs. Imagine that you are viewing the lifted vertices from directly underneath, and larger vertices are closer to you. The number by each vertex is the height (x_{d+1}-coordinate) to which it is lifted (i.e. its distance from you)

A recently proposed way to model domains like these ones and Schönhardt's polyhedron is to generalize simplicial complexes to *pseudosimplicial complexes* composed of nonconvex pseudosimplices. Aichholzer et al. [1] define constrained regular pseudotriangulations that generalize the two-dimensional constrained regular triangulations defined here, and are defined for every choice of vertex weights. Their lifted surface is not necessarily continuous, and is not guaranteed to interpolate all the vertex heights. Aurenhammer and Krasser [3] show that the approach generalizes to higher-dimensional nonconvex polyhedra, but pseudosimplicial complexes representing polyhedra in three dimensions or more must sometimes introduce additional vertices.

Throughout the rest of this article, the terms "PLC" and "CDT" refer to both unweighted and weighted PLCs and CDTs, except where otherwise noted.

3 Foundations

This section proves several fundamental properties of CDTs and weighted CDTs. Among these are the fact that every face of a constrained semiregular simplex is constrained semiregular within some facade (Section 3.1), the fact that constrained regular simplices have disjoint relative interiors and form a complex, and the fact that a generic PLC has at most one CDT (Section 3.3). The Delaunay Lemma offers a powerful alternative characterization of what a CDT is (Section 3.2). Readers who seek the minimum background for understanding the CDT construction algorithms in the sequel articles may safely skip to Section 6.

3.1 Faces of Simplices Inherit Semiregularity and Constrained Semiregularity

CDTs (unweighted and weighted) have properties that allow proofs and algorithms to work in a top-down fashion: if a domain can be filled with a complex of constrained semiregular d-simplices, the lower-dimensional faces "work themselves out."

Let s be a simplex that is constrained semiregular within some PLC X. Let t be a face of s. If t is not included in a constraining facade in X, then t is also constrained semiregular. What if t *is* included in a constraining facade? Then t might not be constrained semiregular within X, but t is constrained semiregular within the lowest-dimensional facade that includes t (and is not a submersible vertex). It follows that the act of filling a d-facade with a complex of constrained semiregular d-simplices automatically fills all of its proper faces with lower-dimensional CDTs.

First consider unconstrained semiregularity.

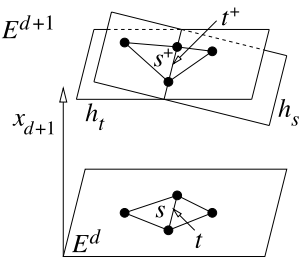


Fig. 19 The simplex s is regular because every lifted vertex lies above some witness d-flat h_s for s^+, except the vertices of s^+. Let t be any face of s. Tilting h_s using t^+ as a hinge yields a witness d-flat h_t that shows that t is regular too

Theorem 4 *Every face of a semiregular simplex is semiregular.*

Theorem 4′ *Every face of a regular simplex is regular.*

Proof Let s be a semiregular simplex, and let t be a face of s as in Fig. 19. Let h_s be a witness to the semiregularity of s. That is, h_s is a d-flat that includes s^+, and no vertex in X lifts to a point below h_s. Clearly, h_s is also a witness to the semiregularity of t, so Theorem 4 holds.

Suppose s is regular. Then every vertex in X lifts above h_s except the vertices of s. Let h_t be a d-flat found by tilting h_s by a tiny amount (as illustrated), so that h_t includes t^+ but lies below the vertices of s^+ not shared by t^+. If the tilt is small enough, the other vertices in X still lift to points above h_t. Hence, h_t is a witness to the regularity of t, and Theorem 4′ holds. □

Theorem 5 *Let X be a PLC. Let s be a simplex, and let t be a face of s that is not included in a (closed) constraining facade in X. If s is constrained semiregular, then t is constrained semiregular.*

Theorem 5′ *Under the assumptions of Theorem* 5, *if s is constrained regular, then t is constrained regular.*

Proof Because s is constrained semiregular, s respects X. As t is a face of s, t also respects X.

Observe that every vertex visible from the relative interior of t is visible from the relative interior of s. Specifically, suppose a vertex v is visible from a point p in the relative interior of t. Because p does not lie in a constraining facade in X, Lemma 6 below implies that some point p' in the relative interior of s sees v.

The rest of the proof is identical to the proof of Theorems 4 and 4′, except that only vertices visible from the relative interior of t are considered, and every occurrence of "semiregular" or "regular" is thus replaced with "constrained semiregular" or "constrained regular." □

The following lemma (which is used frequently in this article) tells us a way to perturb a point p without occluding its visibility from another point q.

Definition 24 (ϵ-Neighbor) A point p' is an ϵ-neighbor[5] of a point p, with respect to a point q and a closed PLC X, if $p' \in |X|$, $|pp'| \leq \epsilon$, and every (closed) constraining facade in X that contains p contains either p' or q.

Lemma 6 *Let p and q be two points that can see each other within a PLC X. There is a positive constant ϵ such that every ϵ-neighbor of p can see q.*

Proof Any facade whose affine hull contains q cannot occlude the visibility between p' and q. Every facade that contains p contains either p' or q, and thus cannot occlude the visibility between p' and q.

What about the other facades? The line segment pq does not intersect any of them. There is a finite gap between pq and any facade that does not intersect pq, and p must move some non-infinitesimal distance to close the gap. A sufficiently small choice of ϵ ensures that every ϵ-neighbor of p is visible from q. □

Observe that if p lies in a constraining facade f, but p' and q do not, then p' is not an ϵ-neighbor of p, and f might occlude the visibility between p' and q.

Next, consider the circumstance where a face of a simplex is included in a constraining facade. The case of a semiregular simplex is considered first (that's *un*constrained semiregular, albeit in the context of a PLC), followed by the case of a constrained semiregular simplex.

Theorem 7 *Let s be a simplex, and let t be a face of s. Suppose a constraining k-facade $f \in X$ includes t. Let Y_f be the k-dimensional facade PLC for f (recall Definition* 20).

If s is semiregular within X, then t is semiregular within Y_f.

Theorem 7′ *Under the assumptions of Theorem* 7, *if s is regular within X, then t is regular within Y_f.*

Proof For intuition's sake, consider first the special case where s is a Delaunay tetrahedron, illustrated in Fig. 20. No vertex lies inside the circumsphere of s. If a triangular face t of s lies within a 2-facade f, then t is Delaunay within the two-dimensional PLC Y_f. Why? Because the circumcircle of t is a cross section of the circumsphere of s, and therefore it encloses no vertex. If s is strongly Delaunay, t is strongly Delaunay.

Figure 21 extends this reasoning to weighted CDTs. Let s be a semiregular simplex. There is a witness d-flat h_s that includes s^+ such that no lifted vertex lies below h_s. Because the face t of s is included in a k-facade f, h_s yields a witness to the fact that t is semiregular within Y_f as follows.

Let F be the affine hull of f. Think of F as the affine space in which Y_f is defined. Let $F^+ = \{\langle p, \alpha \rangle \in E^{d+1} : p \in F, \alpha \in \mathbb{R}\}$. F^+ is a vertical $(k+1)$-flat in E^{d+1}, as Fig. 21 shows. Think of F^+ as the affine space in which witnesses for Y_f are defined. Then $h_t = h_s \cap F^+$ is a witness k-flat within F^+ that includes t^+. Because no vertex

[5] It would be more apt to call this an (ϵ, q, X)-neighbor of p, but it would clutter the writing.

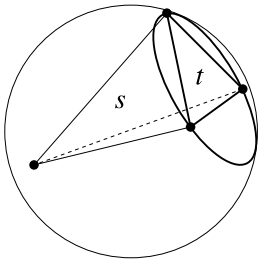


Fig. 20 An unweighted example where $d = 3$. If a tetrahedron s is Delaunay, each of its faces has an empty circumcircle, because each face's circumcircle is a cross section of the tetrahedron's circumsphere

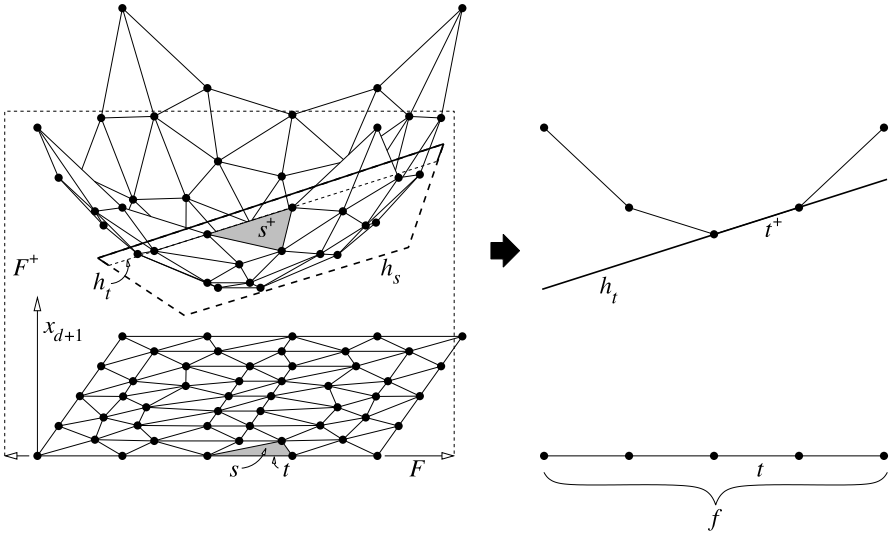


Fig. 21 A weighted example where $d = 2$. If a simplex s is semiregular (no lifted vertex lies below h_s), any face t of s that lies in a facade f is semiregular within f (no lifted vertex lies below h_t)

in X lifts to a point below h_s, no vertex in Y_f lifts to a point below h_t, so t is semiregular within Y_f and Theorem 7 holds.

If s is regular, the lifted companion of every vertex in X lies above h_s, except the vertices of s^+ (which lie on h_s). Thus the lifted companion of every vertex in Y_f lies above h_t, except the vertices of s^+. If every vertex of s in Y_f is also a vertex of t, then h_t is a witness to the regularity of t within Y_f. Otherwise, h_t contains at least one vertex of s^+ that is not a vertex of t^+, but that is no obstacle. By tilting slightly as described in the proof of Theorem 4′, h_t becomes a witness to the regularity of t within Y_f. Thus Theorem 7′ holds. □

The next theorem generalizes Theorem 5, and is the constrained analog of Theorem 7.

Theorem 8 *Let s be a simplex, and let t be a face of s. Let f be the lowest-dimensional facade in X that includes t and is not a submersible vertex.*

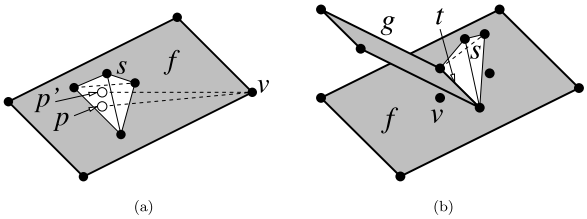


Fig. 22 (a) Example in which f is a 2-facade inside a three-dimensional PLC. The tetrahedron s (which is not a facade) intersects f at a triangular face of s. The point p lies in the relative interior of the triangular face, and p' lies in the interior of s. Both p and p' can see v. (b) Here s intersects f at an edge t. Although v is visible from every point on t, v is not visible from inside s

If s is constrained semiregular within X, then t is constrained semiregular within Y_f.

Theorem 8′ *Under the assumptions of Theorem* 8, *if s is constrained regular within X, then t is constrained regular within Y_f.*

Proof Because s is constrained semiregular, s respects X. As t is a face of s, t also respects X. Moreover, $t \subseteq f = |Y_f|$, so t respects Y_f.

Let v be any vertex in Y_f that is visible from some point p in the relative interior of t. The following reasoning shows that v is also visible from some point in the relative interior of s. See Fig. 22(a). As p can see v, Lemma 6 guarantees that there is an $\epsilon > 0$ such that every ϵ-neighbor of p can also see v. Because t respects f's faces and f is the lowest-dimensional non-submersible facade that includes t, the relative interior of t does not intersect any proper face of f, with the possible exception of submersible vertices. Therefore, every constraining facade that contains p has f for a face and contains v as well. It follows that every point in $|X|$ within a distance of ϵ from p is an ϵ-neighbor of p (with respect to v). Because p is on the boundary of s, v is visible from some point in the relative interior of s.

The rest of the proof is identical to the proof of Theorems 7 and 7′, except that only vertices that are in Y_f and visible from the relative interior of t are considered, and every occurrence of "semiregular" or "regular" is thus replaced with "constrained semiregular" or "constrained regular." □

Figure 22(b) demonstrates why Theorems 8 and 8′ do not apply if f is not the lowest-dimensional facade (other than a submersible vertex) that includes t. In this example, an edge t of a tetrahedron s is a constraining segment on the internal boundary of a 2-facade f. Although s is constrained regular within X, t is not constrained regular within Y_f, because the vertex v is visible from every point on t. The 2-facade g occludes the visibility of v from every point in the interior of s, allowing s to be constrained regular.

The next two theorems simplify proving that a triangulation is a CDT, by putting the burden on the highest-dimensional simplices.

Theorem 9 *Let X be a d-dimensional PLC with no dangling facades* (*i.e. each facade in X is included in a d-facade in X*). *Let T be a simplicial complex that fills*

X. Suppose every d-simplex in T is constrained semiregular. Then T is a CDT of X. Furthermore, for every facade f in X except submersible vertices, $\{t \in T : t \subseteq f\}$ is a CDT of Y_f.

Proof By the definition of "constrained semiregular," every d-simplex in T respects X; and as T is a simplicial complex with no dangling simplices, every simplex in T respects X.

Let t be any simplex in T. Let f be the lowest-dimensional facade in X that includes t. If f is a vertex, then $t = f$ and t is trivially constrained semiregular within Y_f. Otherwise, let s be a d-simplex in T having t for a face. (Some such d-simplex must exist, because T is a simplicial complex filling a PLC with no dangling facades.) By assumption, s is constrained semiregular, so t is constrained semiregular within Y_f by Theorem 8.

Therefore, every simplex in T is constrained semiregular within the lowest-dimensional facade that includes it. By definition, T is a CDT of X. Because T fills and respects X, for every non-submersible facade $f \in X$, the subcomplex $\{t \in T : t \subseteq f\}$ fills and respects Y_f, and thus is a CDT of Y_f. □

The next theorem generalizes Theorem 9 to cover PLCs of mixed dimensionality.

Theorem 10 *Let X be a PLC (possibly with dangling facades). Let T be a simplicial complex that fills X. Suppose that for every $k \geq 1$, for every k-facade $f \in X$ that is not a face of a higher-dimensional facade, every k-simplex of T included in f is constrained semiregular within Y_f. Then T is a CDT of X. Furthermore, for every facade f in X except submersible vertices, $\{t \in T : t \subseteq f\}$ is a CDT of Y_f.*

Proof Identical to the proof of Theorem 9, except that s is a k-simplex in T having t for a face, where k is the dimensionality of the highest-dimensional facade that includes t. (By assumption, s is constrained semiregular within the k-facade that includes s.) □

This section concludes with two corollaries of Theorem 7′.

Corollary 11 *If X is ridge-protected, every facade in X is ridge-protected.* (*That is, if $f \in X$, then its facade PLC Y_f is ridge-protected.*)

Proof Ridge protection holds trivially for a PLC of dimension less than two, so let f be any facade in X of dimension $k \geq 2$. Let Y_f be f's facade PLC. Let d be the dimensionality of X. Because X is ridge-protected and $Y_f \subseteq X$, every constraining facade in Y_f of dimension $d - 2$ or less has a triangulation whose simplices respect X and are regular within X. By Theorem 7′, these simplices are also regular within Y_f. Therefore, every constraining facade in Y_f of dimension $k - 2$ or less has a triangulation whose simplices respect Y_f and are regular within Y_f. □

Corollary 12 *If X is weakly ridge-protected, every facade in X is weakly ridge-protected.*

Proof Let f be a facade in X, and let Y_f be f's facade PLC. It is apparent from Definition 22 that if a face of f is grazeable within Y_f, then the face is grazeable within X too. The rest of the proof is identical to the proof of Corollary 11, except that only every *grazeable* constraining facade in Y_f of dimension $d-2$ or less has a triangulation whose simplices respect X (and therefore Y_f) and are regular within X (and therefore within Y_f). □

3.2 The Delaunay Lemma

A well-known and important property of Delaunay triangulations is that "local optimality" is equivalent to "global optimality," in the following sense. A facet shared by two d-simplices s and t is said to be *locally Delaunay* if the apex of s (not shared by t) is not inside the circumsphere of t (equivalently, the apex of t is not inside the circumsphere of s). If a triangulation is Delaunay, every facet of the triangulation is locally Delaunay. Conversely, if every facet of a triangulation of a point set is locally Delaunay, then the triangulation is Delaunay (i.e. every simplex is Delaunay). Boris Delaunay [14] himself was the first to make this observation.

This section shows that this equivalence generalizes to weighted CDTs, with the change that facets included in constraining facades need not be locally Delaunay (or locally semiregular). This result is valuable because it provides an inexpensive way to test whether a triangulation is a weighted CDT: check that it fills and respects the PLC, check every non-constraining facet for local semiregularity, and check each submerged vertex to ensure it really should be submerged. (A *non-constraining facet* is a facet that is not included in a constraining facade.) The Delaunay Lemma offers an alternative answer to the question, "What does it mean for a PLC X to have no CDT?" It means that no triangulation of X fulfills these requirements.

Definition 25 (Locally Regular; Locally Semiregular) Let T be a triangulation, and let s and t be two d-simplices in T that share a facet f. The facet f is *locally regular* within T if the lifted d-simplices s^+ and t^+ adjoin each other at a dihedral angle, measured from above, of less than 180°. In other words, the apex of t^+ lies above the witness d-flat of s, and vice versa, as illustrated in Fig. 13(a).

The facet f is *locally semiregular* within T if the upper dihedral angle where s^+ meets t^+ is less than or equal to 180°. In other words, either f is locally regular, or s and t have the same witness d-flat.

If a facet f is constrained regular, then f is locally regular, because the spices of s^+ and t^+ lie above some witness d-flat of f. If f is constrained semiregular, f is locally semiregular.

Theorem 13 (Delaunay Lemma) *Let X be a* (*weighted*) *PLC with no dangling facades. A triangulation T is a* (*weighted*) *CDT of X if and only if T has the following four properties*:

A. *T fills X.*
B. *T respects X.*
C. *Every facet in T is either locally semiregular or included in a constraining $(d-1)$-facade of X.*

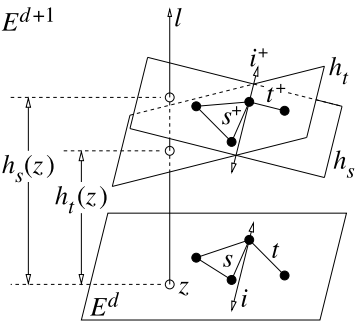


Fig. 23 If s overlaps t from the viewpoint z, then $h_s(z) > h_t(z)$

D. *If a vertex v in X is missing from T (submerged), then v is in a d-simplex s of T such that v^+ lies on or above s^+.*

If X is unweighted, Property D reads, "No vertex is submerged."

The proof of the Delaunay Lemma relies on a lemma that is worth stating separately because it is reused in Sections 3.3 and 5.1. The lemma uses the following definitions.

Definition 26 (Overlaps) Let z be an arbitrary point in E^d. Let s and t be two simplices (each of any dimensionality). Say that *s overlaps t from the viewpoint z* if some point of s not shared by t lies between z and t, as Fig. 23 illustrates. In other words, there exists a point $p_s \in s \backslash t$ and a point $p_t \in t$ such that $p_s \in z p_t$.

Definition 27 (Witness Function) Let $h \subset E^{d+1}$ be a non-vertical d-flat. The *witness function* $h(p)$ is the linear function that maps each point $p \in E^d$ to the x_{d+1}-coordinate such that $\langle p, h(p) \rangle \in h$. In other words, if $\ell \subset E^{d+1}$ is the vertical line (parallel to the x_{d+1}-axis) that contains $\langle p, 0 \rangle$, then $h(p)$ is the x_{d+1}-coordinate of $h \cap \ell$, as Fig. 23 illustrates.

Lemma 14 *Let s and t be two simplices, each of any dimensionality. Suppose there is a non-vertical d-flat h_s that includes s^+ such that every vertex of t^+ lies on or above h_s. Suppose there is a non-vertical d-flat h_t that includes t^+ such that every vertex of s^+ lies* strictly *above h_t, except the vertices shared by t^+. Then the following statements hold*:

- *If s and t are not disjoint, then $s \cap t$ is a face of both s and t.*
- *Let z be an arbitrary point in E^d. If s overlaps t from the viewpoint z, then $h_s(z) > h_t(z)$.*

Proof If s is a face of t, both results follow immediately. (In this case, s does not overlap t from any viewpoint.) Otherwise, s^+ has a vertex that t^+ lacks. This vertex lies on h_s and above h_t, so $h_s \neq h_t$. The d-flats h_s and h_t must intersect, because some vertex of s^+ lies above h_t and some vertex of t^+ lies on or above h_s. Let i^+ be the $(d-1)$-flat $h_s \cap h_t$. Let $i = \{p \in E^d : \ h_s(p) = h_t(p)\}$ be the vertical projection of i^+ into E^d, as illustrated in Fig. 23.

The hyperplane i cuts E^d into two halfspaces. Every vertex of s lies in the closed halfspace $\{p \in E^d\colon\ h_s(p) \geq h_t(p)\}$. Therefore, so does every point in s. Likewise, every point in t lies in the closed halfspace $\{p \in E^d\colon\ h_s(p) \leq h_t(p)\}$. Any vertex v of s that lies on i has a lifted companion v^+ that lies on h_t, so by assumption, v must be a vertex of t. Therefore, $s \cap t$ is the convex hull of the vertices of s that lie on i, which is a face of both s and t. Furthermore, any point of s not shared by t cannot lie on i.

If s overlaps t from the viewpoint z, then some point $p_s \in s \backslash t$ lies between z and t. The point p_s lies in the open halfspace $\{p \in E^d\colon\ h_s(p) > h_t(p)\}$, so z must lie there too. □

Lemma 14 is similar to theorems of Edelsbrunner [15] and Edelsbrunner and Shah [19], which they use to prove the *acyclicity* of every regular triangulation T: for any fixed viewpoint z, the overlap relation among regular simplices is a partial order. The function $h_s(z)$ imposes a total order on the simplices in T such that no simplex overlaps another simplex that appears later in the order. This acyclicity property does not extend to CDTs, but it does apply to the regular simplices that comprise the lower-dimensional facades in a ridge-protected PLC (see Section 5.1).

Proof of the Delaunay Lemma The "only if" implication is straightforward. If T is a CDT of X, Properties A and B follow by the definition of CDT. Property D follows because every d-simplex in a CDT is constrained semiregular. Property C follows because each facet in a CDT is constrained semiregular—unless it is included in a constraining $(d-1)$-facade of X—and every constrained semiregular simplex is locally semiregular.

Not surprisingly, the "if" implication takes more work to prove. Suppose T is a triangulation with all four properties. Let s be any d-simplex in T. The following argument establishes that s is constrained semiregular.

Let v be any vertex in X that is visible from some point p in the interior of s. It is helpful if the line segment vp does not intersect any simplex in T of dimension less than $d-1$, except at the vertex v. If this is not true, then by Lemma 6 there is a neighborhood of p from which every point can see v. Choose from this neighborhood a point p' such that p' is in the interior of s and vp' does not intersect any simplex in T of dimension less than $d-1$, except at v.

T is a simplicial complex that fills X by Property A, so the line segment vp' intersects the interiors of a contiguous sequence of d-simplices $s_1, s_2, \ldots, s_k = s$, with $v \in s_1$. Let f_i denote the facet shared by s_i and s_{i+1}. Because vp' does not intersect any lower-dimensional faces of T (except at v), it passes through the relative interiors of the facets $f_1, f_2, \ldots, f_{k-1}$. Because v is visible from p', none of these facets is included in a constraining facade, so by Property C all of them are locally semiregular.

Because f_1 is locally semiregular, either $h_{s_1} = h_{s_2}$ or f_1 is locally regular. In the latter case, $h_{s_1}(v) > h_{s_2}(v)$ by Lemma 14; in either case, $h_{s_1}(v) \geq h_{s_2}(v)$. The same reasoning holds for $f_2, \ldots, f_{k-1}$, so $h_{s_1}(v) \geq h_{s_2}(v) \geq \cdots \geq h_{s_k}(v) = h_s(v)$. If v is a vertex of s_1, then the height (x_{d+1}-coordinate) of v^+ is $h_{s_1}(v)$; otherwise, v is

submerged, and by Property D the height of v^+ is at least $h_{s_1}(v)$.[6] In either case, $v_{x_{d+1}} \geq h_{s_1}(v) \geq h_s(v)$, so v^+ cannot lie below the witness d-flat h_s. Because this is true of every vertex v that is visible from the interior of s, and because s respects X by Property B, s is constrained semiregular.

By assumption, X has no dangling facades, so by Theorem 9, T is a CDT of X. □

If X has dangling facades, T may be cut into subcomplexes of different dimensionalities so that each subcomplex has no dangling simplices. Then the Delaunay Lemma can be applied to each piece separately, thereby showing the constrained semiregularity of the whole. In a k-dimensional portion of the triangulation, only the local semiregularity of the $(k-1)$-faces needs to be checked.

To make good on the title of this article, the following definition offers the constrained analog of a regular triangulation. A constrained regular triangulation is a projection of a polyhedron whose ridges are locally convex everywhere except where the constraining facades permit them to be reflex.

Definition 28 (Constrained Regular Triangulation) A triangulation T is *constrained regular* relative to an unweighted PLC X if T fills and respects X, and there exists an assignment of weights to the vertices in X such that every non-constraining facet in T is locally regular.[7]

Every generic CDT (recall Definition 19) is a constrained regular triangulation. This fact is a consequence of the Delaunay Lemma and the fact that in a generic CDT, constrained regularity and constrained semiregularity are the same. However, not *every* CDT is a constrained regular triangulation. For example, let T be the triangulation illustrated in Fig. 24, which is not regular. If all the vertex heights are zero, T is a valid weighted Delaunay triangulation and (relative to a compatible PLC) a valid weighted CDT. However, only the simplices on the boundary of T are regular; the rest are only semiregular. No assignment of weights can make every edge of T regular. Nevertheless, if T is a triangulation of a PLC X, and X includes one of the long internal edges as a constraining segment, then T is constrained regular with respect to X.

The viewpoint at the center of the triangulation T in Fig. 24 demonstrates that T does *not* have the acyclic property established by Edelsbrunner and Shah [19] for regular triangulations. However, constrained regular triangulations have a limited acyclicity property. Say that s *visibly overlaps* t from the viewpoint z if there exists a point p_t in t's relative interior that is visible from z, and a point $p_s \in s \backslash t$ such that $p_s \in zp_t$. For any viewpoint z, the visible overlap relation among simplices is a partial order. This fact follows from Lemma 14 by the same inductive step used to prove the Delaunay Lemma, with the inequalities replaced by the strict inequalities $h_{s_i}(z) > h_{s_{i+1}}(z)$.

[6] The vertex v might lie in several d-simplices of T (on a shared boundary), and Property D explicitly applies to only one of them. However, the lifted surface T^+ is continuous where simplices of T meet, so Property D holds for all the simplices in T that contain v.

[7] Obviously, there is always an assignment of weights to the vertices of X missing from T that satisfies Property D of the Delaunay Lemma. Just make their weights really small.

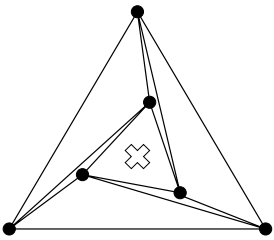


Fig. 24 A triangulation that is not regular. From the viewpoint at the center, the three outer triangles form a mutually overlapping cycle

Linear programming can determine whether a triangulation T that fills and respects X is constrained regular relative to X. The variables of the linear program are the vertex weights and a variable δ. For each non-constraining facet f in T, write a linear constraint enforcing the local regularity of f. Specifically, f is a facet of two d-simplices s and t; the linear constraint requires that the apex of s^+ (not shared by t^+) be a distance of at least δ above t's witness d-flat. The objective is to maximize δ subject to the facet constraints. If this linear program has a feasible point with $\delta > 0$, T is constrained regular relative to X.

3.3 The Omnipresent Complex of Constrained Regular Simplices

A property of every PLC X is that its constrained regular simplices (of all dimensionalities, within all the facades in X) have disjoint relative interiors and form a simplicial complex, even if X has no CDT. Another property is that if X does have a CDT—perhaps several CDTs—then every constrained regular simplex appears in every CDT of X. This property implies that if X is generic, it has at most one CDT.

These properties do not hold for semiregular simplices. If some selection of $d + 2$ or more vertices of a PLC lift to a common non-vertical d-flat, the PLC might have more than one CDT, and its semiregular simplices might have intersecting interiors.

Because the CDT of a generic PLC contains every constrained regular simplex and no other simplex, CDT construction algorithms can work in a bottom-up fashion, from low dimensionalities to high: if an algorithm obtains the CDT of each constraining facade in a generic PLC X (perhaps by calling itself recursively), it can construct the constrained regular d-simplices with confidence that they will match the facade triangulations. For a nongeneric PLC, however, the CDTs of different constraining facades might be incompatible with each other, causing a CDT construction algorithm to fail to find a CDT of the whole PLC even when one exists. Section 6 offers a perturbation method that enforces genericity, so that CDT construction algorithms may avoid this fate.

The proofs rely on the following lemma, which is also used heavily in Section 5.

Lemma 15 *Let P and C be closed, convex polyhedra (not necessarily of the same dimensionality) with $P \subseteq C$. Let m be a point in the relative interior of P. Let C_m be the face of C whose relative interior contains m. Then $P \subseteq C_m$.*

Proof If $C_m = C$ the result follows immediately. Otherwise, by Definition 3, there is a hyperplane h such that $C_m = C \cap h$ and h does not intersect the relative interior of C, which implies that $C \backslash C_m$ lies entirely on one side of h. Clearly, $m \in h$. Suppose

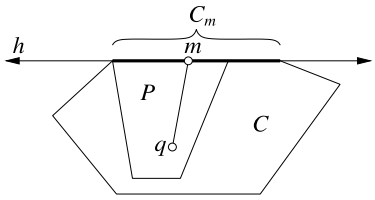


Fig. 25 If $P \subseteq C$ but $P \not\subseteq C_m$, then m is on the boundary of P

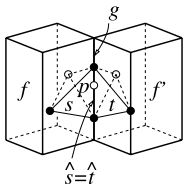


Fig. 26 A constrained semiregular simplex s and a constrained regular simplex t can intersect only at a shared face

for the sake of contradicting the lemma that $P \not\subseteq C_m$, as illustrated in Fig. 25. Then P contains a point q in $C \backslash C_m$. Thus q is on the same side of h as C, and no point in P is on the other side of h. Because P is convex, P includes the line segment qm, but any extension of the line segment qm past m lies outside P. Therefore, m is on the boundary of P. This contradicts the assumption that m is in the relative interior of P, so $P \subseteq C_m$.

□

The following theorem, which generalizes half of Lemma 14, shows that a constrained semiregular simplex and a constrained regular simplex can intersect only at a shared face.

Theorem 16 *Let s and t be simplices. Suppose that s is constrained semiregular within f and t is constrained regular within f', where f and f' are facades in a PLC X (possibly with $f = f'$), and neither f nor f' is a submersible vertex. If s and t are not disjoint, then $s \cap t$ is a face of both s and t.*

Proof Suppose s and t are not disjoint. Let p be a point in the relative interior of $s \cap t$, as illustrated in Fig. 26. Let g be the lowest-dimensional facade in X that contains p and is not a submersible vertex. (Either p is in the relative interior of g, or p coincides with an isolated submersible vertex of g's internal boundary. Note that g might be of any dimension from zero to d.) Because $p \in f$ and $p \in f'$, g is a face (not necessarily a proper face) of both f and f'.

Each of s and t has one face whose relative interior contains p. Call these faces $\hat{s}$ and $\hat{t}$, respectively. Because s and t respect X, so do $\hat{s}$ and $\hat{t}$. It follows that every facade that contains p (and is not a submersible vertex) includes $\hat{s}$ and $\hat{t}$. Three such facades are f, f', and g.

By Theorem 8, $\hat{s}$ is constrained semiregular within g. By Theorem 8′, $\hat{t}$ is constrained regular within g. By Lemma 14 (applied within the facade PLC Y_g), $\hat{s} \cap \hat{t}$ is a face of both $\hat{s}$ and $\hat{t}$. However, p is in the relative interiors of both $\hat{s}$ and $\hat{t}$, and $p \in \hat{s} \cap \hat{t}$, so $\hat{s} = \hat{s} \cap \hat{t} = \hat{t}$.

Because $\hat{s} = \hat{t}$ is a face of both s and t, $\hat{t} \subseteq s \cap t$. By Lemma 15 (substituting p for m, t for C, $\hat{t}$ for C_m, and $s \cap t$ for P), $s \cap t \subseteq \hat{t}$. Therefore, $\hat{s} = \hat{t} = s \cap t$, verifying that $s \cap t$ is a face of both s and t. □

Corollary 17 *The constrained regular simplices of a PLC have disjoint relative interiors.*

Corollary 18 *Let X be a PLC. Let T be the set that contains every simplex that is constrained regular within X or within a constraining facade in X. T is a simplicial complex.*

Proof By Theorem 8′, every face of every simplex in T is constrained regular within some facade that is not a submersible vertex. Therefore, T contains every face of every simplex in T. By Theorem 16, the intersection of any two simplices in T is either empty or a shared face of the two simplices. Hence T is a simplicial complex. □

A consequence of Corollary 18 is that if a PLC does not have a CDT, one or more of its facades has a gap that is not covered by constrained regular simplices. The next theorem shows that if a PLC has several CDTs, they share the same constrained regular simplices, and differ only by the simplices that are constrained semiregular but not constrained regular.

Theorem 19 *Every CDT of a PLC X contains every simplex that is constrained regular within X or within a constraining facade in X.*

Proof Let t be any simplex that is constrained regular within some facade f in X, where f is not a submersible vertex. (If a simplex is constrained regular within X, it is constrained regular within some d-facade in X.) Let p be a point in the relative interior of t.

Let T be a CDT of X. Because T fills X, T contains a simplex s that contains p and is not a submersible vertex. By the definition of CDT, s is constrained semiregular within the lowest-dimensional facade that includes it.

By Theorem 16, $s \cap t$ is a face of both s and t. However, $s \cap t$ contains p, which is in the relative interior of t, so $s \cap t = t$. Therefore, t is a face of s, and $t \in T$. This conclusion holds for every CDT T of X and every constrained regular simplex t. □

Corollary 20 *A generic PLC has at most one CDT.*

Proof By Theorem 19, every CDT of a PLC X contains every simplex that is constrained regular within a facade in X, except perhaps within a submersible vertex. By the definition of CDT, no CDT of X contains a simplex that is not constrained semiregular within a facade in X. If X is generic, constrained regularity and constrained semiregularity are equivalent. Therefore, two CDTs of X can differ from each other only in the choice of submersible vertices. However, a CDT fills X, so the choice of submersible vertices is uniquely determined by the higher-dimensional simplices. Therefore, X has at most one CDT. □

This corollary and Corollary 18 together imply that if a PLC is generic and has a CDT, a CDT construction algorithm can triangulate each facade of the PLC, starting with the 1-facades and working up to the d-facades, and rest assured that the facade triangulations of different dimensions all match.

4 Interpolation Criteria Optimized by CDTs

Among all triangulations of a fixed two-dimensional vertex set, the Delaunay triangulation is optimal by a variety of criteria—maximizing the smallest angle in the triangulation [28], minimizing the largest circumcircle among the triangles [4], and minimizing a property called the *roughness* of the triangulation [35, 37]. A two-dimensional CDT shares these same optimality properties, if it is compared with every other *constrained* triangulation of the same PSLG [4, 29].

Delaunay triangulations in higher dimensions also have optimality properties that generalize to CDTs and offer some of the reasons why higher-dimensional CDTs are such worthy objects of study. Rippa [38] investigates the use of two-dimensional triangulations for piecewise linear interpolation of a bivariate function of the form $Ax^2 + By^2 + Cx + Dy + E$, and concludes that if $A = B$, the Delaunay triangulation minimizes the interpolation error measured in the L_q-norm for every $q \geq 1$ (compared with all other triangulations of the same vertices). Melissaratos [31] generalizes Rippa's result to higher dimensions. D'Azevedo and Simpson [13] show that a two-dimensional Delaunay triangulation minimizes the radius of the largest min-containment circle of its simplices, and Rajan [36] generalizes this result to Delaunay triangulations and min-containment spheres of any dimensionality. The *min-containment sphere* of a simplex is the smallest hypersphere that encloses the simplex. If the center of the circumsphere of a simplex lies in the simplex, then the min-containment sphere is the circumsphere. Otherwise, the min-containment sphere is the min-containment sphere of some face of the simplex.

Rajan's result and a theorem of Waldron [51] together imply a second optimality result related to multivariate piecewise linear interpolation. Suppose you must choose a triangulation to interpolate an unknown function (not necessarily convex), and you wish to minimize the largest pointwise error in the domain. After you choose the triangulation, an adversary will choose the worst possible smooth function for your triangulation to interpolate, subject to a fixed upper bound on the absolute curvature (i.e. second directional derivative) of the function anywhere in the domain. The Delaunay triangulation is your optimal choice.

This section shows that Melissaratos' and Rajan's results generalize to CDTs (when CDTs exist). Melissaratos' result also generalizes to any monotonic norm and, with help from weighted CDTs, to any convex function. Rajan's result is particular to unweighted CDTs—the paraboloid is the right choice of heights to minimize the largest min-containment sphere. The proofs given here are similar to Fortune's presentation for unconstrained Delaunay triangulations [21], and are substantially simpler than Melissaratos' and Rajan's.

Consider multivariate piecewise linear interpolation on a weighted CDT. Let X be a PLC, and let $f(p)$ be a convex scalar function defined over the triangulation domain $|X|$. Assign each vertex $v \in X$ the weight $|v|^2 - f(v)$, so that the x_{d+1}-coordinate

of v^+ is $f(v)$. Let T be a weighted CDT of X, if one exists. The triangulation T and the vertex heights $f(v)$ define a piecewise linear surface $T^+ = \{s^+ : s \in T\}$. By analogy to witness functions (Definition 27), think of T^+ as a continuous piecewise linear function $T^+(p)$, which maps each point $p \in |X|$ to a real value. Because f is convex, every vertex in X is semiregular, so T^+ interpolates the lifted companion of every vertex in X, even if some vertices in X are missing from T.

Let $e(p) = T^+(p) - f(p)$ be the error in the interpolated function T^+ as an approximation of the true function f. At each vertex v in X, $e(v) = 0$. Because f is convex, the error satisfies $e(p) \geq 0$ for all $p \in |X|$.

Consider the unconstrained case first. T is the weighted Delaunay triangulation of the vertices in X, so T^+ is the underside of the convex hull of the lifted vertices. The intuition (formalized in Theorem 21 below) is that for any point $p \in |X|$, there is no way to triangulate the lifted vertices that yields a lesser value of $T^+(p)$ than the underside of the convex hull. Melissaratos' result follows immediately: T minimizes $\|e\|_{L_q}$ for every Lebesgue norm L_q.

The constrained case is only a little more complicated.

Theorem 21 *Let $f(p)$ be a function defined over the domain $|X|$ of a PLC X. Assign each vertex $v \in X$ the height $f(v)$—i.e. the weight $|v|^2 - f(v)$. If X has a weighted CDT, then at every point $p \in |X|$, every weighted CDT T of X minimizes $T^+(p)$ among all triangulations of X.*

Proof Let T be a weighted CDT of X. Suppose for the sake of contradiction that there is a triangulation S of X and a point p such that $S^+(p) < T^+(p)$. Let s be the simplex in S whose relative interior contains p. Let t be a simplex in T that contains p and is not a submersible vertex. Let f be the lowest-dimensional facade in X that includes t. Because t is not a submersible vertex, f is not one either, so s respects f. Because p is in both f and the relative interior of s, $s \subseteq f$. Because s respects X and $p \in s$, the vertices of s are visible from p by Theorem 3.

Define the point $p_s = \langle p, S^+(p) \rangle \in E^{d+1}$. Thus $p_s \in s^+ \in S^+$, and p is the projected companion of p_s. Because $S^+(p) < T^+(p)$, p_s lies below t^+. For every witness d-flat h_t that includes t^+, at least one vertex of s^+ lies below h_t, because s^+ is a simplex that contains p_s. Therefore, t is not constrained semiregular within f. However, by assumption, T is a weighted CDT of X, so t is constrained semiregular within f. By contradiction, there is not a triangulation S and a point p such that $S^+(p) < T^+(p)$. □

Corollary 22 *Let $f(p)$ be a convex function defined over the domain $|X|$ of a PLC X. Assign each vertex $v \in X$ the height $f(v)$. If X has a weighted CDT, then at every point $p \in |X|$, every weighted CDT T of X minimizes the interpolation error $|T^+(p) - f(p)|$ among all triangulations of X.*

Because the weighted CDT minimizes the error $e(p)$ at every point, the weighted CDT minimizes e in every norm that is monotonic in e, including the Lebesgue norms. With the right choice of weights, this result holds for any convex function. Rippa also investigates the special case of interpolating $f(p) = Ax^2 + By^2 + Cx + Dy + E$ where $A \neq B$. For a function of this form, an anisotropic triangulation (with

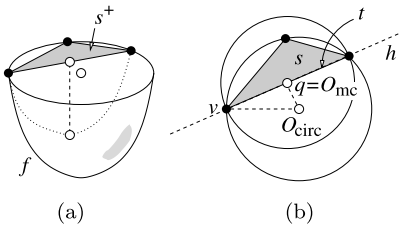


Fig. 27 (a) Within s, the error $e(p)$ is maximized at the point nearest the circumcenter of s. (b) Top view of s, its circumcircle, and its min-containment circle

long, thin triangles) is optimal. Rippa suggests handling such functions by affinely mapping the vertices in E^d to a "stretched" space over which $f(p)$ is isotropic, finding the Delaunay triangulation of the mapped vertices, and mapping the triangulation back to the original space. Corollary 22 suggests an alternative: use weights to achieve the same effect as Rippa's mapping. This approach obtains exactly the same results when $f(p)$ is parabolic, but it is more flexible as it can adapt to other convex functions as well.

Corollary 22 plays a part in showing that Rajan's result generalizes to CDTs.

Theorem 23 *If X has an unweighted CDT, then every unweighted CDT of X minimizes the largest min-containment sphere, compared with all other triangulations of X.*

Proof Recall that $e(p) = T^+(p) - f(p)$. As X is unweighted, $f(p) = |p|^2$.

Over any single d-simplex s, there is an explicit expression for $e(p)$. Recall from the proof of Lemma 1 that the witness d-flat h_s that includes s^+ has the witness function $h_s(p) = 2O_{\text{circ}} \cdot p - |O_{\text{circ}}|^2 + r_{\text{circ}}^2$, where O_{circ} and r_{circ} are the circumcenter and circumradius of s, and $p \in E^d$ varies freely. (The *circumcenter* and *circumradius* of s are the center and radius of s's circumsphere.) Hence, for all $p \in s$,

$$\begin{aligned} e(p) &= h_s(p) - f(p) \\ &= 2O_{\text{circ}} \cdot p - |O_{\text{circ}}|^2 + r_{\text{circ}}^2 - |p|^2 \\ &= r_{\text{circ}}^2 - |O_{\text{circ}}p|^2. \end{aligned}$$

Figure 27(a) illustrates the functions $h_s(p)$ and $f(p)$ over a triangle s. The error $e(p)$ is the vertical distance between the two functions. At which point p in s is $e(p)$ largest? At the point nearest the circumcenter, because $|O_{\text{circ}}p|^2$ is smallest there. (The error is maximized at the circumcenter if the circumcenter is in s; Fig. 27 gives an example where it is not.) Let O_{mc} and r_{mc} be the center and radius of the min-containment sphere of s, respectively. Lemma 24 below shows that the point in s nearest O_{circ} is O_{mc}, and $r_{\text{mc}}^2 = e(O_{\text{mc}})$.

It follows that the square of the min-containment radius of s is $\max_{p \in s} e(p)$, and thus the largest min-containment sphere of the entire triangulation has a squared radius of $\max_{p \in |T|} e(p)$. By Corollary 22, the unweighted CDT T minimizes this quantity among all triangulations of X. □

Lemma 24 *Let O_{circ} and r_{circ} be the circumcenter and circumradius of a d-simplex s. Let O_{mc} and r_{mc} be the center and radius of the min-containment sphere of s. For*

$p \in s$, define the function $e(p) = r_{\text{circ}}^2 - |O_{\text{circ}}p|^2$. Let q be the point in s nearest O_{circ}. Then $O_{\text{mc}} = q$ and $r_{\text{mc}}^2 = e(q)$.

Proof Let t be the face of s whose relative interior contains q. The face t is not a vertex, because the vertices of s are s's furthest points from O_{circ}. Because q is the point in t nearest O_{circ}, and because q is in the relative interior of t, the line segment $O_{\text{circ}}q$ is orthogonal to t. (This is true even if $t = s$, in which case $O_{\text{circ}} - q = \mathbf{0}$.) This fact, plus the fact that O_{circ} is equidistant from all the vertices of t, implies that q is equidistant from all the vertices of t (as Fig. 27 demonstrates). Let r be the distance between q and any vertex of t. Because $q \in t$, there is no containing sphere of t (or s) with radius less than r, because there is no direction q can move without increasing its distance from one of the vertices of t. Therefore, q and r are the center and radius of the min-containment sphere of t.

By the following reasoning, s has the same min-containment sphere as t. If $q = O_{\text{circ}}$, this conclusion is immediate. Otherwise, let h be the hyperplane through q orthogonal to $O_{\text{circ}}q$. Observe that h includes t. No point in s is on the same side of h as O_{circ}: if there were such a point w, there would be a point in s (between w and q) closer to O_{circ} than q, contradicting the fact that q is closest. Observe that h cuts the circumsphere into two pieces, and that the smaller piece encloses s and is enclosed by the min-containment sphere of t. Therefore, q and r are the center and radius of the min-containment sphere of s.

Let v be any vertex of t. Pythagoras' Law on $\triangle O_{\text{circ}}qv$ (see Fig. 27) yields $r_{\text{circ}}^2 = r^2 + |O_{\text{circ}}q|^2$, and therefore $r^2 = e(q)$. □

For an algebraic proof of Lemma 24 (based on quadratic program duality), see Lemma 3 of Rajan [36].

The optimality of the CDT for controlling the largest min-containment radius dovetails nicely with an error bound for piecewise linear interpolation derived by Waldron [51]. Let $\mathcal{C}_c$ be the space of scalar functions defined over $|X|$ that have C^1 continuity and whose absolute curvature nowhere exceeds c. In other words, for every $f \in \mathcal{C}_c$, every point $p \in |X|$, and every unit direction vector $\mathbf{d}$, the magnitude of the second directional derivative $f''_{\mathbf{d}}(p)$ is at most c. This is a common starting point for analyses of piecewise linear interpolation error. In contrast with Corollary 22, $\mathcal{C}_c$ is not restricted to convex functions.

Let f be a function in $\mathcal{C}_c$. Let $s \subseteq |X|$ be a simplex (of any dimensionality) with min-containment radius r_{mc}. Let h_s be a linear function that interpolates f at the vertices of s. Waldron shows that for all $p \in s$, the absolute error $|e(p)| = |h_s(p) - f(p)|$ is at most $cr_{\text{mc}}^2/2$. Furthermore, this bound is sharp: for every simplex s with min-containment radius r_{mc}, there is a function $f \in \mathcal{C}_c$ and a point $p \in s$ such that $|e(p)| = cr_{\text{mc}}^2/2$. (That function is $f(p) = c|p|^2/2$, as illustrated in Fig. 27.)

Theorem 25 *Every unweighted CDT T of X (if any exist) minimizes*

$$\max_{f \in \mathcal{C}_c} \max_{p \in |X|} |T^+(p) - f(p)|,$$

the worst-case pointwise interpolation error, among all triangulations of X.

Proof For any triangulation T, $\max_{f \in \mathcal{C}_c} \max_{p \in |X|} |T^+(p) - f(p)| = cr_{\max}^2/2$, where $r_{\max}$ is the largest min-containment radius among all simplices in T. The result follows immediately from Theorem 23. □

One of the reasons why CDTs are important is because, in the senses of Corollary 22 and Theorem 25, the CDT is an optimal piecewise linear interpolating surface. Of course, $e(p)$ is not the only criterion for the merit of a triangulation used for interpolation. Many applications need the interpolant to approximate the gradient—that is, not only must $T^+(p)$ approximate $f(p)$, but $\nabla T^+(p)$ must approximate $\nabla f(p)$ well too. For the goal of approximating $\nabla f(p)$ in three or more dimensions, the weighted CDT is sometimes far from optimal even for simple functions like the paraboloid $f(p) = |p|^2$. Still, the CDT is a good starting point for mesh improvement algorithms [6, 7, 10, 11, 16, 30, 46, 48] that create a triangulation that is appropriate for approximating both $f(p)$ and $\nabla f(p)$.

5 Proof of the CDT Theorem for Generic PLCs

Theorem 26 *Let X be a generic, weakly ridge-protected, d-dimensional PLC (weighted or not). X has a CDT (a weighted CDT if X is weighted).*

This section is devoted to the proof of Theorem 26, the generic version of the *CDT Theorem*. A lot of ink must be split for it, and readers who are not feeling athletic are invited to skip to Section 6, where the genericity requirement is removed from Theorem 26.

Half the work is already done: Corollary 18 states that the constrained regular simplices form a simplicial complex, and Theorem 10 states that if this complex fills X, it is a CDT of X. The most difficult part of the proof is to show that if X is generic and weakly ridge-protected, the complex fills X. The forthcoming Theorem 30 shows that every point in a weakly ridge-protected PLC lies in some constrained semiregular simplex. Unfortunately, several long proofs are needed to build up to that result.

5.1 Visibility Lemmata

One potential difficulty for the CDT Theorem is illustrated in Fig. 28. Imagine that you are standing at a point p in the interior of a three-dimensional domain, scanning

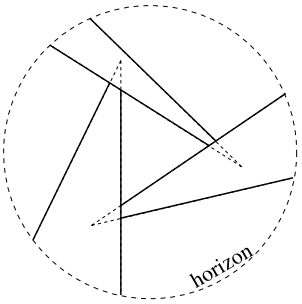


Fig. 28 Spherical projection of the halfspace above your vantage point

the halfspace "above" p for a visible vertex. Looking up into the sky, you see the three illustrated 2-facades, each of which occludes the apical vertex of another. The remaining vertices of these facades are below the horizon (in the halfspace below you). No vertex in the halfspace is visible from your vantage point, so there is no constrained semiregular simplex that contains p.

To prove the existence of a CDT, one must show that this possibility is precluded if X is weakly ridge-protected. Fortunately, Lemma 14 does exactly that. By the definition of "weakly ridge-protected," every grazeable constraining facade in X of dimension $d-2$ or less is a union of regular simplices. In Fig. 28 observe that the inner edges of the three facades form a cycle of overlapping edges. These edges are grazeable. However, Lemma 14 implies that the overlap relation among regular simplices (from a fixed viewpoint) constitutes a partial order. The regular edges bounding the 2-facades cannot form a cycle. This fact is the key to proving two lemmata for weakly ridge-protected PLCs.

For each regular simplex s, let h_s be a witness to the regularity of s. Every lifted vertex lies above h_s, except the vertices of s^+. Recall from Definition 27 the witness function $h_s(p)$, a linear function that maps each point $p \in E^d$ to the x_{d+1}-coordinate such that $\langle p, h_s(p)\rangle \in h_s$. If s is not d-dimensional, it has infinitely many witness d-flats; choose one arbitrarily so that $h_s(p)$ is consistently defined.

Lemma 27 *Let X be a weakly ridge-protected, d-dimensional PLC. Let p be a point in the interior of $|X|$. Let H be an open d-dimensional halfspace whose closure contains p. At least one vertex of X is in H and visible from p.*

Proof Suppose for the sake of contradiction that no vertex of X is in H and visible from p. Let A be the set containing every simplex e that has the following properties:

- e respects X and is regular within X, and
- there is a point m in e's relative interior such that $m \in H$ and m is visible from p.

A is empty—suppose for the sake of contradiction that it is not. Because no vertex of X is in H and visible from p, A contains no vertex. Let e be the simplex in A that maximizes $h_e(p)$. Let m be a point in e's relative interior that is in H and visible from p. Because e is a simplex that intersects H, at least one vertex v of e is in H, as Fig. 29(a) shows. (The other vertices of e might lie below the horizon, outside H.)

By assumption, v is not visible from p, although m is. Let n be the point nearest m on the line segment mv that is not visible from p. In other words, n is the first occluded point encountered on a "walk" from m to v.[8] The line segment pn must intersect some occluding facade of X at some point m'. If several facades occlude the

[8]How do we know that there *is* a first occluded point on the walk from m to v, rather than a last visible point? On the walk, there is at least one transition from points p can see to points p cannot see. Let n be the point where the first such transition occurs. Is n visible from p? There are two ways that a transition might occur. One possibility is an interposing facade that occludes the visibility of n, as in Fig. 29(a). The second possibility is that n lies on a facade f and is visible from p, but the points following n on the walk are occluded by f. To exclude this possibility, observe that e is convex, m is in e's relative interior, $v \in e$, $n \in mv$, and $n \neq v$. Therefore, n must lie in e's relative interior. Because e respects X, every facade that contains n includes e, and therefore f cannot occlude the visibility of any point in e from anywhere. These details are dreary, but the proof depends on them.

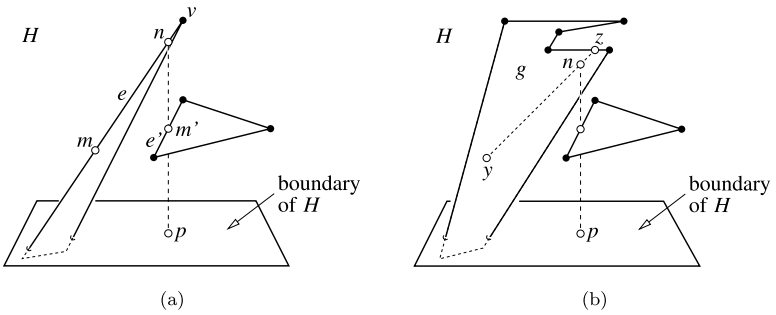


Fig. 29 The supposition that no vertex in H is visible from p leads to a contradiction

view of n from p, consider only the facade that intersects pn closest to p, so that m' is visible from p.

Let f be the face of that facade whose relative interior contains m'. (In Fig. 29, f is the edge e'.) Because n is the first occluded point on the walk from m to v, f must have dimension $d-2$ or less (i.e. m' cannot lie in the relative interior of a $(d-1)$-facade). Because no vertex is in H and visible from p, f is not a vertex. The grazing triangle $\triangle pnm$ demonstrates that f is grazeable. As X is weakly ridge-protected, f has a triangulation whose simplices respect X and are regular within X. Let e' be the simplex in that triangulation whose relative interior contains m'. Because n lies in H and p lies in its closure, m' lies in H, so $e' \in A$ (by the definition of A). Because $n \in e$, e' overlaps e from the viewpoint p, and therefore $h_{e'}(p) > h_e(p)$ by Lemma 14.

However, this contradicts the assumption that e maximizes $h_e(p)$ among all members of A. It follows that A is empty.

Because p is in the interior of $|X|$, at least one facade in X intersects H. Let g be the lowest-dimensional facade in X whose relative interior contains a point y that is in H and visible from p. By assumption, g is not a vertex.

Because g intersects H, at least one vertex of g is in H. Imagine shooting a ray from y toward that vertex. Let z be the first point on the boundary of g struck by the ray, as illustrated in Fig. 29(b). As g might not be convex, z might not be the vertex, but z is in H. Because g is the lowest-dimensional facade whose relative interior contains a point in H visible from p, and z lies in the relative interior of a proper face of g, z is not visible from p. Let n be the first occluded point encountered on a "walk" from y to z. By a repetition of the reasoning above, some simplex in A is interposed between p and n, but A is empty, so this is a contradiction.

It follows that some vertex of X is in H and visible from p. □

A second lemma reveals a more subtle (and barely comprehensible) property of visibility in PLCs.

Lemma 28 *Let X be a weakly ridge-protected, d-dimensional PLC. Let $h \subset E^{d+1}$ be a non-vertical d-flat. Let $V_h = \{v \in X\colon\ v$ is a vertex and v^+ is on or below $h\}$, and let $C_h = \mathrm{conv}(V_h)$. (See Fig. 30. Note that V_h and C_h are sets of points in E^d, not E^{d+1}.)*

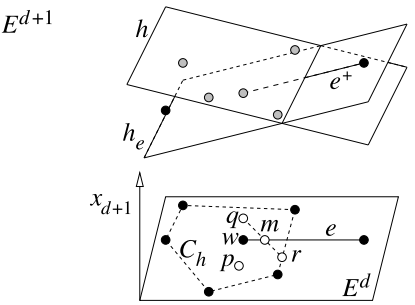


Fig. 30 Because m is between q and r, $h(m) > h_e(m)$

Let p be a point in C_h. Suppose that no vertex in X visible from p lifts to a point below h.

Let $f \in X$ be a grazeable constraining facade of dimension $d-2$ or less. Suppose that some point $m_f \in f \cap C_h$ is visible from p.

Then f includes the face of C_h whose relative interior contains m_f. (This face may be C_h itself.)

Proof Because X is weakly ridge-protected, f has a triangulation whose simplices respect X and are regular within X. Let t be the simplex in this triangulation whose relative interior contains m_f.

Let A be the set containing every simplex e that has the following properties:

- e respects X and is regular within X, and
- there is a point m in e's relative interior such that
  - — m is visible from p,
  - — $m \in C_h$, and
  - — e does not include the face of C_h whose relative interior contains m.

If A is empty, then $t \notin A$, so t includes the face of C_h whose relative interior contains m_f, and the lemma holds. Suppose for the sake of contradiction that A contains at least one simplex.

Let e be the simplex in A that maximizes $h_e(p)$. As $e \in A$, there is a point m in the relative interior of e such that $m \in C_h$ and m is visible from p. Because e is regular, there is a witness d-flat $h_e \subset E^{d+1}$ that includes e^+, as illustrated in Fig. 30. Each vertex of e lifts to a point on h_e. Every other vertex in X lifts to a point above h_e.

For each vertex $v \in V_h$, v^+ lies on or below h, and on or above h_e, so $h(v) \geq h_e(v)$. If v is in V_h but not in e, then v^+ lies strictly above h_e, so $h(v) > h_e(v)$.

Because C_h is the convex hull of V_h, and h and h_e are linear functions, it follows that for each point $q \in C_h$, $h(q) \geq h_e(q)$, and if q is not in e, then $h(q) > h_e(q)$.

Let C_m be the face of C_h whose relative interior contains m. By assumption, e does not include C_m, so some point $q \in C_m$ is not in e. Because m is in the relative interior of C_m, there is a point $r \in C_m$ such that m is between q and r. (See Fig. 30.) Thus $h(q) > h_e(q)$ and $h(r) \geq h_e(r)$, so by the linearity of h and h_e, $h(m) > h_e(m)$.

Because e is a simplex that contains m, there must be at least one vertex w of e for which $h(w) > h_e(w)$. Because w^+ lies on h_e, w^+ lies below h, so $w \in V_h$ (by the definition of V_h). By assumption, no vertex visible from p lifts to a point below

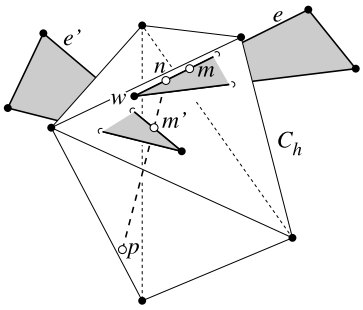


Fig. 31 Because m is visible from p and w is not, some simplex e' must overlap e

h, so w is not visible from p. However, recall that $m \in e$ is visible from p. Can m be visible from p if w is not?

Let n be the point nearest m on the line segment mw that cannot see p, as illustrated in Fig. 31. The line segment pn must intersect some facade in X at some point m'. If there are several facades occluding the view of n from p, consider only the facade that intersects pn closest to p, so that m' is visible from p.

Let g be the face of that facade whose relative interior contains m'. (In Fig. 31, g is the edge e'.) Because n is the first occluded point on the walk from m to w, g must have dimension $d-2$ or less (i.e. m' cannot lie in the relative interior of a $(d-1)$-facade). The grazing triangle $\triangle pnm$ demonstrates that g is grazeable. As X is weakly ridge-protected, g has a triangulation whose simplices respect X and are regular within X. Let e' be the simplex in that triangulation whose relative interior contains m'. Observe that $n \in C_h$ because n lies between m and w, which are both in C_h. Moreover, $m' \in C_h$ because m' lies between n and p. Let $C_{m'}$ be the face of C_h whose relative interior contains m'. By Lemma 15 (substituting C_h for C, $C_{m'}$ for C_m, and pn for P), $pn \subseteq C_{m'}$. Because g occludes the visibility between p and n, e' contains neither p nor n. It follows that $C_{m'} \not\subseteq e'$.

By the definition of A, $e' \in A$. Because e' overlaps e from the viewpoint p, $h_{e'}(p) > h_e(p)$ by Lemma 14. However, this contradicts the assumption that e maximizes $h_e(p)$ among all members of A. It follows that A is empty, and the lemma holds. □

5.2 Ridge-Protected PLCs Are Filled

This section completes the proof of Theorem 26. Most of the effort is spent proving that if a PLC is weakly ridge-protected, every point in the triangulation domain lies in some constrained semiregular simplex. The proof is made easier by considering a subset of the triangulation domain first—a set of points from which visibility is particularly well behaved.

For a d-dimensional PLC X, let N be the set containing every point in the interior of $|X|$ that is not cohyperplanar with any d affinely independent vertices in X. No point in N lies on any constraining facade, nor on any k-simplex whose vertices are in X for $k < d$, nor on their affine hulls. The closure of N is the union of the d-facades in X.

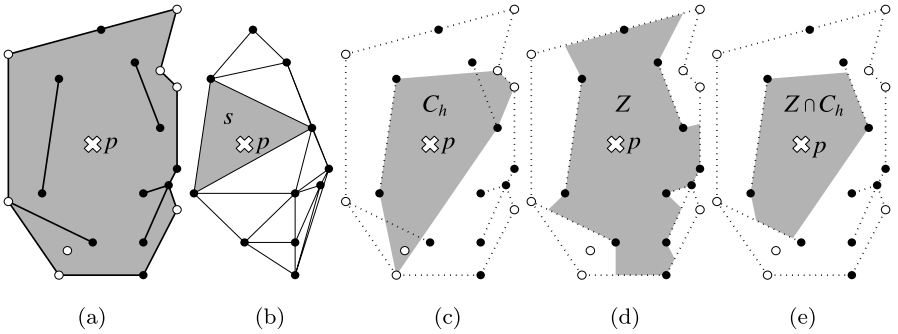


Fig. 32 From left to right: a PLC X wherein the vertices visible from p (the set W) are colored black, and the vertices not visible from p are colored white. The weighted Delaunay triangulation of W, and the simplex s therein that contains p. C_h is the convex hull of the vertices whose lifted companions lie on or below the witness d-flat for s. Z is the closure of the set of points visible from p. $Z \cap C_h$ is convex and respects X

Lemma 29 *Let X be a weakly ridge-protected, d-dimensional PLC. Define N as above, and let p be a point in N. Some constrained semiregular d-simplex contains p.*

Proof Let W be the set of all vertices in X visible from p—the black vertices in Fig. 32(a). The following reasoning establishes that p is in $\text{conv}(W)$. Suppose for the sake of contradiction that it is not. Then there is an open halfspace H such that p lies on the boundary of H and $W \cap H = \emptyset$.[9] However, by Lemma 27, some vertex of X is in H and visible from p. This vertex is in $W \cap H$, a contradiction.

Let s be the d-simplex that contains p in a weighted Delaunay triangulation of W (Fig. 32(b)). Because $p \in \text{conv}(W)$, some such simplex must exist. The rest of this proof shows that s is constrained semiregular within X, so the lemma holds.

Let h_s be the unique witness to the semiregularity of s within W. No vertex in W lifts to a point below h_s, so no vertex in X visible from p lifts to a point below h_s. Let $V_h = \{v \in X\colon\ v \text{ is a vertex and } v^+ \text{ lies on or below } h_s\}$. Let C_h be the convex hull of V_h (Fig. 32(c)). Observe that the vertices of s are in V_h, so $s \subseteq C_h$ and $p \in C_h$.

Let Z be the closure of the set of all points that p can see in the triangulation domain $|X|$ (Fig. 32(d)). Because $p \in N$, p lies in the interior of $|X|$, which implies that Z is d-dimensional with p in its interior. The vertices of s are in Z.

Because Z is the closure of points visible from p, the shadows cast by constraining facades of dimension $d-2$ or less have no effect on Z. Z is a star-shaped polyhedron (not generally convex) with two types of facets: portions of $(d-1)$-facades, and *shadow facets* that are cohyperplanar with p because they are boundaries of shadows cast by occluding $(d-1)$-facades.

The rest of this proof is a sequence of claims and their justifications.

Claim *No constraining $(d-1)$-facade in X intersects the interior of Z.* Because $p \in N$, p is not cohyperplanar with any $(d-1)$-facade, so every $(d-1)$-facade casts

[9] This claim is intuitive, but its formal proof is tricky. It is the well-known Farkas Lemma; see Ziegler [54] for a proof.

a shadow (occludes visibility from p) and no $(d-1)$-facade intersects the interior of Z.

Claim *$Z \cap C_h$ is a star-shaped d-polyhedron.* Because $p \in N \cap C_h$, p is in the interior of C_h. Because Z and C_h are both closed star-shaped d-polyhedra with p in their kernels and in their interiors, so is $Z \cap C_h$.

Claim *No constraining facade in X intersects the interior of $Z \cap C_h$.* Suppose for the sake of contradiction that a constraining facade f intersects the interior of $Z \cap C_h$. Let m be a point in the intersection of f's relative interior and the interior of $Z \cap C_h$. Assume without loss of generality that m is visible from p—if it is not, then m's visibility is occluded by some other constraining facade that intersects the interior of $Z \cap C_h$ closer to p (because $Z \cap C_h$ is star-shaped with p in its kernel), so f and m can be replaced by the occluding facade and the closer intersection point.

Because no constraining $(d-1)$-facade intersects the interior of Z, f must have dimension $d-2$ or less. To show that f is grazeable, choose an open grazing triangle L that does not intersect any constraining facade, such that one boundary edge of L contains m. Does such a triangle always exist? If L has m on its boundary and is sufficiently small, the only constraining facades that can intersect L are those that contain m. These facades intersect the interior of Z, so they have dimension $d-2$ or less. Almost every plane (2-flat) through m intersects these facades only at the point m. (Here, "almost every" is used in the analytic sense: for any $(d-2)$-facade g that contains m, the set of planes through m that intersect $g \backslash \{m\}$ has measure zero in the space of planes through m.) Therefore, almost every sufficiently small open triangle with m on its boundary intersects no constraining facade, so f has a grazing triangle.

By Lemma 28, $f \supseteq C_h$. This contradicts the fact that C_h is d-dimensional and f is at most $(d-2)$-dimensional, so no constraining facade intersects the interior of $Z \cap C_h$.

Claim *$Z \cap C_h$ is convex.* See Fig. 32(e). Suppose for the sake of contradiction that $Z \cap C_h$ is not convex. Then there exist two points q and r in the interior of $Z \cap C_h$ such that $qr \not\subseteq Z \cap C_h$. Because $Z \cap C_h$ is star-shaped with p in its kernel, $Z \cap C_h$ includes both pq and pr, so the three points p, q, and r cannot be collinear. Continuously move q and r toward p until $\triangle pqr \subset Z \cap C_h$, but qr still intersects the boundary of $Z \cap C_h$, as illustrated in Fig. 33. Let m be the point nearest q on qr that lies on the boundary of $Z \cap C_h$. (That point is neither q nor r, which are in the interior.) Loosely speaking, $Z \cap C_h$ is locally reflex at m. Because C_h is convex with q and r in its interior, m also lies in the interior of C_h, so m must lie on the boundary of Z.

Because the open triangle $L = \triangle pqr$ is included in the interior of $Z \cap C_h$, which intersects no constraining facade, L is a grazing triangle for m, and m is visible from p. Because m lies on the boundary of Z, but the open line segment pm does not intersect Z's boundary, m lies on at least one facet of Z that is not a shadow facet. Therefore, m lies on some $(d-1)$-facade g, as illustrated. Because g intersects neither the open triangle L nor the open line segment qm, m must lie on the boundary of g.

Let $\hat{g}$ be the face of g whose relative interior contains m. Because m is on g's boundary, $\hat{g}$ has dimension $d-2$ or less. L demonstrates that $\hat{g}$ is grazeable. By

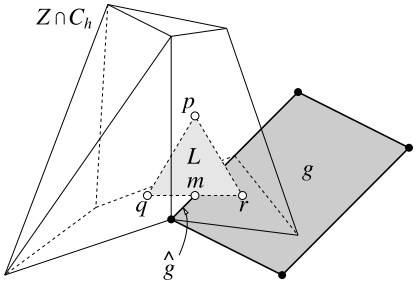


Fig. 33 If $Z \cap C_h$ is not convex, its boundary incorporates a grazeable facade $\hat{g}$

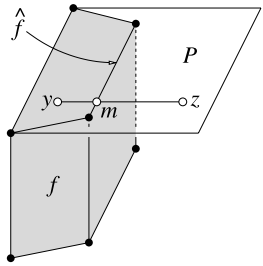


Fig. 34 The circumstance depicted here, where a facade f intersects the interior of a face P of $Z \cap C_h$ but does not include P in its entirety, cannot happen in a weakly ridge-protected PLC

Lemma 28, $\hat{g} \supseteq C_h$. This contradicts the fact that C_h is d-dimensional and $\hat{g}$ is at most $(d-2)$-dimensional, so $Z \cap C_h$ is convex.

Claim *$Z \cap C_h$ has no shadow facets.* This claim follows because shadow facets are cohyperplanar with p, but $Z \cap C_h$ is a convex d-polyhedron with p in its interior.

Claim *$Z \cap C_h$ respects X.* Suppose for the sake of contradiction that some facade $f \in X$ (that is not a submersible vertex) intersects the relative interior of a face P of $Z \cap C_h$, but f does not include P. Let y be a point in the intersection of f with the relative interior of P, as illustrated in Fig. 34.

Because f is closed and does not include P, there is a point z in the relative interior of P that is not in f. Let m be the point nearest z in $f \cap yz$, as illustrated. Because y and z are in the relative interior of P, so is m. Let y' be a point in P such that m is between y' and z. (The choice $y' = y$ will do if $y \neq m$; but if $y = m$, choose y' just past m on the ray $z\vec{m}$.) Let $\hat{f}$ be the face of f (possibly f itself) whose relative interior contains m. This choice guarantees that y' and z do not lie on the affine hull of $\hat{f}$, and $\hat{f}$ cannot have dimension d.

The facade $\hat{f}$ cannot have dimension $d-1$, either. If it did, then it would intersect the interior of $Z \cap C_h$, because m is in the relative interiors of both $\hat{f}$ and $y'z$, $y'z$ is on the boundary of $Z \cap C_h$, and $y'z$ does not lie on the same hyperplane as $\hat{f}$. However, no $(d-1)$-facade intersects the interior of Z. Therefore, $\hat{f}$ has dimension $d-2$ or less. To show that $\hat{f}$ is grazeable, choose an open grazing triangle L such that L is included in the interior of $Z \cap C_h$, and $y'z$ is an edge of (the closure of) L. No constraining facade intersects the interior of $Z \cap C_h$, so L is indeed a grazing triangle.

Let C_m be the face of C_h whose relative interior contains m. By Lemma 28, $\hat{f} \supseteq C_m$. Recall that m lies in the relative interior of P, which is a face of $Z \cap C_h$,

which implies that $P \subseteq C_h$. By Lemma 15, $P \subseteq C_m$. Thus $z \in P \subseteq C_m \subseteq \hat{f} \subseteq f$, contradicting the fact that z is not in f. The claim that $Z \cap C_h$ respects X follows.

Claim $s \subseteq Z \cap C_h$. This claim follows because both Z and C_h contain all the vertices of s, and $Z \cap C_h$ is convex.

Claim *s respects X.* Let t be any face of s, and let m be any point in the relative interior of t. Suppose some facade $f \in X$ (that is not a submersible vertex) contains m. As $m \in t \subseteq s \subseteq Z \cap C_h$, let C_m be the face of $Z \cap C_h$ whose relative interior contains m. Because t and $Z \cap C_h$ are convex with $t \subseteq Z \cap C_h$, it follows from Lemma 15 (substituting t for P and $Z \cap C_h$ for C) that $t \subseteq C_m$.

Recall that $Z \cap C_h$ respects f: if f intersects the relative interior of a face of $Z \cap C_h$, then f includes the whole face. Because f intersects the relative interior of C_m (at m), f includes C_m, which implies that $t \subseteq f$. This relationship holds for any face t of s, any point m, and any facade $f \in X$ that satisfy the assumptions, so s respects X.

Claim *Every point in $Z \cap C_h$ can see every other point in $Z \cap C_h$, but no point in the interior of $Z \cap C_h$ can see any vertex of V_h not in $Z \cap C_h$.* The first half of this claim follows from Theorem 3 because $Z \cap C_h$ respects X. For the second half of the claim, let q be a point in the interior of $Z \cap C_h$, and let v be a vertex in V_h that is not in $Z \cap C_h$. Some facet F of $Z \cap C_h$ lies between q and v. Because v is in C_h (which is convex) and q is in its interior, F is not on the boundary of C_h. Thus F must lie on the boundary of Z. Because $Z \cap C_h$ has no shadow facets, F must be included in some $(d-1)$-facade in X, which occludes the visibility of v from q. Therefore, no point in the interior of $Z \cap C_h$ can see any vertex of V_h not in $Z \cap C_h$.

Claim *s is constrained semiregular.* Because $p \in Z \cap C_h$, p sees every vertex in $Z \cap C_h$. By construction, no vertex visible from p has a lifted companion below the witness h_s; therefore, no vertex in $Z \cap C_h$ has one. By the definition of V_h, every vertex in X whose lifted companion is below h_s is in V_h. By the previous claim, no point in the interior of $s \subseteq Z \cap C_h$ can see any vertex of V_h not in $Z \cap C_h$. Therefore, no point in the interior of s can see any vertex whose lifted companion is below h_s. Moreover, s respects X, so s is constrained semiregular. □

Theorem 30 *Let X be a weakly ridge-protected, d-dimensional PLC. Let p be a point in a d-facade in X. Some constrained semiregular d-simplex contains p.*

Proof If $p \in N$, the result follows from Lemma 29. What about points not in N? Every point in N lies in some closed constrained semiregular d-simplex, and the closure of N is the union of all the d-facades in X. It follows that every point in every d-facade in X lies in some constrained semiregular d-simplex. □

Theorem 30 provides the machinery to prove Theorem 26: if X is a generic, weakly ridge-protected, d-dimensional PLC, then X has a CDT.

Proof of Theorem 26 Let T be the set that contains every simplex that is constrained semiregular within X or within a constraining facade in X. Because X is generic,

every constrained semiregular simplex is constrained regular, and Corollary 18 guarantees that T is a simplicial complex.

Let p be any point in the triangulation domain $|X|$. Let f be the highest-dimensional facade in X that contains p, and let k be the dimensionality of f. If $k = d$, Theorem 30 states that there exists a constrained semiregular d-simplex that contains p. By the definition of T, this d-simplex is in T.

If $k < d$, f is a dangling facade. Let Y_f be the k-dimensional facade PLC for f. By Corollary 12, Y_f is weakly ridge-protected. Therefore Theorem 30 applies, with Y_f substituted for X and k substituted for d. In this case the theorem states that some k-simplex exists that contains p and is constrained semiregular within Y_f. This k-simplex is in T.

Because such a simplex exists for every point $p \in |X|$, T fills X. By Theorem 10, T is a CDT of X. □

Theorem 26 requires X to be generic only to ensure that Corollary 18 applies. If X is nongeneric, T may contain constrained semiregular simplices whose interiors overlap. Theorem 30, however, holds even for nongeneric X.

6 Nongeneric PLCs, Weight Perturbations, and the CDT Theorem

It is well known that the Delaunay triangulation is not unique when $d + 2$ or more vertices lie on a common empty hypersphere. Every affinely independent subset of these cospherical vertices yields a Delaunay simplex. Some of the Delaunay simplices have mutually overlapping relative interiors, so some Delaunay simplices must be omitted to form a proper triangulation. Different choices yield different Delaunay triangulations. Likewise, a weighted Delaunay triangulation is not unique when the underside of the convex hull of the lifted vertices has a facet that is not a simplex.

The story is a bit more complicated for CDTs and weighted CDTs. The triangulation domain of a PLC might have a polyhedral gap (not necessarily convex) that is not covered by constrained regular simplices. Sometimes this happens simply because the PLC has no CDT, but sometimes the gap can be triangulated with constrained semiregular simplices. A gap might have several such triangulations, yielding multiple CDTs of one PLC. If a gap is shaped like Schönhardt's polyhedron, it cannot be triangulated at all.

A generic PLC has at most one weighted CDT (by Corollary 20), consisting of every constrained regular simplex (by Theorem 19), so it is pleasingly unambiguous. A nongeneric PLC raises the question of whether there exists a set of constrained semiregular simplices that fill the gaps and complete the triangulation. Because there may be several choices of constrained semiregular simplex to cover any point in a gap, determining whether a CDT exists is like solving a jigsaw puzzle with extra, useless pieces included.

Surprisingly, the problem of determining whether a three-dimensional nongeneric PLC has a CDT is NP-complete [24], even for an unweighted PLC. By contrast, it is always possible to determine whether a generic PLC has a CDT in polynomial time—by attempting to construct it. (See the second article in this series for further discussion.)

This section removes the genericity requirement from the CDT Theorem by perturbing the vertex weights so that no $d + 2$ vertices lift to a common non-vertical d-flat. The vertex coordinates are not perturbed. If the perturbed PLC has a CDT, the latter is also a CDT of the original, unperturbed PLC. The method works even for unweighted PLCs, by temporarily assigning each vertex a tiny weight. This idea first appears in the work of Edelsbrunner and Mücke [17, Section 5.4].

The weight perturbation method serves a practical function as well as a theoretical one. The third article in this series describes an easy way to implement the perturbations to ensure the correctness of algorithms for constructing and updating CDTs. There is a catch, though. Will a PLC that has a CDT still have a CDT after it is perturbed? Not necessarily. Perturbations cannot circumvent the NP-hardness result.

The perturbations are *symbolic*—the magnitudes of the perturbations are not explicitly specified. Following Edelsbrunner and Mücke, the ith vertex weight could be perturbed by ϵ^{2^i} for a sufficiently small ϵ, but the proofs are simpler if the perturbations are implicitly chosen by the following procedure instead.

Let X be a d-dimensional PLC. Let V be the set of vertices in X. Consider all the $(d+1)$-simplices, including degenerate ones, that can be defined by taking subsets of $d + 2$ lifted vertices from V^+. Call these the *orientation simplices*. Assume that the vertices of each orientation simplex are listed in some canonical order. The signed volume of an orientation simplex $\langle v_0^+, v_1^+, \ldots, v_{d+1}^+ \rangle$ is $1/(d + 1)!$ times the determinant of the matrix with column vectors $v_1^+ - v_0^+, v_2^+ - v_0^+, \ldots, v_{d+1}^+ - v_0^+$. Each signed volume varies linearly with the vertex weights. Every question about whether a lifted vertex lies above a witness d-flat for a d-simplex is a question about the sign of the volume of an orientation simplex. A volume of zero indicates cohyperplanarity.

Perturb the weights of the vertices in V one at a time, in some arbitrary order, each by a tiny negative or positive amount (different for each vertex). To perturb the weight of a vertex v, choose the magnitude of the perturbation to be sufficiently small that no orientation simplex's signed volume changes from positive to nonpositive, or from negative to nonnegative. Some signed volumes may change from zero to nonzero—that is the goal of the perturbations. Once a signed volume becomes nonzero, subsequent perturbations are not permitted to change its sign. The idea is to move vertices off of witnesses, but never to move a vertex from above a witness to below, nor vice versa. For each vertex in turn, it is always possible to choose a nonzero perturbation small enough to satisfy these restrictions. Perturb every vertex once.

Theorem 31 *Let X be a PLC. Let X' be a weighted PLC defined by perturbing every vertex weight in X as described above.* (*If X is unweighted, assign each vertex a weight of zero before perturbing it.*) *The following statements hold*:

A. *If a simplex s is regular within X, it is regular within X'.*
B. *If a simplex s is constrained regular within a facade in X, it is constrained regular within the same facade in X'.*
C. *If a simplex s is regular within X', it is semiregular within X.*
D. *If a simplex s is constrained regular within a facade in X', it is constrained semiregular within the same facade in X.*
E. *X' is generic and has at most one CDT.*

F. *If X is ridge-protected, so is X'.*
G. *If X is weakly ridge-protected, so is X'.*
H. *If X' has a CDT, the CDT of X' is a CDT of X.*
I. *If X is generic, X and X' have the same CDT* (*or lack thereof*).

Proof A lifted vertex lies above, on, or below the witness for a d-simplex according to whether the signed volume of some orientation simplex is positive, zero, or negative. Similarly, the regularity of any lower-dimensional simplex depends on the volumes of certain orientation simplices all having the right sign. Because a perturbation never changes the volume of any orientation simplex from positive to nonpositive or from negative to nonnegative, Statements A, B, C, and D hold by induction on the sequence of perturbations.

Perturbing the height of a vertex v moves v^+ off of any non-vertical d-flat that it lay on before the perturbation. No perturbation, of v or any other vertex, can move v^+ onto a witness d-flat that v^+ did not lie on before the perturbation, because that would imply that the volume of some orientation simplex changes from nonzero to zero. Therefore, v^+ does not lie on any witness immediately after it is perturbed, except the witnesses that by definition pass through v^+; and subsequent perturbations preserve this claim. By induction on the sequence of vertex perturbations, the claim holds for every vertex in X', and X' is generic. By Corollary 20, X' has at most one CDT.

Statements F and G follow from Statement A. Statement H follows from Statement D and the genericity of X'.

If X is generic, then constrained regularity and constrained semiregularity are equivalent. Thus, Statement B implies that any CDT of X is a CDT of X', just as Statement H says that any CDT of X' is a CDT of X. Either X and X' both have the same CDT, or both have no CDT. □

A CDT of X' is a CDT of X, but if X is nongeneric, different perturbations of X (i.e. perturbing the vertices in a different order, or using different mixtures of positive and negative perturbations) may yield different CDTs of X, or no CDT at all. Nevertheless, any choice of perturbation faithful to the procedure described above suffices to excise the genericity requirement from Theorem 26.

Theorem 32 (CDT Theorem) *Let X be a weakly ridge-protected, d-dimensional PLC* (*weighted or not*). *X has a CDT* (*a weighted CDT if X is weighted*).

Proof Let X' be the perturbed weighted PLC defined in Theorem 31. By the theorem, X' is generic and weakly ridge-protected, so by Theorem 26, X' has a CDT T. By Theorem 31, T is a CDT of X. □

7 Conclusions

In their article on two-dimensional conforming Delaunay triangulations, Edelsbrunner and Tan [20] write:

A seemingly difficult open problem is the generalization of our polynomial bound to three dimensions. The somewhat easier version of the generalized problem considers a graph whose vertices are embedded as points in $\mathbb{R}^3$, and edges are represented by straight line segments connecting embedded vertices. More relevant, however, is the problem for the crossing-free embedding of a complex consisting of vertices, edges, *and* triangles.

Three-dimensional CDTs shift the emphasis back to the former of these two problems. An algorithm that could create a Steiner CDT by inserting only a polynomial number of additional vertices would be an exciting development.

Some applications of finite element methods use meshes that have open slits, which are infinitesimally thin fissures across which information does not flow. The ideas in this article seem to extend in a straightforward way to topological PLCs wherein open slits are modeled by topological holes in the domain. Unfortunately, it is difficult to describe these PLCs in simple geometric terms, because of the need to distinguish topologically distinct points that have the same coordinates. For example, an internal $(d-1)$-facade can be converted into an open slit by making a topologically distinct copy of the facade that coincides with the original. Both the original and the copy adjoin the exterior domain (the infinitesimally thin hole), but they adjoin each other only along their external boundaries. The internal vertices in the original facade are topologically distinct from the internal vertices in the copy (and may or may not coincide), thereby supporting the interpolation of discontinuous functions as illustrated in Fig. 1(b). The open question is how to formulate these topological PLCs rigorously, and how to extend the results in this article to them.

Several other questions deserve investigation. Is there a simply stated and tested condition that is both sufficient and necessary for a generic PLC to have a CDT? The NP-hardness result suggests that there is no such condition for nongeneric PLCs. Is there a less conservative definition of "constrained Delaunay" (perhaps giving more power to constraining facades of dimension less than $d-1$) that admits useful, well-defined triangulations over a larger class of PLCs? Is there a better approach to assuring the existence of a CDT than to make a PLC weakly ridge-protected? Finally, when do curved manifold complexes (e.g. the stratifications mentioned in Section 2.1) have CDTs?

Acknowledgements I thank Dafna Talmor and Herbert Edelsbrunner for helpful discussions. In particular, Dafna pointed out the duality between degenerate faces of the Voronoi diagram and simplices that are Delaunay but not strongly Delaunay.

References

1. Aichholzer, O., Aurenhammer, F., Krasser, H., Brass, P.: Pseudotriangulations from surfaces and a novel type of edge flip. SIAM J. Comput. **32**(6), 1621–1653 (2003)
2. Aurenhammer, F.: Power diagrams: properties, algorithms, and applications. SIAM J. Comput. **16**(1), 78–96 (1987)
3. Aurenhammer, F., Krasser, H.: Pseudo-tetrahedral complexes. In: Proceedings of the Twenty-First European Workshop on Computational Geometry, Eindhoven, The Netherlands, March 2005, pp. 85–88 (2005)
4. Bern, M., Eppstein, D.: Mesh generation and optimal triangulation. In: Du, D.-Z., Hwang, F. (eds.) Computing in Euclidean Geometry. Lecture Notes Series on Computing, vol. 1, pp. 23–90. World Scientific, Singapore (1992)

5. Brown, K.Q.: Voronoi diagrams from convex hulls. Inf. Process. Lett. **9**, 223–228 (1979)
6. Cheng, S.-W., Dey, T.K.: Quality meshing with weighted Delaunay refinement. In: Proceedings of the Thirteenth Annual Symposium on Discrete Algorithms, San Francisco, CA, January 2002, pp. 137–146. Association for Computing Machinery, New York (2002)
7. Cheng, S.-W., Dey, T.K., Edelsbrunner, H., Facello, M.A., Teng, S.-H.: Sliver exudation. J. ACM **47**(5), 883–904 (2000)
8. Cheng, S.-W., Poon, S.-H.: Graded conforming Delaunay tetrahedralization with bounded radius-edge ratio. In: Proceedings of the Fourteenth Annual Symposium on Discrete Algorithms, Baltimore, MD, January 2003, pp. 295–304. Society for Industrial and Applied Mathematics, Philadelphia (2003)
9. Chew, L.P.: Constrained Delaunay triangulations. Algorithmica **4**(1), 97–108 (1989)
10. Chew, L.P.: Guaranteed-quality triangular meshes. Technical Report TR-89-983, Department of Computer Science, Cornell University (1989)
11. Chew, L.P.: Guaranteed-quality Delaunay meshing in 3D. In: Proceedings of the Thirteenth Annual Symposium on Computational Geometry, Nice, France, June 1997, pp. 391–393. Association for Computing Machinery, New York (1997)
12. Cohen-Steiner, D., de Verdière, É.C., Yvinec, M.: Conforming Delaunay triangulations in 3D. In: Proceedings of the Eighteenth Annual Symposium on Computational Geometry, Barcelona, Spain, June 2002, pp. 199–208. Association for Computing Machinery, New York (2002)
13. D'Azevedo, E.F., Simpson, R.B.: On optimal interpolation triangle incidences. SIAM J. Sci. Stat. Comput. **10**, 1063–1075 (1989)
14. Delaunay, B.N.: Sur la sphère vide. Izv. Akad. Nauk SSSR, VII Ser. **7**, 793–800 (1934)
15. Edelsbrunner, H.: An acyclicity theorem for cell complexes in d dimension. Combinatorica **10**(3), 251–260 (1990)
16. Edelsbrunner, H., Guoy, D.: An experimental study of sliver exudation. In: Tenth International Meshing Roundtable, Newport Beach, CA, October 2001, pp. 307–316. Sandia National Laboratories (2001)
17. Edelsbrunner, H., Mücke, E.P.: Simulation of simplicity: a technique to cope with degenerate cases in geometric algorithms. ACM Trans. Graph. **9**(1), 66–104 (1990)
18. Edelsbrunner, H., Seidel, R.: Voronoi diagrams and arrangements. Discrete Comput. Geom. **1**, 25–44 (1986)
19. Edelsbrunner, H., Shah, N.R.: Incremental topological flipping works for regular triangulations. Algorithmica **15**(3), 223–241 (1996)
20. Edelsbrunner, H., Tan, T.S.: An upper bound for conforming Delaunay triangulations. Discrete Comput. Geom. **10**(2), 197–213 (1993)
21. Fortune, S.: Voronoi diagrams and Delaunay triangulations. In: Du, D.-Z., Hwang, F. (eds.) Computing in Euclidean Geometry. Lecture Notes Series on Computing, vol. 1, pp. 193–233. World Scientific, Singapore (1992)
22. George, P.-L., Borouchaki, H.: Delaunay Triangulation and Meshing: Application to Finite Elements. Hermès, Paris (1998)
23. Gomes, A.J.P.: A concise B-rep data structure for stratified subanalytic objects. In:Kobbelt, L., Schröder, P., Hoppe, H. (eds.) Eurographics Symposium on Geometry Processing, Aachen, Germany, June 2003, pp. 83–93 (2003)
24. Grislain, N., Shewchuk, J.R.: The strange complexity of constrained Delaunay triangulation. In: Proceedings of the Fifteenth Canadian Conference on Computational Geometry, Halifax, Nova Scotia, August 2003, pp. 89–93 (2003)
25. Grünbaum, B., Shephard, G.C.: A new look at Euler's theorem for polyhedra. Am. Math. Mon. **101**(2), 109–128 (1994)
26. Hadwiger, H.: Vorlesungen über Inhalt, Oberfläche und Isoperimetrie. Springer, Berlin (1957)
27. Hazlewood, C.: Approximating constrained tetrahedrizations. Comput. Aided Geom. Des. **10**, 67–87 (1993)
28. Lawson, C.L.: Software for C^1 surface interpolation. In: Rice, J.R. (ed.) Mathematical Software III, pp. 161–194. Academic Press, New York (1977)
29. Lee, D.-T., Lin, A.K.: Generalized Delaunay triangulations for planar graphs. Discrete Comput. Geom. **1**, 201–217 (1986)
30. Li, X.-Y., Teng, S.-H.: Generating well-shaped Delaunay meshes in 3D. In: Proceedings of the Twelfth Annual Symposium on Discrete Algorithms, Washington, DC, January 2001, pp. 28–37. Association for Computing Machinery, New York (2001)
31. Melissaratos, E.A.: L_p optimal d dimensional triangulations for piecewise linear interpolation: a new result on data dependent triangulations. Technical Report RUU-CS-93-13, Department of Computer Science, Utrecht University, Utrecht, April 1993

32. Miller, G.L., Talmor, D., Teng, S.-H., Walkington, N., Wang, H.: Control volume meshes using sphere packing: generation, refinement and coarsening. In: Fifth International Meshing Roundtable, Pittsburgh, PA, October 1996, pp. 47–61 (1996)
33. Murphy, M., Mount, D.M., Gable, C.W.: A point-placement strategy for conforming Delaunay tetrahedralization. In: Proceedings of the Eleventh Annual Symposium on Discrete Algorithms, January 2000, pp. 67–74. Association for Computing Machinery, New York (2000)
34. Pav, S.E., Walkington, N.J.: Robust three dimensional Delaunay refinement. In: Thirteenth International Meshing Roundtable, Williamsburg, VA, September 2004, Sandia National Laboratories, pp. 145–156 (2004)
35. Powar, P.L.: Minimal roughness property of the Delaunay triangulation: a shorter approach. Comput. Aided Geom. Des. **9**(6), 491–494 (1992)
36. Rajan, V.T.: Optimality of the Delaunay triangulation in $\mathbb{R}^d$. In: Proceedings of the Seventh Annual Symposium on Computational Geometry, North Conway, NH, June 1991, pp. 357–363 (1991)
37. Rippa, S.: Minimal roughness property of the Delaunay triangulation. Comput. Aided Geom. Des. **7**(6), 489–497 (1990)
38. Rippa, S.: Long and thin triangles can be good for linear interpolation. SIAM J. Numer. Anal. **29**(1), 257–270 (1992)
39. Ruppert, J.M.: Results on triangulation and high quality mesh generation. PhD thesis, University of California at Berkeley, Berkeley, CA (1992)
40. Ruppert, J., Seidel, R.: On the difficulty of triangulating three-dimensional nonconvex polyhedra. Discrete Comput. Geom. **7**(3), 227–253 (1992)
41. Schönhardt, E.: Über die Zerlegung von Dreieckspolyedern in Tetraeder. Math. Ann. **98**, 309–312 (1928)
42. Seidel, R.: Voronoi diagrams in higher dimensions. Diplomarbeit, Institut für Informationsverarbeitung, Technische Universität Graz (1982)
43. Seidel, R.: Constrained Delaunay triangulations and Voronoi diagrams with obstacles. In: Poingratz, H.S., Schinnerl, W. (eds.) 1978–1988 Ten Years IIG, pp. 178–191. Institute for Information Processing, Graz University of Technology (1988)
44. Shewchuk, J.R.: Delaunay refinement mesh generation. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, May 1997. Available as Technical Report CMU-CS-97-137
45. Shewchuk, J.R.: A condition guaranteeing the existence of higher-dimensional constrained Delaunay triangulations. In: Proceedings of the Fourteenth Annual Symposium on Computational Geometry, Minneapolis, MN, June 1998, pp. 76–85. Association for Computing Machinery, New York (1998)
46. Shewchuk, J.R.: Mesh generation for domains with small angles. In: Proceedings of the Sixteenth Annual Symposium on Computational Geometry, Hong Kong, June 2000, pp. 1–10. Association for Computing Machinery, New York (2000)
47. Shewchuk, J.R.: Constrained Delaunay tetrahedralizations and provably good boundary recovery. In: Eleventh International Meshing Roundtable, Ithaca, NY, September 2002, pp. 193–204. Sandia National Laboratories (2002)
48. Shewchuk, J.R.: Delaunay refinement algorithms for triangular mesh generation. Comput. Geom. Theory Appl. **22**(1–3), 21–74 (2002)
49. Si, H., Gärtner, K.: Meshing piecewise linear complexes by constrained Delaunay tetrahedralizations. In: Hanks, B.W. (ed.) Proceedings of the Fourteenth International Meshing Roundtable, San Diego, CA, September 2005, pp. 147–163 (2005)
50. Sibson, R.: A brief description of natural neighbor interpolation. In: Barnett, V. (ed.) Interpreting Multivariate Data, pp. 22–36. Wiley, New York (1981)
51. Waldron, S.: The error in linear interpolation at the vertices of a simplex. SIAM J. Numer. Anal. **35**(3), 1191–1200 (1998)
52. Weatherill, N.P., Hassan, O.: Efficient three-dimensional grid generation using the Delaunay triangulation. In: Hirsch, Ch., Périaux, J., Kordulla, W. (eds.) Proceedings of the First European Computational Fluid Dynamics Conference, Brussels, Belgium, September 1992, pp. 961–968 (1992)
53. Weatherill, N.P., Hassan, O., Marcum, D.L., Marchant, M.J.: Grid Generation by the Delaunay Triangulation. Von Karman Institute for Fluid Dynamics 1993–1994 Lecture Series (1994)
54. Ziegler, G.M.: Lectures on Polytopes, 1st edn. Springer, New York (1995)